David Pelegrín García

The role of classroom acoustics on vocal intensity regulation and speakers’ comfort

PhD thesis, September 2011
The role of classroom acoustics on vocal intensity regulation and speakers’ comfort

PhD thesis by
David Pelegrín García

Technical University of Denmark
2011
This thesis was submitted to the Technical University of Denmark (DTU) as partial fulfillment of the requirements for the degree of Doctor of Philosophy (Ph.D.) in Electronics and Communication. The work presented in this thesis was completed between June 15, 2008 and September 13, 2011 at Acoustic Technology, Department of Electrical Engineering, DTU, under the supervision of Associate Professors Jonas Brunskog and Torben Poulsen. The project was funded by AFA Försäkring; reference 070142. All the experiments presented in this thesis were approved by the Science-Ethics Committee for the Capital Region of Denmark; reference H-KA-04149-g.

Cover illustration:
Composition with school teaching
at Nyvångskolan, Dalby\textsuperscript{1} and
the anechoic chamber at DTU\textsuperscript{2}
\textsuperscript{1}© Jonas P. Andersson, Skånska Dagbladet
\textsuperscript{2}© Sylvain Favrot,
modified with permission.
Abstract

Teachers are one of the professional groups with the highest risk of suffering from voice disorders. Teachers point out classroom acoustics among the potential hazards affecting their vocal health, together with air dryness, background noise, and other environmental factors. The present project has investigated the relationships between the classroom acoustic condition and teachers’ voice, focusing on their vocal intensity, and between the classroom acoustic condition and the sensation of acoustic comfort for a speaker.

In the presence of low background noise levels, teachers were found to adjust their vocal intensity according to the room gain or voice support of the classroom, which are equivalent objective measures that quantify the amplification of one’s own voice in a room due to the reflections at the room boundaries. Most of the vocal intensity variation among classrooms was due to differences in average teacher-to-student distance, but some of the variation was due to the room acoustic condition. The amount of vocal intensity variation with the room acoustic condition increased with the distance between teacher and student. In field measurements performed during typical working days, teachers with and without self-reported voice problems reacted identically to variations in noise, whereas they reacted differently to the voice support of the classrooms where they taught, suggesting that teachers with voice problems are more sensitive to the working environment than their healthy colleagues.

The acoustic conditions that conveyed the highest comfort for a speaker were derived from laboratory experiments in virtual classrooms and corresponded to values of the reverberation time between 0.45 and 0.55 s, calculated from the decay between -5 and -35 dB of the backward integrated energy curve of an impulse response measured between the mouth and the ears of a dummy head.

Prediction models for the reverberation time (calculated in the way described above) and the voice support were obtained, linking these measures to the volume and the traditional reverberation time of the room. Combining these models with the knowledge obtained during the project, speaker-oriented classroom acoustic design recommendations are given. These recommendations suggest that classrooms for flexible teaching should not have more than fifty students if optimum acoustic conditions for a speaker are to be met, and that, in smaller classrooms, the voice support should be between -12 and -8 dB.
Acknowledgments

It has been a long way toward the completion of the PhD, and even though in some moments the feeling of loneliness in research was unavoidable, many people have truly contributed to develop the work that is presented in these pages and deserve my utmost gratefulness.

My supervisors Jonas Brunskog and Torben Poulsen at the Technical University of Denmark (DTU) have always been supporting me, encouraging my work and showing me the light when I was in the dark. Now I can see that the challenges they put on me were the way to make me grow.

My fellow PhD student, Viveka Lyberg Åhlander from Lund University, proved to be an excellent team player, making easy, smooth, and flowing those tasks for which I lacked skills. Her supervisors Roland Rydell and Anders Löfqvist brought a lot of their expertise in voice to Viveka’s thesis, which extends to some of the papers in my dissertation.

My former thesis students Bertrand Smits and Oier Fuentes Mendizabal did two very valuable pieces of work, which after many days of looking at from many different perspectives, turned out to be of great importance to develop my thesis.

I am very grateful to Pasquale Bottalico and Arianna Astolfi, room acousticians of the Politecnico di Torino, with whom I share feelings of stepping in the unknown, when it comes to the effects of classroom acoustics on teachers’ voice health.

I got a lot of inspiration from Valdis Jónsdóttir and the members of the Nordic Group on Voice Ergonomics, who forced me to open my mind and go beyond the boundaries of room acoustics.

I really admire the genuine, honest, professional, and especially fast feedback of Finn Jacobsen, Head of Acoustic Technology at DTU. Through his vast experience and openness, Finn has also acted as a “third supervisor” during the PhD.

Cheol-Ho Jeong, Associate Professor at DTU, gave also much feedback on the
In this dissertation.

I want to thank Jesper Hagedorn, who conducted some laboratory experiments at DTU on my behalf while I was in Lund. I would have never had the time to conduct them myself.

The enthusiasm of Tobias Olesen and his work in classroom acoustics, validating one of the prediction models in my thesis, have been a big support and encouragement. Furthermore, his master thesis helped me to refresh some of the latest literature on classroom acoustics in a comprehensive way.

I had a great fun with the loudspeaker-based system to generate virtual acoustic environments. That would not have been possible without the work of David Santos Domínguez in an early stage, and without the work of Jörg Buchholz and Sylvain Favrot in the facility now called “Space Lab,” at DTU.

Many of the technical aspects when it came to experimental setups, or to having a proper computer system, were solved by my colleagues Jørgen Rasmussen and Tom Arents Petersen, whom I will remember for their kindness and exemplary professionalism.

Also many thanks go to all the colleagues at Acoustic Technology and fellow group Hearing Systems, with whom I had the pleasure of working together and learning from through interesting discussions, joining for lunch, and for some coffee-and-cake sessions on Fridays.

The experiments would have not been possible without the voluntary participation of the test subjects, teachers and students, who often gave part of their spare time for the research. I hope the results of this and further research will contribute to their well-being.

I want to recognize the inspiration and interesting insights of Anders Christian Gade on many of the steps in this project. His critical point of view has motivated different sections in this thesis.

Important contributors to this work are Antonio Carrión Isbert and the former colleagues at the acoustic consultant firm Audioscan SL, who guided me in my first steps into the world of acoustics and increased my longing for a deeper knowledge in this area.

Not uniquely related to the world of acoustics, the fellow singers in Akademisk Kor brought harmony and a musical counterpoint to the physical study of voice in rooms.

I have been lucky to share my spare time with many friends, from whom I had the
chance to learn, grow, laugh, listen to some cool jazz, and get a huge support. To them I owe much of this work.

I also owe much to Neringa, who endlessly encouraged me and brought the best out of me to work on this project, reminding me that there are more important things in life.

My family is doubtlessly who had the most difficult time, having me four years away from home at a distance of more than two thousand kilometers, but who nevertheless lovely cared for me and believed in me always.
List of publications

During the course of the PhD, and as a result of the work carried out in this area, the following papers were produced:

**Paper A** Comment on “Increase in voice level and speaker comfort in lecture rooms”  

**Paper B** Vocal effort with changing talker-to-listener distance in different acoustic environments  

**Paper C** Equal autophonic level curves under different room acoustics conditions  

**Paper D** Measurement and prediction of voice support and room gain in school classrooms  

**Paper E** Influence of classroom acoustics on the voice levels of teachers with and without voice problems: a field study  

**Paper F** Loudspeaker-based system for real-time own-voice auralization  
Manuscript (2010)

**Paper G** Speakers’ comfort and voice level variation in classroom: Laboratory research  
Manuscript (2011)

Several other papers were published in the course of the Ph.D., although they are not explicitly cited in this thesis due to overlapped content with the papers A to G.
List of publications

- D. Pelegrin-Garcia, and J. Brunskog, *Development of an auditory virtual environment to measure the speakers’ comfort and increase of voice levels in lecture rooms*. Proceedings of the First Nordic Conference of Voice Ergonomics and Treatment, Helsinki, Finland (2009)


Thesis at a glance

The thesis is a continuation of the work published by Brunskog et al. [16]. It contributes to increasing the understanding of the relationship between room acoustics, vocal intensity regulation, and speaking comfort, particularly applied to the teaching scenario. This knowledge can be used to design classrooms that maximize the acoustic comfort for speaking and that prevent teachers from using excessively high levels of vocal intensity, in order to improve the working conditions of teachers and as a step toward reducing the prevalence of voice problems among teachers.

Chapter 1, Introduction, describes the voice problems experienced by teachers; their causes, consequences, and the current preventive actions taken to minimize their prevalence.

Chapter 2, Interaction between room acoustics and the voice of a speaker, collects the knowledge about the dependence between room acoustics, vocal intensity adjustment, and speakers’ comfort, gained from the experiments and measurements in papers A to G, and puts it into context with previous investigations.

Chapter 3, Implications for classroom acoustics design, briefly reviews the traditional approach to classroom acoustic design based on the optimization of the conditions for listeners and uses the results presented in chapter 2 to propose alternative design strategies focused on the requirements of speakers.

Chapter 4, General discussion, reviews the main factors connected to the chosen methodology potentially affecting the findings of the study, evaluated the effectiveness of the design measures suggested in chapter 3, and gives directions for future research.

Finally, chapter 5 summarizes the main findings and conclusions of the work.

Summary of publications

Seven papers, either published in international scientific journals, in manuscript form under editorial process, or submitted as articles in conference proceedings, constitute the core of this thesis, and are included at the end of the dissertation:
Paper A: Comment on “Increase in voice level and speaker comfort in lecture rooms” [J. Acoust. Soc. Am. 125, 2072-2082 (2009)]

This paper revises the work of Brunskog et al. [16] on the vocal intensity adjustment of teachers in classrooms with different acoustic conditions and low levels of background noise. Paper A suggests an improved measurement method of classroom acoustic properties relevant for a speaker (called room gain and voice support) and presents corrected and simplified empirical models based on the measurement data from Brunskog et al. [16], describing vocal intensity variations as a function of room gain or voice support. These models are used as a reference scenario to compare with following studies.

Paper B: Vocal effort with changing talker-to-listener distance in different acoustic environments

The effects of talker-to-listener distance are separated from the effects of room acoustics on the vocal effort\(^1\) of speakers. The paper shows that the main factor affecting vocal intensity is distance, but nevertheless room acoustic conditions play an important role and explain the observations in paper A. Other voice parameters, as the fundamental frequency and the duration of phonated segments, also vary with the distance and the room acoustic conditions.

Paper C: Equal autophonic level curves under different room acoustic conditions

This investigation shows the vocal intensity needed in different acoustic conditions to keep the voice of a speaker equally loud at his/her own ears. The effect of different room acoustic conditions is here related to that of sidetone amplification [54] and to studies of the Lombard effect [55]. It is observed that room acoustics have a systematic effect on voice adjustment. However, the magnitude of the changes in vocal intensity is smaller than 2.3 dB in typical rooms.

\(^1\) Vocal effort, according to Traunmüller and Eriksson [112], is a physiological magnitude different from vocal intensity, which accounts for the changes in voice production required for the communication at different distances. Some descriptors of vocal effort are vocal intensity, fundamental frequency, phonation time, and spectral distribution.
Paper D: Measurement and prediction of voice support and room gain in school classrooms

Paper D presents the measurements of voice support and room gain in 30 primary and secondary school classrooms and proposes a prediction model for these parameters, based on geometrical properties and the reverberation time of the classrooms. The prediction model can be used during the design phase of educational spaces or rooms for speech, as a tool to assess the additional vocal loading experienced by a speaker due to the environment.

Paper E: Influence of classroom acoustics on the voice levels of teachers with and without voice problems: a field study

A field study of teachers with and without self-reported voice problems was carried out, in which their voice levels were monitored during real teaching and were related to the acoustics of the classrooms where they taught. The results show that both groups of teachers reacted identically to the noise present (according to the Lombard effect) and that the groups reacted significantly different to the voice support of the classrooms, suggesting that teachers with self-reported voice problems are more sensitive to changes in their working environment.

Paper F: Loudspeaker-based system for real-time own-voice auralization

A laboratory facility was specially built for this project, which allowed to generate the acoustics of virtual classrooms; this is, to induce in a speaker the auditory sensation of being talking in a space different from the actual laboratory room. Paper F describes its technical and design details.

Paper G: Speakers’ comfort and voice level variation in classroom: Laboratory research

Laboratory experiments were carried out using the setup described in Paper F, where the aim was to investigate further the relationship between voice support and vocal intensity, and to find optimum acoustic conditions for a speaker by means of questionnaires. The studies show that the performance in laboratory is highly dependent on the instruction. With the proper instruction, speakers react to the acoustic environment similarly to the findings of Brunskog et al. [16] and paper A. When speakers are not engaged into the communication task, they react to the acoustic environment keeping the loud-
ness of their own voice constant. The most preferred acoustic conditions for a speaker are indicated by a reverberation time between 0.45 and 0.55 s, derived from the decay between -5 and -35 dB of the backward integrated energy curve of an impulse response measured between the mouth and the ears of a dummy head.
Contents

Abstract iii
Acknowledgments v
List of publications ix
Thesis at a glance xi
List of abbreviations xix

1 Introduction 1
1.1 Voice problems in teachers ................................. 2
1.1.1 Definitions ........................................ 2
1.1.2 Prevalence ........................................ 2
1.1.3 Consequences ...................................... 3
1.1.4 Causes ............................................. 3
1.2 Current preventive actions ................................. 6
1.2.1 Voice training programs ............................. 7
1.2.2 Use of electroacoustic amplification ............... 7
1.2.3 Teaching methods and classroom management .... 8
1.2.4 Voice rest ........................................ 8
1.2.5 Classroom acoustic design ......................... 8

2 Interaction between room acoustics and the voice of a speaker 11
2.1 Definition of vocal parameters .............................................. 13
2.2 Speaker’s decision on the desired vocal parameters ................. 14
  2.2.1 Public loop feedback mechanisms in speech ..................... 16
  2.2.2 Private loop feedback mechanisms in speech ..................... 16
2.3 Components of one’s own voice ........................................... 18
2.4 Room Gain and Voice Support ............................................. 19
  2.4.1 Definition ........................................................................ 20
  2.4.2 Measurement ..................................................................... 21
  2.4.3 Bias factors affecting voice support and room gain ............. 25
  2.4.4 Prediction model ............................................................... 27
2.5 Reverberation time at the ears .............................................. 29
  2.5.1 Measurement ..................................................................... 30
  2.5.2 Prediction .......................................................................... 31
2.6 Vocal intensity under different room acoustic conditions: the room effect
  2.6.1 Distance factor ................................................................. 33
  2.6.2 Equal autophonic level curves ......................................... 34
  2.6.3 Field study: interaction of classroom acoustics and teachers’
       voice health ....................................................................... 38
  2.6.4 Laboratory experiments .................................................. 42
2.7 Other changes in voice production .......................................... 47
2.8 Acoustic comfort for a speaker .............................................. 49
2.9 Summary of findings ............................................................... 52

3 Implications for classroom acoustics design ............................... 55
  3.1 Listener-oriented classroom acoustics design ......................... 56
  3.2 Speaker-oriented classroom acoustics design ......................... 58
  3.3 Combined design for listeners and speakers ......................... 62
  3.4 Summary ............................................................................. 64

4 General discussion .................................................................... 67
  4.1 Factors affecting the characterization of classroom acoustic conditions . 68
     4.1.1 Measurement equipment dependency .............................. 68
Contents

4.1.2 Body conduction ................................................. 68
4.1.3 Lack of validation ............................................. 69
4.2 Sources of bias in vocal intensity ................................. 70
  4.2.1 Vocal intensity measures .................................. 70
  4.2.2 Intersubject variation ................................. 71
4.3 Sources of bias in the judgment of acoustic comfort ....... 72
4.4 Factors affecting the room effect ................................. 73
4.5 About the statistical methods .................................. 74
4.6 Effectiveness of the suggested measures ...................... 75
4.7 Future directions ................................................ 76

5 Conclusions .................................................. 79

A Optimum acoustic conditions for a speaker .................. 83
  A.1 Student-activity noise level ................................. 84
  A.2 Voice power level of a speaker .............................. 84
  A.3 Speech SPL in a room ...................................... 86
  A.4 Criteria and recommendations for optimum acoustic design for a speaker 86
  A.5 Signal-to-noise ratio ....................................... 92

Bibliography .................................................. 95

Papers A-G .................................................. 105
List of abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANCOVA</td>
<td>Analysis of covariance</td>
</tr>
<tr>
<td>ANOVA</td>
<td>Analysis of variance</td>
</tr>
<tr>
<td>B&amp;K</td>
<td>Brüel &amp; Kjær</td>
</tr>
<tr>
<td>$C_{50}$</td>
<td>Early-to-late ratio</td>
</tr>
<tr>
<td>CI</td>
<td>Confidence interval</td>
</tr>
<tr>
<td>F0</td>
<td>Fundamental frequency</td>
</tr>
<tr>
<td>$G_{RG}$</td>
<td>Room Gain</td>
</tr>
<tr>
<td>HaTS</td>
<td>Head and Torso Simulator</td>
</tr>
<tr>
<td>HRTF</td>
<td>Head-related transfer function</td>
</tr>
<tr>
<td>IR</td>
<td>Impulse response</td>
</tr>
<tr>
<td>$L_{50}$</td>
<td>SPL exceeded 50% of the time</td>
</tr>
<tr>
<td>$L_W$</td>
<td>Voice power level</td>
</tr>
<tr>
<td>$\Delta L_W$</td>
<td>Relative voice power level</td>
</tr>
<tr>
<td>MLS</td>
<td>Maximum Length Sequence</td>
</tr>
<tr>
<td>PTR</td>
<td>Phonation time ratio</td>
</tr>
<tr>
<td>SD</td>
<td>Standard deviation</td>
</tr>
<tr>
<td>SNR</td>
<td>Signal-to-noise ratio</td>
</tr>
<tr>
<td>SPL</td>
<td>Sound pressure level</td>
</tr>
<tr>
<td>STI</td>
<td>Speech transmission index</td>
</tr>
<tr>
<td>$ST_V$</td>
<td>Voice support</td>
</tr>
<tr>
<td>$T_{30}$</td>
<td>Reverberation time</td>
</tr>
<tr>
<td>$T_{30,ears}$</td>
<td>Reverberation time at the ears</td>
</tr>
<tr>
<td>$U_{50}$</td>
<td>Useful-to-detrimental ratio</td>
</tr>
<tr>
<td>VHI-T</td>
<td>Voice Handicap Index with Throat subscale</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

Teaching is one of the occupations with highest vocal demands and represents an important share of the total workforce of a country. The prevalence of voice problems among teachers is much higher than the average among other occupations due to their use of voice at work. Teachers have to teach in a variety of rooms of different dimensions and acoustic conditions, which modify their voice behaviors and the behaviors of students. Even though teachers claim that classroom acoustics is a factor that affects their voice, very little research has been done to understand the cause-effect relationship between voice use and room acoustics.

This dissertation is part of a larger project, named *Speakers’ comfort and voice disorders in classrooms*, carried out in collaboration with the Department of Logopedics, Phoniatrics and Audiology at Lund University. In the project, the focus was to examine the voice behavior of teachers—one of the professional voice user groups at larger risk of suffering from voice disorders—at work and its relation to room acoustics. The work carried out by the partners in Lund was focused on health aspects and led to the doctoral dissertation of Lyberg-Åhlander [63] *Voice use in teaching environments: Speakers’ comfort*.

The understanding of the interaction between room acoustics and voice use is necessary in order to define preventive actions based on classroom acoustic design and planning.
1. Introduction

1.1 Voice problems in teachers

1.1.1 Definitions

Voice problems is a term widely used, although there seems not to be consensus on the definition [69], and here is used with the same meaning as Lyberg-Åhlander [63] does. It is necessary to define the terms vocal loading and vocal fatigue to understand voice problems. Vocal loading refers to the natural adaptation in the phonatory apparatus during voice production. Signs of a tolerable voice loading during a working day are a rise in F0, a rise in SPL, a rise in phonatory threshold pressure, and a change towards hyperfunction. Vocal fatigue is a term used to refer to the negative sensations experienced after a period of vocal loading, accompanied with physiological, perceptual, and subjective changes. However, some of the fatiguing changes are necessary to avoid physiological overstraining [116, p.36]. In any case, the symptoms of vocal fatigue indicate the necessity of having a period of vocal rest. Titze [108] suggests a model in which prolonged periods of vocal loading without enough vocal rest can result into permanent damage of the vocal folds.

1.1.2 Prevalence

Teachers are one of the professional groups who suffer more frequently from voice problems:

1. Teachers are overrepresented in voice clinics: according to Fritzell [24], there were 16.3% of teachers as patients in Swedish voice clinics in 1990, but only represented 5.9% of the working population. According to Titze et al. [110], the percentages in the US were in the same order of magnitude: 19.6% of the patients in voice clinics were teachers, whereas they constituted only the 4.2% of the working population.

2. Teachers suffer from voice problems twice as much as other professional groups during their careers [89, 96].

3. The prevalence of voice problems among teachers in the present study, regionally located in southern Sweden, was 13% [66]. This quantity referred to those teachers reporting having voice problems occurring sometimes, often, or always. The prevalence is similar to that reported by Russell et al. [90], who found that
1.1 Voice problems in teachers

16% of the teachers self-reported voice problems at the moment of the study, and by Roy et al. [89] (11%).

1.1.3 Consequences

The voice problems can cause teachers to be absent from work in order to receive treatment and be able to recover. Smith et al. [97] reported that over 20% of teachers had missed some working days due to voice problems, but none of the non-teachers had. Moreover, McAleavy et al. [71] stated that 32% of the days of teachers’ absence leave were due to voice problems, whereas 30% of the days were due to stress. Verdolini and Ramig [113] estimated the social costs in the US due to teachers’ voice problems—sick-leave and treatment—to amount approximately $2.5 billion per year. Voice problems can become so frequent and serious that can turn into permanent damage of the vocal organ and working disability [69, 91]. Furthermore, employers can be liable for negligences leading to permanent vocal damage in employees, such as the case of Joyce Walters [30], who received a payout of £156.000 in compensation for permanent voice problems that terminated her career as a teacher. Furthermore, Rogerson and Dodd [87] pointed out that teachers’ dysphonic voices have a detrimental effect on children’s performance.

1.1.4 Causes

It is generally agreed upon that voice problems result from the combination of vocal loading and the individual capacity of coping with loading [115]. Individual risk factors affecting the capacity of coping with voice loading, according to Vilkman [115], are

**Gender**  The prevalence of voice problems among females is much higher than among males [24]. This can be due to a higher fundamental frequency in adult females [107], related to the gender differences in size and physiology of the larynx, and some more subtle anatomical changes in the vocal folds and the surrounding tissues [19].

**Health condition**  Teachers who suffer or have suffered from hearing problems, allergic reactions, or respiratory infections are more affected from vocal fatigue than their healthy colleagues [29].
Life habits Smoking or drinking coffee can be detrimental for vocal health [84]. Voice-demanding activities outside work, such as solo or choir singing can prevent voice from having enough rest.

Vocal skill/experience Kooijman et al. [50] found an indication that young teachers experienced more vocal problems than more experienced teachers, due to longer working hours and less vocal hygiene habits than their more experienced colleagues. However, other studies [89], do not agree and point out group ages between 40 and 59 years old as more likely to experience voice problems.

Psychosocial and personality factors Kooijman et al. [51] and Roy et al. [88] point out that psycho-emotional factors, such as introversion/extroversion, play an important role in the development and consolidation of voice problems. Gassull et al. [27] points out that teachers with voice problems have a greater reactivity to stress than their healthy colleagues, and McAleavy et al. [71] found out the presence of trait anxiety to be relevant in voice health.

Other factors contributing to the vocal loading, not dependent on the individual, but linked to the work environment are

Duration of phonation By definition, vocal loading increases with phonation time and can lead to vocal fatigue if there is not enough time of vocal rest [116]. According to Vilkman [115], the teaching profession has very high demands regarding vocal endurance, because teachers need to use their voice for several hours every working day.

Intensity of phonation According to Titze [109], the mechanical stress suffered by the vocal folds—suggested to be cause for different voice disorders—is dependent on the amplitude of vibration, which determines the intensity of phonation. The teaching methods and the need of applying discipline can require using high vocal intensity.

Long speaking distance Speakers raise their vocal intensity with increasing distance to the listener [58, 73, 112, 119]. This effect is explained in more detail in section 2.2.

Air quality High air humidity contributes to the lubrication of the vocal folds. Dry vocal folds can become easily irritated [45]. Dust in the air is often reported as a
1.1 Voice problems in teachers

risk factor [114], but there is no evidence that it leads to an increased prevalence of voice complaints [84, 95].

Other ergonomic factors The talking posture, or whether a speaker is seating or standing, and how the head, the back and the neck muscles are positioned, are regarded as important in the quality and efficiency of the voice produced [45, 115, 116].

Psychosocial factors The teaching environment is regarded often as stressful. In Poland, teachers are the professional group who feel most stressed at work (34% versus an average 26.6%) [22].

Background noise Teachers raise their voice levels in the presence of noise to make themselves understood. This is known as the Lombard effect [55] and is described more in detail in section 2.2. Furthermore, there are some indications—though not conclusive—that noise in classrooms induces stress in teachers, even at moderate levels [104]. A study by Schönwälder et al. [94] reported that 80% of 1200 teachers considered pupils’ noise as a stress factor.

Room acoustics Tiesler and Oberdörster [104] also indicated that classrooms with short reverberation times led to lower stress levels on teachers. In terms of vocal intensity, talkers tend to speak louder in acoustically dry rooms, whereas they tend to speak softer in more “live” rooms [5]. However, more “live” rooms increase the activity levels caused by students, which are usually the main source of “noise”. Hodgson et al. [33] proposed an empirical model based on measurements in university classrooms, which took into account the student activity noise. Brunskog et al. [16] found that teachers modified their vocal intensity according to the objective measure room gain, which indicates the degree of amplification offered by the room to the voice of the speaker at his ears. Kob et al. [49] noted that teachers with voice problems are more affected by unfavorable room acoustic conditions than their voice healthy colleagues. Determining the actual relationship between the classroom acoustic conditions and the variations in vocal intensity experienced by teachers in the line started by Brunskog et al. [16] was one of the goals in this project. Chapter 2 is dedicated to this topic.

Teaching methods Oberdörster and Tiesler [78] found that different teaching methods such as frontal lessons (in which the teacher addresses the pupils) or student-
centered lessons (with focus in group-work and discussion) result in different noise levels, and more important, in different interactions with room acoustics. I.e., there was a significant reduction in sound pressure level (SPL) during student-centered lessons after refurbishment of a classroom (which reduced the reverberation time), whereas the SPL during frontal lessons did not change significantly.

Despite the knowledge of the risk factors that eventually can lead to voice problems, these are not yet understood, because self-reported voice problems do not always have a correlation with objective acoustic measures [56] or features assessed with laryngological examinations [65]. For other occupational diseases, as for example occupational hearing loss, the relationship between exposure to noise, recovery periods, and hearing damage is well documented (see, e.g., Gelfand [28]).

Titze [108], from the observation of other studies in occupational health, suggested to establish safety vocalization limits. Some steps have been taken in this direction, by introducing the so-called vocal doses to quantify the exposure of the vocal folds to vibration [111], determining vocal recovery trajectories [36], and comparing the use in occupational and non-occupational settings [37]. Yet, Hunter [35] points out that more research is needed in order to distinguish with confidence the effects of environmental factors on voice use, and to determine how non-occupational voice use affects vocal rest. Hunter [35] states that the main problem in the research area is the current lack of a real metric to show vocal impairment.

1.2 Current preventive actions

Voice problems or disorders are seen as an issue of Occupational Health and Safety (OSH) by many scientists. The term vocoergonomics or voice ergonomics [46, 115] is used to refer to the actions taken to prevent and treat voice disorders as a consequence of its use in the working environment. However, only in Poland voice disorders are listed among occupational diseases [80].

Preventive actions aim at improving the personal ability of coping with vocal loading or at reducing vocal loading in itself. On the one hand, voice training programs aim at improving the individual capacity of coping with vocal loading. On the other hand, actions that aim at reducing vocal loading are the use of electroacoustic amplification,
1.2 Current preventive actions

pedagogic instruction on how to deal with noise in classrooms, teaching schedules that include possibility of rest, and classroom layout and acoustic design.

1.2.1 Voice training programs

Ilomäki et al. [39] found that long-term voice training programs were an effective tool to reduce the prevalence of voice disorders and suggested that short-term programs, such as vocal hygiene lectures, could be useful to raise awareness of vocal symptoms, but were not efficient to improve vocal endurance. Furthermore, long-term voice training programs seemed to increase vocal endurance and well-being [38], which can lead to higher satisfaction of teachers at work. Timmermans et al. [106] also found beneficial effects of training programs in future teachers four months after the instruction.

In addition, there are different initiatives to encourage good vocal hygiene among teachers, much of which include voluntary work (e.g., the Voice Care Network in the UK [118]) and guidelines (e.g., [103]). It is commonly pointed out at the necessity of developing good voice use through education programs early at university and through the availability of voice care initiatives for the support of teachers during their careers, which have to be jointly arranged by higher education institutions, schools, local authorities and speech and language therapists. However, voice education in the current university programs for future teachers is not sufficient [103].

1.2.2 Use of electroacoustic amplification

The use of electroacoustic amplification (or sound-field amplification) is beneficial for both students and teachers, as has been widely documented in the literature (see, e.g., the review article by Millett [74] or the PhD thesis of Jónsdóttir [45]). The sound-field amplification increases the speech SPL from the teacher across the classroom, which results in increased signal-to-noise ratio (SNR) and increased speech intelligibility from the students. This feature is particularly helpful in classrooms, because children require higher SNR than adults to achieve the same speech intelligibility scores [8, 12, 98]. Furthermore, children learning second languages require even higher SNR than native speakers of that language [70]. Children with temporary or permanent hearing loss also require higher SNR than normal-hearing children [75].

Sapienza et al. [92] reported that teachers using sound-field amplification during teaching lowered their vocal intensity by 2.4 dB when compared to teaching in a non-
1. Introduction

amplified setting. Jónsdóttir [44] found that most of the teachers using sound field amplification claimed that voice production became easier, that vocal endurance was improved, and that the need for repetition diminished.

In the present dissertation, however, it is assumed that the voice of the teacher or the students is not amplified by electroacoustic means unless it is explicitly stated.

1.2.3 Teaching methods and classroom management

The teaching or pedagogical method is a factor contributing to voice loading, as it is linked to the time that the teacher spends speaking and to the noise that the students produce. Pedagogical methods are rapidly changing and more importance is given to co-operative and group-work oriented approaches rather than to traditional lecturing in primary and secondary schools [105].

In addition, good classroom management skills can help the teacher to keep pupils focused and engaged on learning rather than on the noise of competing distractions [14] and therefore keep the noise levels under control.

1.2.4 Voice rest

Titze [108] compared the process of vocal loading to that of tissue injury of muscles in athletes, pointing out the importance of periods for voice recovery, and showing that continuous vocal loading can lead to permanent injuries. In this context, teaching schedules can be adjusted in order to introduce regular pauses to allow for short recovery and longer breaks after voice demanding activities such as lecturing style lessons.

1.2.5 Classroom acoustic design

Classroom acoustic design has an important role on voice production. The introduction of acoustically absorptive material in the classroom can, on the one hand, effectively reduce the noise from the students and increase the length of the periods with silence [105]. On the other hand, it can lead to increased vocal intensity of teachers, who perceive their voice damped and raise it in consequence [16]. Chapter 2 aims at describing this second effect through the links between voice production and the perception of one’s own voice. Chapter 3 combines the two effects to suggest classroom acoustic
1.2 Current preventive actions

designs where teachers do not have to raise their voice much, either due to excessive absorption or to excessive noise from the students.
1. Introduction
Chapter 2
Interaction between room acoustics and the voice of a speaker

One of the definitions of voice is the “expiration of air with the vocal cords drawn close so as to vibrate audibly” [117]. Moreover, voice is the acoustic output of a muscular activity—involving respiratory, laryngeal, and articulatory muscles—triggered by motor commands produced in the brain. The factors influencing the motor commands in the brain that result in a particular utterance or vocal sound are shown in figure 2.1. In the same figure, the factors potentially affected by the room characteristics are shown in red.

Figure 2.1 illustrates the process of voice adjustment as a closed loop system; a system with feedback, which continuously monitors the results of the actual output voice to fit it to the “desired” output, i.e., the intention of the talker to speak a particular utterance.

Section 2.1 introduces the main vocal parameters used to characterize average properties of the utterances. Important parameters to describe the utterance include vocal intensity, fundamental frequency F0, spectral content, and duration of phonation; however, the present study focuses on vocal intensity. Section 2.2 examines the factors affecting the intention to speak a particular utterance, describing the public and private (or personal) feedback mechanisms available to a speaker for monitoring and adjusting the vocal parameters.

Section 2.3 presents the paths that one’s own voice follows in order to produce
2. Interaction between room acoustics and the voice of a speaker

Figure 2.1: Conceptualization of the motor control feedback systems in speech available to a speaker, adapted from Borden et al. [7] (on gray background). Other elements show factors contributing to the intention to speak a particular utterance, including the feedback from the listener. The factors potentially affected by the room characteristics are shown in red.
an auditory sensation. Sections 2.4 and 2.5 introduces different objective parameters that are related to the acoustic feedback of one’s own voice. Section 2.6 describes the changes of vocal intensity as a response to different room acoustic conditions, whereas section 2.7 describes the changes in other voice parameters. Section 2.8 describes the effect of different room acoustic conditions on the subjective perception of the environment in terms of reverberance and acoustic comfort for speaking. Finally, section 2.9 summarizes the main findings.

### 2.1 Definition of vocal parameters

To characterize long intervals of continuous utterances or speech, different parameters are used.

The **vocal intensity** is the magnitude at focus in the present dissertation and refers to the vibratory amplitude of the vocal folds, which is correlated with the sound power radiated from the mouth of a speaker. The concept of vocal intensity is also referred to as **voice level** in a qualitative way, and it is quantified with different physical measures in the different papers:

- **Voice power level** ($L_W$), or sound power level of the voice, which is the sound power radiated from the mouth of a speaker.

- **Sound pressure level** (SPL) at a microphone position close to the mouth. If the microphone is close enough to the mouth of the speaker, the increase of SPL due to the reflections of the room is negligible.

- **Equivalent on-axis, free-field SPL** at 1 m in front of the speaker. This measure is derived from the previous one, using a correction measurement that accounts for the SPL difference between the SPL at the microphone close to the mouth and the on-axis, free-field SPL at 1 m in front of the speaker.

The term **speech SPL** is reserved for the SPL that the voice of a speaker arises at the listener position, including the effect of the reflections in the room.

The **fundamental frequency** (F0) describes the number of vibrations per second performed by the vocal folds when producing a voiced sound. In the present work, F0 is calculated in intervals of 10 ms to obtain a time sequence of F0 values. Only the
2. Interaction between room acoustics and the voice of a speaker

Phonated segments of speech are taken into account to derive the F0 sequence. From the F0 sequence, two quantities are given:

- F0 mean, as the average of the F0 sequence
- F0 standard deviation (SD), as the sample standard deviation of the F0 sequence

The phonation time defines the time that the vocal folds are under vibration. Usually, it is expressed as a relative value, i.e., duration of phonation per unit of time. In paper B, the term Phonation Time Ratio (PTR) is used to define the relative duration of phonated segments in running speech, i.e., the speech signal processed to remove those relatively long lapses of silence.

There is a large amount of information about the changes in vocal effort in the speech spectrum (frequency representation of speech). It is sensitive to the individual characteristics of the subject and to the speech material used (e.g., a single vocalization, reading of a phonetically balanced sentence, natural speech).

2.2 Speaker’s decision on the desired vocal parameters

As figure 2.1 illustrates, the process of voice production starts with the intention of speaking a particular utterance or voice sound—as a part of speech at a higher cognitive level—with different parameters. The intention is dominated by a strong desire of establishing a successful communication with the listener. There are other factors which influence the parameters of the utterance to be produced.

The first factor is the context in which the voice is used. For example, a speaker might use his voice to establish authority, to give clear instructions, to sing, or to act.

Secondly, the intention of a speaker of using different vocal parameters can be motivated by the knowledge of listener characteristics: additional clarity should be conveyed to speech—through articulation, rate of speech, and other voice quality variations—if the listener has some speech perception deficits. These deficits can be due to hearing impairment, but also to underdevelopment in the speech perception abilities, which is normal in the case of children and of people who are not native speakers of the language used by the talker.

Third, the background noise that a speaker hears influences his vocal parameters in a reflexive act called the “Lombard effect,” named in tribute to Étienne Lombard, a French otolaryngologist who first reported that speakers raise their voice level in the
2.2 Speaker’s decision on the desired vocal parameters

presence of noise [61]. The acoustic changes in the utterances produced under background noise are increased voice level, increased F0, a shift in energy towards higher frequencies, increase in vowel duration, spectral tilting, and shift in the two first formant frequencies of vowels [47]. In a summary of different works, Lazarus [57] reported that speakers raised their voice levels by 0.3 to 0.6 dB for each dB of increase in noise level above 45 dB during face-to-face conversations or in telecommunication systems. However, Pearson et al. [81] found that teachers rose their voice level up to 1 dB per dB of noise under actual teaching. Even though the Lombard effect is dominated by the premium on successful communication, its reflexive nature makes it difficult to be inhibited [82]. In situations with multiple speakers, there is one speaker of interest for a particular listener and a number of interfering speakers. The speech from interfering speakers is perceived as noise, therefore a speaker will raise his vocal intensity to overcome the noise. At the same time, other speakers will raise their vocal intensity as a response to increased interferer speech SPL. This feedback loop affecting the speakers is a common effect found in, for example, cocktail parties [67] and is more commonly described as café effect [120]. Therefore, addition of sound absorbing materials in rooms is an efficient way of reducing conversational noise [52].

Fourth, the distance to the listener strongly influences the voice parameters used, specially the vocal intensity. Different studies on this topic [31, 43, 58, 68, 73, 112, 119, 122], including paper B in the thesis, point out that the voice level increases almost linearly with the logarithm of the distance between speaker and listener.1 The results of different studies on this topic found in the literature are summarized in table 2.1. The effect of distance on voice level adjustment varies across studies and is likely an effect of instruction. When the instruction was to provide a constant level at the listener, or when the listener gave feedback to the speaker, the variations were close to 6 dB per double distance [112, 119, 122]. In other cases in which the speaker was given no feedback by the listener—in an unsupervised condition—the effect of distance on voice level was much lower, between 1 and 2 dB per double distance [43, 68, 73]. In paper B, the measured effect of distance on voice level, averaged across subjects was between 1.3 dB per doubling distance in a reverberation room, and 2.2 dB per doubling distance in an anechoic room, whereas more common spaces had intermediate effects.

---

1 In paper B, the amount of voice level variation per double distance is called compensation rate, instead of effect
The variation in the effect of distance on voice level under different room acoustic conditions is discussed in section 2.6.1.

The intention of speaking a particular utterance is adjusted with feedback mechanisms, which can form part of a public loop or a private loop. The feedback mechanisms of the public loop involve the presence of other people, whereas private loop feedback mechanisms are those used by the speaker himself.

2.2.1 Public loop feedback mechanisms in speech

The only public loop feedback mechanism shown in figure 2.1 is the interaction between the listener and the speaker following the reactions of the listener after hearing the voice of the speaker. The listener can express himself verbally (through request) or non-verbally (facial expression, lack of attention) so that the speaker gains valuable information on how to modify his vocal parameters. For example, in a large room, listeners located far away may ask the speaker to increase his vocal intensity in order to hear him better. The acoustic conditions of the room where communication takes place can affect listeners’ perception of speech, demanding raised voice levels or improved articulation from the speaker.

2.2.2 Private loop feedback mechanisms in speech

The private loop feedback mechanisms indicated by Borden et al. [7] are shown on gray background in figure 2.1, together with the general mechanisms that result in voice production.

The intention to speak a particular utterance activates a series of motor commands through the central nervous system. The central nervous system might retrieve these commands from stored spatial-temporal speech patterns (in order to produce complex speech sounds) in the so-called internal feedback. The motor commands in the central nervous system are sent to different muscles and muscle spindles through efferent nerves, and result in the activity of respiratory, laryngeal, and articulatory muscles. Along the same muscle spindles, there are afferent neurons and nerves that detect changes in the length of the muscles and send information back to the central nervous system—this is the so-called proprioceptive feedback, which is the primary response feedback.

The coordinated activity of the respiratory, laryngeal, and articulatory muscles re-
### 2.2 Speaker's decision on the desired vocal parameters

<table>
<thead>
<tr>
<th>Paper</th>
<th>Remarks</th>
<th>Footnote Remarks</th>
<th>Room Properties</th>
<th>Distance (m)</th>
<th>Effect (dB per double distance)</th>
<th>Range of distances (m)</th>
<th>Room properties</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Describe a map to a listener</td>
<td>Room A, a recreation room, four corners, an audience</td>
<td>Provide a common level at the listener position</td>
<td>0.6</td>
<td>1.5 — 2</td>
<td>1.2 — 2.2</td>
<td>1.5 — 12</td>
</tr>
<tr>
<td>B</td>
<td>Describe a map to a listener</td>
<td>Room B, a recreation room, four corners, an audience</td>
<td>Provide a common level at the listener position</td>
<td>0.6</td>
<td>1.5 — 2</td>
<td>1.2 — 2.2</td>
<td>1.5 — 12</td>
</tr>
<tr>
<td>C</td>
<td>Provide a common level at the listener position</td>
<td>Room C, a recreation room, four corners, an audience</td>
<td>Provide a common level at the listener position</td>
<td>0.6</td>
<td>1.5 — 2</td>
<td>1.2 — 2.2</td>
<td>1.5 — 12</td>
</tr>
<tr>
<td>D</td>
<td>Provide a common level at the listener position</td>
<td>Room D, a recreation room, four corners, an audience</td>
<td>Provide a common level at the listener position</td>
<td>0.6</td>
<td>1.5 — 2</td>
<td>1.2 — 2.2</td>
<td>1.5 — 12</td>
</tr>
</tbody>
</table>

**Table 2.1**: Summary of the investigations that have examined the increase in vocal intensity with increasing distance between speaker and listener.
sults in movements of the different elements in the phonatory apparatus, including the lungs, the vocal folds, the articulators (such as pharynx, tongue, lips, teeth, palate, and alveolar ridge), which results in the physical production of sound waves (voice) and in mechanical vibrations transmitted through the body. The articulatory contacts and the changes in air pressure inside the phonatory apparatus activate the surface receptors that send tactile feedback to the central nervous system. This kind of feedback is called external feedback, because the receptors involved in it are sensible to external stimuli. The generated sound waves propagate through the air, while the mechanical vibrations propagate through the body, and they reach the cochlea, producing an auditory sensation that is used in the central system to monitor the voice parameters—constituting another path of external feedback. The auditory feedback is explained in more detail in section 2.3, and becomes modified under different room acoustic conditions.

2.3 Components of one’s own voice

After the voice is produced, a speaker hears his voice via the mechanical vibrations transmitted through the body (body-conducted sound) or via the sound waves propagated through the air (airborne sound) which arrive at the cochlea. Studies by Békésy [4], Pörschmann [83], Reinfeldt et al. [85] have shown that the body conducted and the airborne components of one’s own voice have about the same importance, though Reinfeldt et al. [85] showed that the relative importance varies for different vocalizations and sounds.

At the same time, the airborne transmission of one’s own voice consists of the direct airborne sound path between the mouth and the ears—which is affected by the diffraction around the head and all the scattering at the pinna—and the sound which is radiated away from the speaker, reflected at the environment boundaries, and returning to the ears of the listener. This last component is called reflected sound or indirect sound, and can also be affected by the presence of electroacoustic amplification systems. The three identified components of one’s own voice—body-conducted sound, direct airborne sound and reflected sound—are shown schematically in figure 2.2.

One’s own voice is sometimes called sidetone. This term is found in the literature as early as in 1893 [99] to denominate the loud sound of one’s own voice echoed by the first telephone systems. Psychoacoustic research in this area was directed toward determining the vocal behavior of a speaker under different sidetone amplification levels,
2.4 Room Gain and Voice Support

including lack of sidetone. Lombard [61] noted that people who deafened raised their voices abnormally. Black [6] measured the effect of hearing loss on voice level by inducing temporary threshold shift on male college students after exposure to loud noise, and found that speakers raised their voice 0.58 dB per dB of the induced temporary threshold shift. Lane et al. [54] determined that speakers varied their vocal intensity by -0.46 dB per each dB of sidetone amplification to hold the perceived loudness of their own voice constant and called this effect sidetone compensation. Lane and Tranel [55] said that the sidetone compensation and the Lombard effect were two sides of the same coin at the light of the observed results.

The reflected component of one’s own voice depends on the acoustic environment, but it is usually much lower in magnitude than the body-conducted and the direct airborne components of one’s own voice. The next section introduces two measures (room gain and voice support) to quantify the relative importance of the reflected component and the direct airborne component of one’s own voice. Further below, section 2.6.2 reports the results of paper C, which studies the room acoustic conditions as a special case of sidetone, and presents the variations of voice level that keep the autophonic level constant under different room acoustic conditions.

2.4 Room Gain and Voice Support

The importance of the reflected component of one’s own voice is judged with two alternative measures introduced by Brunskog et al. [16]: the room gain and the voice support. The perceived loudness of one’s own voice is also known as autophonic rating, and the perceived loudness level of one’s own voice is called autophonic level.

---

2 The perceived loudness of one’s own voice is also known as autophonic rating, and the perceived loudness level of one’s own voice is called autophonic level.
2. Interaction between room acoustics and the voice of a speaker

This importance is reflected through the continuous use of these measures in papers A to E and G.

2.4.1 Definition

The room gain $G_{RG}$ was the first measure introduced by Brunskog et al. [16] and is defined as the degree of amplification offered by the room to the speaker’s voice at his ears, considering only the airborne paths. Let the airborne direct sound reaching the ears have energy $E_D$ and the reflected sound have energy $E_R$, then

$$G_{RG} \approx 10 \log \left( \frac{E_D + E_R}{E_D} \right) \text{[dB]}, \quad (2.1)$$

assuming that the total energy is the sum of the energies of the direct sound and the reflected sound. The voice support $ST_V$ is an alternative measure that is defined as the energy ratio (in dB) between the reflected sound and the airborne direct sound,

$$ST_V = 10 \log \left( \frac{E_R}{E_D} \right) \text{[dB]} \hspace{1cm} (2.2)$$

The nomenclature $ST_V$ used for the voice support is defined after the work of Gade [25, 26] on the acoustics of stages in concert halls, where the measure objective support $ST_{early}$ is used to assess the acoustical quality from the performers’ point of view.

The purpose of the room gain and the voice support is to establish a metric that ranks rooms in terms of natural amplification offered to the voice of a speaker. The relationship between the two measures is

$$G_{RG} \approx 10 \log \left( 10^{\frac{ST_V}{10}} + 1 \right) \text{[dB]}, \quad (2.3)$$

with the same assumption regarding energy summation as in Eq. (2.1). This relationship is illustrated in Fig. 2.3. Paper D reported measured values of $ST_V$ in rooms in the range between -20 dB and -5 dB. The room gain, on the other hand, is between 0.045 dB and 1.2 dB. The higher range of values for the voice support makes it a more suitable parameter in architectural acoustics than the room gain.

The room gain and the voice support can also be defined in terms of energy level differences. Given the total energy level $L_E$, the energy level of the direct sound $L_D$,
2.4 Room Gain and Voice Support

Figure 2.3: Relation between voice support and room gain (black, bold). The dotted line $G_{RG} = ST_V$ is shown to illustrate the asymptotic value of room gain for high values of voice support.

and the energy level of the reflected sound $L_R$,

$$L_E = 10 \log \left( \frac{E_D + E_R}{E_0} \right) \text{[dB]} \quad (2.4a)$$

$$L_D = 10 \log \left( \frac{E_D}{E_0} \right) \text{[dB]} \quad (2.4b)$$

$$L_R = 10 \log \left( \frac{E_R}{E_0} \right) \text{[dB]}, \quad (2.4c)$$

where $E_0$ is an arbitrary energy reference, the room gain is alternatively defined as

$$G_{RG} = L_E - L_D \text{[dB]}, \quad (2.5)$$

and the voice support as

$$ST_V = L_R - L_D \text{[dB]}. \quad (2.6)$$

2.4.2 Measurement

Brunskog et al. [16] initially proposed a measurement method for the room gain. Paper A proposed an alternative measurement method which was refined in paper D regarding the frequency weighting.

These methods are based on the measurement of impulse responses (IRs) between the mouth and the ears of a dummy head. A simplified representation of the setup used to measure these IRs is shown in figure 2.4, and corresponds specifically to the
The method proposed by Brunskog et al. [16] required the measurement of two IRs between the mouth and the ears of a head and torso simulator: one in an anechoic chamber and another one in the room of interest. The energy of the direct sound $E_D$ was extracted from the IR measured in the anechoic chamber, and the total energy including direct sound and reflections $E_{D+R}$ was extracted from the measurement in the room of interest. Finally, the room gain was calculated with the formula of Eq. (2.1), assuming that $E_{D+R} \approx E_D + E_R$ (i.e., energy summation).

The measurement of room gain and voice support as proposed by the author was carried out in a different way, although conceptually equivalent to Brunskog et al. [16] (for a discussion about the differences in the methods, see paper A). The proposed method calculates the energy of the direct sound and the reflections from a single IR in the room of interest.
2.4 Room Gain and Voice Support

It is assumed that the mouth and the ears are at least 1 m away from every reflecting and scattering surface. In this case, the direct sound $h_D(t)$ can be extracted from an IR $h(t)$ by multiplying it with a window function $w(t)$

$$h_D(t) = h(t)w(t)$$  \hspace{1cm} (2.7)

because all the reflections will have a delay of at least 5.8 ms from the arrival time of the direct sound. The window function is

$$w(t) = \begin{cases} 
1 & t < 4.5 \text{ ms} \\
0.5 + 0.5 \cos \left( 2\pi (t - t_0)/T_W \right) & 4.5 \text{ ms} < t < 5.5 \text{ ms} \\
0 & t > 5.5 \text{ ms}
\end{cases}$$  \hspace{1cm} (2.8)

with $t_0 = 4.5$ ms and $T_W = 2$ ms. The window function $w(t)$ is flat at unity from 0 to 4.5 ms and decays smoothly following half a period of a raised cosine function until it reaches a value of 0 at 5.5 ms. The reflected sound $h_R(t)$ is obtained by multiplying the IR by the complementary window function $1 - w(t)$,

$$h_R(t) = h(t)(1 - w(t)).$$  \hspace{1cm} (2.9)

An IR and the windowing functions are shown in figure 2.5.

![Figure 2.5: Example of an IR measured between the mouth and the ears $h(t)$ and the windowing applied to extract the direct and the reflected sound.](image-url)

The signals of the direct sound and the reflected sound are filtered with octave band
2. Interaction between room acoustics and the voice of a speaker

filters \( h_{F,i}(t) \) that have center frequencies 125 Hz, 250 Hz, 500 Hz, 1 kHz, 2 kHz, and 4 kHz (see table 2.2). Therefore, \[ h_{D,i}(t) = h_D(t) * h_{F,i}(t), \quad i = 1 \ldots 6 \] \[ h_{R,i}(t) = h_R(t) * h_{F,i}(t), \quad i = 1 \ldots 6 \] (2.10a) (2.10b)

where * is the symbol of the convolution operator.

Table 2.2: Center frequencies for the octave band filters \( h_{F,i}(t) \)

<table>
<thead>
<tr>
<th>( i )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Center frequency of ( h_{F,i}(t) ), Hz</td>
<td>125</td>
<td>250</td>
<td>500</td>
<td>1000</td>
<td>2000</td>
<td>4000</td>
</tr>
</tbody>
</table>

The energy of the direct sound and the reflected sound is calculated in octave bands as

\[ E_{D,i} = \int_0^\infty h_{D,i}(t) \, dt, \quad i = 1 \ldots 6 \] \[ E_{R,i} = \int_0^\infty h_{R,i}(t) \, dt, \quad i = 1 \ldots 6. \] (2.11a) (2.11b)

The voice support \( STV,i \) is calculated in each of the octave bands using Eq. (2.2) with the energies for the direct sound \( E_{D,i} \) and the reflected sound \( E_{R,i} \) of the \( i \)-th octave band.

In order to obtain a single value descriptor of voice support, a frequency weighting is applied to the \( STV,i \) values in the octave band. The reference spectrum is the typical speech level at the ears \( L_{ref,ears} \), indicated in table 2.3. The overall speech-weighted

<table>
<thead>
<tr>
<th>Center frequency, Hz</th>
<th>125</th>
<th>250</th>
<th>500</th>
<th>1000</th>
<th>2000</th>
<th>4000</th>
</tr>
</thead>
<tbody>
<tr>
<td>( L_{ref,ears} ), dB SPL</td>
<td>58.0</td>
<td>69.1</td>
<td>73.5</td>
<td>71.7</td>
<td>69.0</td>
<td>63.0</td>
</tr>
</tbody>
</table>

Table 2.3: Typical speech levels (SPL) at the eardrum \( L_{ref,ears} \) in octave bands
2.4 Room Gain and Voice Support

direct sound level $\tilde{L_D}$ and the overall speech-weighted reflected sound level $\tilde{L_R}$ are

\[
\tilde{L_D} = 10 \log \left( \sum_{i=1}^{6} 10^{\frac{L_{ref,ears,i}}{10}} \right) \text{[dB]} \quad (2.12)
\]

\[
\tilde{L_R} = 10 \log \left( \sum_{i=1}^{6} 10^{\frac{L_{ref,ears,i} + STV,i}{10}} \right) \text{[dB]}, \quad (2.13)
\]

from which the overall speech-weighted voice support $\tilde{ST}_V$ (or simply, voice support) is finally calculated as

\[
\tilde{ST}_V = \tilde{L_R} - \tilde{L_D} = 10 \log \left( \frac{\sum_{i=1}^{6} 10^{L_{ref,ears,i} + STV,i}}{\sum_{i=1}^{6} 10^{L_{ref,ears,i}}} \right) \text{[dB]} . \quad (2.14)
\]

This process is illustrated and summarized in the block diagram in figure 2.6.

Finally, the overall room gain is calculated from the overall speech-weighted voice support by using Eq. (2.3).

2.4.3 Bias factors affecting voice support and room gain

One of the potential drawbacks of the voice support and the room gain is that they are equipment-dependent. Therefore, the measured values of these parameters reported in
2. Interaction between room acoustics and the voice of a speaker

the papers A–E and G are likely to vary if measured with different equipment. Sources of bias are indicated in figure 2.7.

Airborne direct sound  The propagation of the airborne direct sound between the mouth and the ears (figure 2.7a) depends on the distance between them, the diffraction around the head (affected by its geometry), the multiple reflections and scatter at the pinna, and the acoustic response of the ear canal. Additionally, the design of the torso can affect the direct sound, due to scatter and diffraction at the edges. For the HaTS B&K type 4128, the difference $\Delta L_D$ between the SPL at the ears and the on-axis free-field SPL at 1 m in octave bands, when pink noise is reproduced through its mouth in an anechoic chamber, is shown in table 2.4.

<table>
<thead>
<tr>
<th>Center frequency, Hz</th>
<th>125</th>
<th>250</th>
<th>500</th>
<th>1000</th>
<th>2000</th>
<th>4000</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta L_D$, dB</td>
<td>13.1</td>
<td>11.8</td>
<td>11.7</td>
<td>13.5</td>
<td>15.3</td>
<td>14.1</td>
</tr>
</tbody>
</table>

Radiation characteristics  The radiation characteristics of a source are important because they determine the relative levels of the first reflections as a function of the angle of emission. This means that the early reflections at the boundaries in the directions of maximum radiation will have more weight than the early reflections at the boundaries in the directions of minimum radiation. In this perspective, the voice support will be affected by the orientation and the radiation characteristics of the source. Ideally, the
2.4 Room Gain and Voice Support

directivity pattern of the source should be similar to that of a human speaker. Measurements by Chu and Warnock [17] show that the directivity characteristic of the HaTS B&K type 4128 is fairly similar to that of an average human speaker (shown in figure 2.7b).

**Head-related transfer function** The head-related transfer function (HRTF) indicates the increase in sound pressure level that is obtained when measuring at the ears of a human-like receiver instead of measuring with an omnidirectional microphone at the center of the head (in an undisturbed sound field, without the head and the torso). The increase is due to the effect of the head, the torso, the pinna, and the ear canal. The HRTF depends on the angle of incidence of the sound, but manufacturers usually provide the diffuse-field HRTF, which is an average of HRTFs over all the possible directions of incidence (see figure 2.7c). The diffuse-field HRTF $\Delta L_{\text{HRTF}}$ for the HaTS B&K type 4128 is shown in table 2.5.

<table>
<thead>
<tr>
<th>Center frequency, Hz</th>
<th>125</th>
<th>250</th>
<th>500</th>
<th>1000</th>
<th>2000</th>
<th>4000</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta L_{\text{HRTF}}$, dB</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>11</td>
<td>13</td>
</tr>
</tbody>
</table>

2.4.4 Prediction model

A prediction model for the average voice support in a room is presented in paper D. The model disregards the importance of the surroundings of the speaker in determining the actual voice support at the speaker position and provides a unique value for a room, averaged across positions. Nevertheless, a model for the average voice support in a room can be a useful tool during the design process.

The final prediction model (see the complete derivation in paper D) is formulated as

$$ST_V = 10 \log \left[ \left( \frac{cT}{6V \ln 10} - \frac{4}{S} + \frac{Q^*}{4\pi(2d)^2} \right) S_{\text{ref}} \right] + \Delta L_{\text{HRTF}} - K \ [\text{dB}] \quad (2.15)$$

The symbols in this equation are the following:
2. Interaction between room acoustics and the voice of a speaker

\[ c \] Speed of sound in the air (\( \approx 343 \text{ m/s} \))
\[ T \] Reverberation time (in s)
\[ V \] Volume (in \( \text{m}^3 \))
\[ S \] Total surface area (in \( \text{m}^2 \))
\[ Q^* \] Directivity of the source in the downward direction
\[ d \] Distance from the mouth to the floor (\( = 1.5 \text{ m} \))
\[ S_{\text{ref}} \] Reference area (\( \approx 1 \text{m}^2 \))
\[ \Delta L_{\text{HRTF}} \] Diffuse-field HRTF (in dB)
\[ K \] Difference between sound pressure level at the eardrum and source sound power level (in dB)

This model contains the following terms:

- Diffuse-field attenuation of sound, indicated by the term \( [(cT)/(6V \ln 10) − 4/S] \) inside the \( 10 \log \), which is written sometimes as \( 4/R \) in the room acoustics literature (where \( R \) is the so-called room constant, a corrected version of the total absorption area).

- Floor reflection, given by the term \( Q^*/[4\pi(2d)^2] \) inside the \( 10 \log \). The floor reflection is considered present in all measurements, and it is assumed that the floor is totally reflective and that the mouth and the ears are at a height of 1.5 m above the floor. All the early reflections from the walls, when averaged across positions in a room, are included in the diffuse-field attenuation term. The reflection from the ceiling is included in the diffuse-field attenuation term because it is attenuated by the typical presence of an absorbing ceiling in classrooms and because the height varies across rooms.

- Diffuse-field HRTF (\( \Delta L_{\text{HRTF}} \)), accounting for the increase in level associated to the use of a dummy head instead of a small microphone for the measurement of the sound reflections.

- Direct sound characterization with the term \(-K\).

The average voice support does not depend on the orientation of the room and is less sensitive to the radiation characteristics than the local voice support obtained with measurements. For the prediction of local values, Olesen [79, pp.17–19] suggests a method to calculate voice support from computerized room acoustic simulations.
2.5 Reverberation time at the ears

The dependence of the voice support with the volume and the reverberation time of a room is illustrated in figure 2.8, considering a flat reverberation time across frequency. The voice support decreases almost linearly with the logarithm of the volume (except for the largest volumes at low reverberation times) and increases with the reverberation time.

![Figure 2.8: Voice support versus room volume according to the predictions of the model, for different values of reverberation time (labeled on the right), considering a flat reverberation time across frequency](image)

2.5 Reverberation time at the ears

The reverberation time at the ears $T_{30, ears}$ is another magnitude derived from an IR measured between the mouth and the ears of a dummy head. The $T_{30, ears}$ is neither a new concept, nor it is intended to be a new measure of reverberation time, but instead it is a specification of the conditions used to determine the reverberation time. Furthermore, a prediction model showing the dependence of $T_{30, ears}$ with the traditional reverberation time and the room volume is described.

The reverberation time at the ears is used in paper G and section 2.8 to study the perceived acoustic conditions for a speaker.
2. Interaction between room acoustics and the voice of a speaker

2.5.1 Measurement

The reverberation time at the ears $T_{30,\text{ears}}$ is measured from an IR between the mouth and ears of a dummy head, which simulates a human speaker. In this project, the setup of figure 2.4 was used, including a HaTS B&K type 4128 with a left ear simulator type 4159 and a right ear simulator type 4158. The HaTS contained a loudspeaker at its mouth and microphones at the position of the eardrums. The measurement MLS signal was produced with the software dBBATI32 and digitalized with the interface 01dB Symphonie. The IR determined in this way is strongly influenced by the direct sound that propagates from the mouth to the ears. An example IR from the mouth to the ears (in logarithmic scale) is shown in figure 2.9.

![Figure 2.9: Example of an impulse response measured between the mouth and the ears of a dummy head (in gray). The corresponding backwards integrated decay curve is shown as a solid line. The reverberation time at ears $T_{30,\text{ears}}$ is defined as twice the time between the decays at -5 dB and -35 dB and is indicated with the dashed line. A more representative measure of the “traditional” reverberation time $T_{20}$ is obtained by evaluating the slope of the decay curve in the decay from -25 to -45 dB (shown with a dash-dot line).](image)

The $T_{30,\text{ears}}$ is defined as twice the time that it takes for the backwards-integrated energy curve of the IR measured between the mouth and the ears to decay from -5 to -35 dB. Figure 2.9 illustrates the procedure used to evaluate the $T_{30,\text{ears}}$, as compared to a measure of the “traditional”, or “diffuse-field”, reverberation time $T_{20}$, which ignores the effect of the direct sound. The reverberation time at the ears is particularly different...
2.5 Reverberation time at the ears

from the traditional reverberation time because the first one is very sensitive to the ratio between direct sound and reflected sound, whereas the second one is independent of it. However, the reverberation time at the ears does not have any physical meaning.

With the present definition of $T_{30, \text{ears}}$, there is a clear dependence on the equipment used. As happens with the room gain and the voice support, the bias factors are not in the definition of the parameters, but on the method to acquire the IRs between the mouth and the ears. One must be aware of the existing bias factors described in section 2.4.3: the reverberation time at the ears is affected by the path of the direct sound, the directivity of the sound source, and the diffuse-field response of the dummy head.

2.5.2 Prediction

Paper G also presents in detail a prediction model for the average $T_{30, \text{ears}}$ in rooms, as a function of the volume and the diffuse-field reverberation time. This model includes the same elements as the prediction model for the voice support but with temporal considerations. As a difference, the prediction model for $T_{30, \text{ears}}$ does not have a closed mathematical expression, and has to be calculated by means of an algorithm that has the following steps:

1. Modeling of a parametric IR from the mouth to the ears
2. Calculation of the backward integrated energy curve
3. Search the time instants where the backward integrated energy curve decays -5 dB and -35 dB relative to the level at the time of arrival of the direct sound
4. Finally, the $T_{30, \text{ears}}$ is calculated as twice the absolute value of the difference between the two time instants found in the previous step

The prediction model for the average $T_{30, \text{ears}}$, analogously to the prediction model for $STV$, assumes an IR from the mouth to the ears of a dummy head with the following components: direct sound, a floor reflection, and a reverberation tail. These components are illustrated in figure 2.10 in the form of an energy density time curve. The direct sound and the floor reflection are modeled as Dirac delta functions and the reverberation tail as a decaying exponential function.

Assuming this parametric energy density time curve, in which the amplitude and decay constants of the reverberation tail vary with the volume of the room and the
2. Interaction between room acoustics and the voice of a speaker

![Diagram of room acoustics](image)

Figure 2.10: Energy density time curve assumed for the prediction of average $T_{30, ears}$, showing the main components in the airborne acoustic path between the mouth and the ears: the direct sound, the floor reflection, and the reverberation tail.

Reverberation time, the $T_{30, ears}$ is calculated identically as in measurements (section 2.5.1). For this, the backward integrated energy curve is firstly found. Then, the time instants where the backward integrated energy curve decays -5 dB and -35 dB relative to the level at the time of arrival of the direct sound are found, and lastly, the $T_{30, ears}$ is calculated as twice the absolute value of the difference between these two time instants.

Figure 2.11 shows the output of the prediction model for different values of volume and reverberation time. The predicted $T_{30, ears}$ decays with the volume of the room and increases with the reverberation time.

![Graph of reverberation time vs. volume](image)

Figure 2.11: Reverberation time at ears versus volume according to the predictions of the model, for different values of diffuse-field reverberation time.
2.6 Vocal intensity under different room acoustic conditions: the room effect

Brunskog et al. [16] studied the voice power levels used by speakers in rooms of different acoustic conditions, including an anechoic chamber and normal teaching rooms ranging from a small meeting room to a large lecture hall. They noted that speakers adjusted their voices according to the room gain at the position of the speaker. However, there were some incorrect measurements of room gain, which were corrected in paper A using a different method than Brunskog et al. [16]. These corrected values were used together with the original voice power level measurements to propose two simple linear regression models. The first one (see figure 2.12a) describes the variations in voice power level $\Delta L_{W}$—relative to the voice power level in the anechoic chamber for each of the subjects—as a function of the room gain

$$\Delta L_{W} = 0.5 - 13.5G_{RG} \text{ [dB]}.$$  \hfill (2.16)

This model can also be represented as a function of the voice support, showing a nonlinear relationship (dotted line in figure 2.12b). A simplification of this model contains an asymptote for very low values of voice support (dashed line in figure 2.12b) and a linear relationship between the variations in voice power level and the voice support,
but excluding the anechoic chamber (solid line in figure 2.12b). This linear relationship is at focus in the second model, 

\[ \Delta L_W = -13 - 0.78ST_V \text{ [dB]}, \]

which is only valid in 'typical' rooms in a limited range of voice support, approximately between -18 dB and -8 dB.

The term room effect refers to the variations in vocal intensity as a function of the acoustics of the room, represented by the room gain or the voice support. It is specifically the slope of the linear relationship between the two magnitudes. In the first model, the room effect is \(-13.5 \text{ dB} dB_G\) (dB of voice power level for dB of room gain, indicated with the subscript \(G\)). In the second model, the room effect is \(-0.78 \text{ dB} dB_S\) (dB of voice power level for dB of voice support, indicated with the subscript \(S\)). The two versions of room effect describe indeed the same effect, but have different scales and apply to different ranges of acoustic conditions. The use of one or another depends on the conditions tested. If there is an exceptionally damped room (e.g., an anechoic room) among the conditions, the room gain is more suitable than the voice support to define the room effect, because the precise value of the voice support measure is highly unimportant in a very damped room (i.e., it does not make a difference for a speaker whether a room has a \(ST_V\) of -25 dB or -35 dB).

### 2.6.1 Distance factor

Paper B argued that the room gain is correlated to the volume, as small rooms tend to have high room gain and large rooms tend to have lower room gain values, and that the volume is correlated to the average distance of the audience in rooms. The importance of the communication distance on the voice power level has been presented in section 2.2. The goal of paper B was to determine whether there was an effect of the acoustic condition on the voice power level variations of paper A, or these variations could be explained only with the changes in distance. For that matter, 13 male speakers had to describe the contents of a map [2] to a listener located at the distances of 1.5, 3, 6, and 12 m, and they repeated the operation in four acoustically different rooms: an anechoic room, a lecture hall, a reverberation room, and a long corridor. The voice power levels were calculated from the recordings of a small head-worn microphone that the speakers wore.
2.6 Vocal intensity under different room acoustic conditions: the room effect

Figure 2.13a shows that voice power levels used by speakers increased linearly with the logarithm of the distance to the listener, and that the slopes varied between the least steep in the reverberation room (1.3 dB per double distance) and the steepest in the anechoic room (2.2 dB per double distance). Furthermore, the voice power level in the anechoic room was significantly higher than in the other rooms at all distances.

![Figure 2.13a: Average voice power levels obtained in paper B, plotted as a function of the communication distance for different environments.](image)

Figure 2.13b shows the voice power levels as a function of the room gain in the rooms, for different distances. A linear model was fit to the values obtained at each distance, obtaining a total of four linear models. The slopes of these linear models (room effect) were -1.6 dB/dB\(_G\) at 1.5 m (orange), -2.6 dB/dB\(_G\) at 3 m (red), -3.6 dB/dB\(_G\) at 6 m (blue), and -3.7 dB/dB\(_G\) at 12 m (green). If there was no room effect, the lines would be horizontal and parallel to each other. Nevertheless, there is a room effect and one of its characteristics is that its importance increases with the average distance to the listeners.

An alternative analysis in figure 2.13b shows the voice power levels as a function of the room gain in the rooms, for different distances. A linear model was fit to the values obtained at each distance, obtaining a total of four linear models. The slopes of these linear models (room effect) were -1.6 dB/dB\(_G\) at 1.5 m (orange), -2.6 dB/dB\(_G\) at 3 m (red), -3.6 dB/dB\(_G\) at 6 m (blue), and -3.7 dB/dB\(_G\) at 12 m (green). If there was no room effect, the lines would be horizontal and parallel to each other. Nevertheless, there is a room effect and one of its characteristics is that its importance increases with the average distance to the listeners.

On the same figure 2.13b, the results of Brunskog et al. [16]—as presented in paper A—are shown (in black). The voice power level measurements of Brunskog et al. [16] contained the combined effect of distance variation and room acoustic quality in the measure room gain. The intersection of the black regression line (Brunskog et al. [16], variable distances to the audience) with the other color lines (fixed distance to

---

\(^3\) In paper B, only the regression line for a distance of 6 m is shown. However, the regression lines at the different distances are relevant and this figure is included in the introduction of paper G.
2. Interaction between room acoustics and the voice of a speaker

the listener) occur at different distances. At 12 m, the intersection occurs at \( G_{RG} = 0.17 \) dB, and increases up to \( G_{RG} = 0.68 \) dB at 1.5 m. These variable intersections show how the average distance to the listeners and the room gain are interrelated. A possible retrospective interpretation of the results in Brunskog et al. [16] is that the average distance to the audience was high for low \( G_{RG} \) values and decreased for high values of \( G_{RG} \).

Although the room effect is measurable on average, it is highly variable across individuals. Figure 2.14 shows a box-plot of the room effects calculated on an individual basis (i.e., calculating the lines on figure 2.13b for each individual), to illustrate the spread of this magnitude. It can be seen that the room effect becomes more negative with increasing distance and that only at 12 m it might have lower spread than at shorter distances.

![Box-plot of room effects](image)

Figure 2.14: Comparison of the room effects (related to room gain) obtained at the different communication distances in paper B

2.6.2 Equal autophonic level curves

Sections 2.2.2 and 2.3 argued that room acoustic conditions can modify the auditory feedback of one’s own voice and influence the vocal intensity used for speaking. Additionally, the results of paper B showed that speakers modify their vocal intensity with the room gain. Paper C was aimed at analyzing whether the room effect of paper B was
2.6 Vocal intensity under different room acoustic conditions: the room effect

an adaption to changes in auditory feedback of the same kind as the Lombard effect (section 2.2) or the sidetone compensation (section 2.3).

Paper C determined the variations in voice level that kept the loudness level of one’s own voice (autophonic levels) constant under different room acoustic conditions, characterized with the room gain or the voice support. Subjects were presented a reference signal of short duration at a fixed SPL and were asked to produce a vocalization that evoked the same loudness sensation as the reference signal. This was repeated under ten different acoustic conditions applied to the voice of the speaker (but not for the reference signal). Speakers performed—linking terms to previous research—a sidetone compensation, being the room acoustic condition the source of sidetone alteration. The acoustic conditions were produced artificially in laboratory with a real-time convolution system.

The average variations in non-weighted SPL \( \Delta L_Z \) and A-weighted SPL \( \Delta L_A \) at the microphone position (taking as a reference the SPL without a simulated room condition) that kept the autophonic level constant under different room gain conditions were

\[
\Delta L_Z = 8.4 \times e^{-0.24G_{RG}} - 8.9 \ [\text{dB}], \quad (2.18a)
\]
\[
\Delta L_A = 6.4 \times e^{-0.25G_{RG}} - 6.9 \ [\text{dB}] \quad (2.18b)
\]

or alternatively, as a function of the voice support,

\[
\Delta L_Z = 8.4 \times \left(10^{STV_{10}} + 1\right)^{-1.05} - 8.9 \ [\text{dB}], \quad (2.19a)
\]
\[
\Delta L_A = 6.4 \times \left(10^{STV_{10}} + 1\right)^{-1.10} - 6.9 \ [\text{dB}] \quad (2.19b)
\]

These curves, and similar curves for the voice level variations in octave bands, are shown in figure 2.15 (the plot at the left shows the voice level variations as a function of the room gain and the plot at the right shows the voice level variations as a function of the voice support).

For a range of room gain between 0 and 0.8 dB, as measured in paper B, Eq. (2.18a) predicts a voice level variation of -1.46 dB, which corresponds to an average room effect of -1.8 dB/dB\( _{RG} \). This value is very similar to the room effect of -1.6 dB/dB\( _{RG} \) for talkers speaking to listeners at 1.5 m (from experiments in paper B, or yellow curve in figure...
2. Interaction between room acoustics and the voice of a speaker

2.13b). Therefore, it is reasonable to affirm that talkers speaking to listeners at a short distance adjust their voice to hear themselves equally loud in different environments.

At further talker-to-listener distances, the room effect is not uniquely explained by the sidetone compensation, and in the rooms with low room gain speakers feel compelled to raise their voices more than in the rooms with high room gain, most likely because speakers want to compensate for a higher attenuation of sound with distance in the rooms with low room gain. Thus, speakers use acoustic cues other than loudness to adjust their voice under different room acoustic conditions.

2.6.3 Field study: interaction of classroom acoustics and teachers’ voice health

Two groups of teachers, one of 13 teachers with voice problems (test group) and another one of 14 teachers with healthy voices (control group), were selected for a field study, which is described in paper E. The study analyzed the reactions of the two groups to classrooms of different acoustic conditions while teaching. The teachers were initially selected from a questionnaire study [66] and were assigned into the test or the control groups according to their rating of the question “I have voice problems”. Teachers in the test group rated experiencing voice problems sometimes, often, or always. Teachers in the control group rated experiencing voice problems never or only occasionally. A later study [65] showed that the two groups did not differ in objective measurable features,
2.6 Vocal intensity under different room acoustic conditions: the room effect

but differed on their ratings to the VHI-T questionnaire [64] and on the fact that the teachers with voice problems reported significantly longer times for vocal recovery.

During one working day, teachers were equipped with a sound level meter which had a lapel microphone positioned at about 15 cm from their mouth and determined the sound pressure level at their position while teaching. Using statistical methods, the sound pressure level corresponding to the teacher (voice level) and to the activity noise (noise level) were estimated separately. The estimated sound pressure level corresponded to the statistical level $L_{50}$ in one lesson, i.e., the level that was exceeded 50% of the time, which was noted as $L_{50,S}$ for the voice level and $L_{50,N}$ for the noise. In addition, objective acoustic parameters in a total of 30 classrooms where the teachers had been teaching were measured and reported in paper D. These parameters included the physical dimensions of the room, the background noise levels in the empty rooms, the reverberation time, the speech transmission index (STI), the room gain, and the voice support. Of these parameters, only the voice support showed a significant correlation with the voice level measurements during teaching.

The voice levels of teachers in the test and control groups were described with the multiple regression models dependent on the noise level and the voice support:

$$L_{50,S}(\text{test}) = 81.3 - 3.87 \times \sqrt{75 - L_{50,N}} - 0.72 \times ST_V \ [\text{dB}],$$  \hspace{1cm} (2.20a)

$$L_{50,S}(\text{control}) = 102.9 - 3.87 \times \sqrt{75 - L_{50,N}} + 0.84 \times ST_V \ [\text{dB}].$$  \hspace{1cm} (2.20b)

For the average measured voice support in the classrooms (-13 dB), the model in Eq. (2.20) reduces to

$$L_{50,S}(\text{test}) = 90.6 - 3.87 \times \sqrt{75 - L_{50,N}} \ [\text{dB}],$$  \hspace{1cm} (2.21a)

$$L_{50,S}(\text{control}) = 92.0 - 3.87 \times \sqrt{75 - L_{50,N}} \ [\text{dB}].$$  \hspace{1cm} (2.21b)

This model is shown in figure 2.16a together with the individual measured values of voice level and noise level. The plot shows identical responses of the teachers in the two groups toward noise, following the Lombard effect, although teachers in the control group used non-significantly higher voice levels than in the test group. The increase in voice level as the noise level increases becomes higher at high noise levels, and it is less important at low noise levels, as indicated by Lazarus [57]. The average slope for noise
levels between 55 and 75 dB is 0.86 dB/dB, which is similar to the 1 dB/dB reported by Pearson et al. [81] for teachers at work.

![Graph](image)

Figure 2.16: (a) Voice levels used by teachers versus student-activity noise levels. (b) Voice levels used by teachers versus voice support measured in the empty classrooms. The solid lines show the regression models.

For the average noise level (≈ 66 dB), the model for the voice levels in Eq. (2.20), reduces to

\[
L_{50,S}^{(\text{test})} = 69.8 - 0.72 \times ST_V \text{ [dB]}, \tag{2.22a}
\]

\[
L_{50,S}^{(\text{control})} = 91.4 + 0.84 \times ST_V \text{ [dB]}, \tag{2.22b}
\]

which depends only on the voice support. These regression lines are shown in figure 2.16b together with the points corresponding to individual measurements of voice level at rooms with particular voice support values. Despite the scattered values, the difference in slopes between the two groups was significant. According to Eq. (2.22), the room effect for the test group was \(-0.72 \text{ dB/dB}_S\), a result which is close to the \(-0.78 \text{ dB/dB}_S\) reported in paper A (section 2.6). This means that teachers with voice problems softened their voice level in more supportive rooms. However, the room effect for the control group was \(+0.84 \text{ dB/dB}_S\), meaning that teachers even raised their voice levels with increasing voice support in the rooms. This effect might be due to a possible increase in activity noise levels with increasing voice support, which in turn makes teachers raise their voice (derivations in Appendix A used this hypothesis, based on observations by Hodgson et al. [33]). Nevertheless, this hypothesis is not proved in the measurements of paper E, because there were no indications of correlation between
2.6 Vocal intensity under different room acoustic conditions: the room effect

noise level and voice support, as shown in figure 2.17. However, the number of students in each classroom was not controlled.

![Scatter plot of activity noise levels versus voice support measured in the classrooms of paper E](image)

Figure 2.17: Scatter plot of activity noise levels versus voice support measured in the classrooms of paper E

The different reaction of teachers with and without voice problems to the voice support might indicate a higher sensitivity of teachers with voice problems toward their working environment, who may lower their voice as an adaptive mechanism to preserve their vocal health. In the same field study (as described in [63]), teachers wore a skin accelerometer glued on their neck, which calculated the fundamental frequency, the vocal intensity, and the phonation time. There were significant differences in the way that the two groups adjusted their vocal intensity as a function of the fundamental frequency. On the one hand, teachers with healthy voices raised their fundamental frequency as they increased their vocal intensity, which is said to be a natural reaction to cope with vocal loading. On the other hand, teachers with voice problems lowered their fundamental frequency with increased vocal intensity, suggesting a reduced vocal flexibility.

The room effects derived from Eq. (2.22) have to be assessed with caution, because in most of the cases, there were only two samples per teacher in the same environment. Therefore, the reaction of individuals to classrooms with different acoustic conditions is largely unknown.
2. Interaction between room acoustics and the voice of a speaker

2.6.4 Laboratory experiments

Methodological aspects

There are physical limitations to test the influence of room acoustic conditions on the voice of a speaker. The researcher must find rooms with different acoustic conditions that are located close to each other, and this is not always possible. For this purpose, paper F describes a laboratory setup that was especially designed to emulate the acoustic conditions of different rooms perceived by a speaker without having the visual influence of the room.

As explained in section 2.3, a person can hear his own voice by body conduction, direct airborne transmission, or through reflections at the environment boundaries. The laboratory setup explained in paper F aimed at preserving the body conduction and the direct airborne transmission of the voice of the speaker, minimize the reflections produced by the actual laboratory room, and simulate the reflections that would occur in another room (virtual room). Figure 2.18 shows a simplified representation of the laboratory setup. The voice of a speaker located in a damped room was picked with a head-worn microphone, digitalized and sent to computer system, which applied an equalizer filter—that adjusted the spectral balance of the speech signal to match that of the speech in front of the talker—and the acoustic effect of the virtual room (i.e., the room impulse response) by means of convolution. The resulting signal was amplified and reproduced through the 29 loudspeakers. The talker in the room perceived the signal played back from the loudspeakers as if it were the reflections of his own voice in a room different from the laboratory room. The acoustics of the virtual room were calculated beforehand (at the left on figure 2.18), starting from a computer model of the virtual room, which was loaded in an acoustic simulation software to extract the information about the reflection paths between the mouth and the ears. The information about the reflection patterns was processed with the LoRA (Loudspeaker-based Room Auralization system) toolbox of Favrot and Buchholz [23], which produced a room impulse response output suitable for the actual loudspeaker reproduction layout in the form of 29 WAV files containing impulse responses.

The body conduction and the direct airborne transmission of the voice of the talker were preserved due to the use of distant loudspeakers while the installed sound absorbing materials in the room attenuated and minimized the effect of the actual room reflections. A picture of the actual laboratory room is shown in figure 2.19.
2.6 Vocal intensity under different room acoustic conditions: the room effect

Figure 2.18: Block diagram of the laboratory setup used to emulate the acoustic conditions of different rooms.

Figure 2.19: Picture of the laboratory facility designed for emulating the acoustic conditions for a talker seating in the middle.
44 2. Interaction between room acoustics and the voice of a speaker

Table 2.6: Summary of the experiments, identifying the group of subjects and its number \( N_S \), the number of conditions \( N_C \), whether there was a questionnaire, the kind of instruction used, and the technical setup.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Subjects</th>
<th>( N_S )</th>
<th>( N_C )</th>
<th>Quest.</th>
<th>Instruction</th>
<th>Setup</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pilot PRE1</td>
<td>Mixed teachers-students</td>
<td>5</td>
<td>5</td>
<td>NO</td>
<td>Simulated lecture</td>
<td>Loudspeaker</td>
</tr>
<tr>
<td>Pilot PRE2</td>
<td>Teachers</td>
<td>5</td>
<td>10</td>
<td>NO</td>
<td>Simulated lecture</td>
<td>Loudspeaker</td>
</tr>
<tr>
<td>A</td>
<td>Teachers</td>
<td>13</td>
<td>10</td>
<td>YES</td>
<td>Reading a text</td>
<td>Loudspeaker</td>
</tr>
<tr>
<td>A2</td>
<td>Students</td>
<td>13</td>
<td>10</td>
<td>YES</td>
<td>Reading a text</td>
<td>Loudspeaker</td>
</tr>
<tr>
<td>B</td>
<td>B1 Teachers (healthy voice)</td>
<td>11</td>
<td>10</td>
<td>YES</td>
<td>Describing a map</td>
<td>Earphones</td>
</tr>
<tr>
<td>B2</td>
<td>Teachers (non-healthy voice)</td>
<td>10</td>
<td>10</td>
<td>YES</td>
<td>Describing a map</td>
<td>Earphones</td>
</tr>
</tbody>
</table>

Paper G reports the actual laboratory experiments carried out with the setup described in paper F. There were a total of four experiments, which are summarized in table 2.6. There were two pilot experiments (named PRE1 and PRE2), a experiment (named A) to test the differences between experienced speakers (teachers, group A1) and unexperienced speakers (students, group A2), and a experiment (named B) to test the differences between the teachers with healthy voices (group B1) and with voice problems (group B2) that had participated in the field study reported in section 2.6.3 and paper E.

The acoustic conditions in experiments PRE1, A, and B corresponded to simulated classrooms of different size, with different reverberation times and different absorption treatments. In this way, the simulated acoustic conditions corresponded to realistic scenarios. The acoustic conditions in experiment PRE2 were obtained from a single impulse response by modifying the gain applied at the output, after the convolution. Thus, the reverberation time (or decay rate of the IR) was fixed, but the voice support was varied. Moreover, impulse responses generated in this way do not correspond to any existing physical space.

The instruction varied across experiments. In the pilot experiments PRE1 and PRE2, subjects were requested to give a lecture, prepared beforehand, about a familiar topic to an imaginary group of 30 students. Speakers could use the most comfortable language for them. In experiment A, speakers were asked to read an English text aloud (Goldilocks’ passage [100]) for a “listener” at a distance of 2 m, indicated with a dummy head at that particular position. Most of the subjects were non-native English speakers. In experiment B, speakers had to describe the elements of a map [2] in
2.6 Vocal intensity under different room acoustic conditions: the room effect

Swedish, their native language, to a listener—who did not understand Swedish—seating at 3 m in front of them.

In the pilot experiments and experiment A, the setup of figure 2.18 at the Technical University of Denmark was used, whereas experiment B was performed at Lund University with a portable setup described in paper C. The portable setup had the same functionality as the setup of figure 2.18 but used earphones—especially designed to minimize changes in body-conducted and airborne direct sound of one’s own voice—instead of loudspeakers.

In experiments A and B, questionnaires were handed in. These questionnaires had the aim of investigating the subjective preference and the impressions that talking in a particular acoustic environment produced on a speaker. The results of the questionnaire study are reported in section 2.8.

Results: performance in laboratory

The laboratory experiments showed that speakers lowered their voice level as a function of the voice support. Figure 2.20 shows the regression lines (solid lines, with the confidence intervals shown as dashed curves) of the linear models describing the voice level as a function of the voice support. The voice level is expressed as on-axis, free-field SPL at 1 m in front of the speaker.

Figure 2.20: Average regression lines for voice levels as a function of the voice support (solid lines). The dashed curves indicate the confidence intervals considering a simple linear regression model with all the measurements.

Figure 2.21 summarizes in a box plot the room effects calculated for different individuals in different experimental groups; i.e., the slopes in figure 2.20 that would be obtained with the measurement data of each individual.

The average room effect in pilot experiment PRE1 was -0.89 dB/dB, whereas in experiment PRE2 was -0.96 dB/dB. These two values were not significantly different
2. Interaction between room acoustics and the voice of a speaker

![Figure 2.21: Comparison of the room effects (related to voice support) obtained in the different experimental groups and methods in laboratory](image)

from each other and are relatively close to the room effect of -0.78 dB/dB_S of paper A. This suggests that speakers imagined being talking in rooms of different sizes and addressing an audience of 30 students distributed through the floor area of the imagined room.

The average room effect for teachers in group A1 was -0.35 dB/dB_S while reading a text, whereas for students, it was -0.11 dB/dB_S. The difference between the groups was not significant, although there were a few outliers (see figure 2.21) that indicated that some subjects could react strongly to changes in voice support.

The average room effect for voice-healthy teachers in group B1 was -0.12 dB/dB_S, whereas for teachers with voice problems it was -0.07 dB/dB_S. The room effect had a much lower variance than in the pilot experiments and experiment A, suggesting that the map task was better controlled than the other tasks. The difference between the two groups was not significant, but it was indeed very different from the performance of the two groups in the field study (section 2.6.3 and paper E).

The equal autophonic level curve of Eq. (2.19a) predicts an average room effect of -0.1 dB/dB_S for values of voice support between -23 and -6 dB, which is the range of voice support tested in experiment B. The value of -0.1 dB/dB_S is very close to the results of groups B1 and B2, suggesting that teachers in both groups talked to hear themselves equally loud.
2.7 Other changes in voice production

Although the main body of the work analyzed the variation of vocal intensity as a function of voice support in rooms, other speech parameters do vary too, for example, the average speech spectrum, the mean fundamental frequency (F0), the long-term standard deviation of the fundamental frequency (F0 SD), and the phonation time.

In paper B, there was an increase of F0 with the distance from the speaker to the listener and F0 in the anechoic chamber was significantly higher than in the other environments. F0 SD also increased with the distance between speaker and listener, and varied across environments: F0 SD was highest in the anechoic room and lowest in the reverberation room. The phonation time increased slightly with distance but changed significantly across environment. The phonation time was 10% higher in the anechoic room and the reverberation room (the two least comfortable rooms for speaking) than in the corridor and the lecture room (the two most comfortable rooms for speaking in paper B). This last observation suggested that an increase in phonation time is either a side-effect of increased vocal intensity or a way to enhance speech intelligibility.

Paper C reported some of the spectral changes that occurred in vocalizations when subjects were asked to produce a vocalization of the same loudness as a reference signal. The main changes induced by the acoustic environment occurred at high frequencies. It was observed that, when the spectra of the vocalizations under the different acoustic environments were compensated for the spectral auditory changes induced in sidetone (i.e., the room gain), then the spectra turned out to be similar. I.e., it appears that speakers keep the voice quality at their ears constant when they are asked to keep the loudness constant. Nevertheless, the actual spectral changes in the natural voice adjustment under different room acoustic conditions were not measured.

The parameters F0, F0 SD, and phonation time were analyzed in the laboratory experiments in a way analogous to the vocal intensity. For each parameter, the results of one subject at each condition were used to fit a linear model. This model indicated the average trend of the parameter with the voice support for that particular subject. The slopes for different individuals in the simple linear models for the parameters F0, F0 SD, and phonation time as a function of the voice support are shown in figures 2.22, 2.23, and 2.24, respectively.

The slopes for F0 as a function of voice support in figure 2.22 have a correlation to the room effects shown in figure 2.21, as would be expected from the natural covariance
of vocal intensity and fundamental frequency in normal speech [109, pp.243–280]. As in the room effect, the slopes for the pilot experiment PRE2 showed the largest variation of F0 with voice support.

![Box-plot with the slopes of the individual linear regressions of mean F0 versus voice support obtained in the different experimental groups and methods in laboratory](image)

**Figure 2.22**: Box-plot with the slopes of the individual linear regressions of mean F0 versus voice support obtained in the different experimental groups and methods in laboratory

The slopes for F0 SD as a function of voice support in figure 2.23 are not significantly different from 0; therefore, it is likely that F0 SD does not change at all with voice support under laboratory conditions.

![Box-plot with the slopes of the individual linear regressions of F0 SD versus voice support obtained in the different experimental groups and methods in laboratory](image)

**Figure 2.23**: Box-plot with the slopes of the individual linear regressions of F0 SD versus voice support obtained in the different experimental groups and methods in laboratory

More revealing, though, are the slopes of phonation time as a function of voice support, shown in figure 2.24. In experiments A and B, the average slopes were not significantly different from 0 and therefore, speakers did not vary the phonation time with the voice support. However, in the pilot tests PRE1 and PRE2, which are the
ones that best recreate a real teaching scenario, speakers increased the phonation time as the voice support increased. Rooms with highest voice support might have been perceived as more reverberant by the speaker. The speaker, knowing about the possible detrimental effects of the reflected sound on speech intelligibility, can decide to increase the duration of vowels (or voiced segments in general) as a means to increase speech intelligibility and compensate for the detrimental effect of reverberation. These results completely agree with the observations in paper B, where speakers increased their phonation time in the reverberation room.

Figure 2.24: Box-plot with the slopes of the individual linear regressions of phonation time versus voice support obtained in the different experimental groups and methods in laboratory

Nevertheless, the significant differences between the room effects measured in laboratory and during real teaching described in previous sections suggest that the voice parameters reported in this section might as well change differently during teaching and under laboratory conditions.

### 2.8 Acoustic comfort for a speaker

When studying the effect of room acoustic conditions on the voice production, it is not only important to determine objective changes, but also to relate how these conditions are perceived by the speaker, specifically regarding the sensation of comfort. The acoustic comfort for a speaker is defined as the overall sensation of well-being transmitted by a room to a speaker through the acoustic feedback of his own voice. A subjective study was performed together with the laboratory experiments and was reported in paper G.

In the experiments A and B described in section 2.6.4 and summarized in table
2. Interaction between room acoustics and the voice of a speaker

2.6, the subjects had to rate a set of questions or statements regarding the experience of talking under a certain acoustic condition:

1. I would feel exhausted if I were talking in this classroom for a whole lesson
2. The classroom is good to speak in
3. The classroom enhances and supports my speech
4. I must raise my voice in order to be heard in the classroom
5. The sound system makes my voice sound unnatural
6. I noticed echo phenomena in the classroom
7. Rate the degree of reverberance that you perceived in the classroom
8. Rate how you perceive your voice now

In questions 1 to 6, the extremes ratings were totally disagree and strongly agree. In question 7, the extremes were very low and very high. Question 8 had extremes no voice problems and extremely severe problems. This last question had the aim of detecting anomalous performance in certain conditions and was not used for further analysis.

The answers to the questions were not independent, but were highly correlated among them, by groups. Answers to questions 1 to 4 were strongly correlated among them and were included in one principal component, which was related to the acoustic comfort for a speaker. The answers to questions 5 to 7 were also correlated among them and were included in a second principal component linked to the sensation of reverberance.

The answers of the questionnaires were analyzed using different acoustic parameters, including the voice support $ST_v$, the reverberation time $T_{20}$, and the reverberation time at the ears $T_{30, ears}$ described in section 2.5.

The reverberation time at the ears $T_{30, ears}$ was of particular importance in the study because it presented the strongest correlation with the subjective impressions of acoustic comfort for a speaker and reverberance. Figure 2.25 shows the answers of the questionnaires (in the form of principal components) as a function of $T_{30, ears}$.

The acoustic comfort for a speaker (top row in figure 2.25) had a non-linear dependence with the $T_{30, ears}$ for speakers with healthy voices. The maximum of comfort was located for $T_{30, ears}$ between 0.45 and 0.55 s, probably because environments with
2.8 Acoustic comfort for a speaker

Figure 2.25: Top row: principal component related to the acoustic comfort for a speaker, as a function of the reverberation time at the ears. Bottom row: principal component related to the reverberance of the room, as a function of the reverberation time at the ears. Each column shows a different experimental group. Individual answers are shown in gray, average values at each condition shown with black dots. Best fitting first or second order polynomials are overlaid on the plots.
2. Interaction between room acoustics and the voice of a speaker

$T_{30,\text{ears}} < 0.45$ were perceived as too dry and environments with $T_{30,\text{ears}} > 0.55$ degraded speech intelligibility. However, for teachers with voice problems, the comfort increased linearly with $T_{30,\text{ears}}$, showing that they felt more comfortable in rooms that amplified their voices.

The sensation of reverberance (bottom row in figure 2.25) increased linearly with $T_{30,\text{ears}}$, with very similar slopes for all the different groups of subjects.

2.9 Summary of findings

The process of voice adjustment in a speaker is a complex mechanism that starts with the intention of generating a voice with certain parameters that would ensure a successful communication, involves motor actions that result in the production of voice, which in turn triggers a series of feedback mechanisms that allow the speaker to continuously monitor the vocal output and to adapt his voice. Some of the factors that influence the intention of the talker of speaking with certain vocal parameters are the distance to the listeners, the background noise, the knowledge of some special requirements about the listener (e.g., hearing, age, or mother tongue) and the intention to use the voice for particular purposes (e.g., to sound authoritative, for instructing, or for singing).

One of the feedback mechanisms influencing the voice adjustment is the audition of one’s own voice perceived through body-conduction, through direct airborne sound propagation, and through the reflections of sound at the environment boundaries. The acoustic conditions of a room determine this last path and affect the auditory feedback available to adjust one’s own voice.

The relative importance of the direct airborne sound and the reflected sound components of one’s own voice are quantified with the objective parameters room gain and voice support defined in paper A (although the room gain was introduced by Brunskog et al. [16]). The room gain is defined as the difference between the total energy level in a room and the energy level of the direct sound, i.e., the gain applied by the room to one’s own voice. The voice support is defined as the difference between the energy level of the reflections and the energy level of the direct sound of one’s own voice.

Brunskog et al. [16] and paper A determined that, under realistic teaching situations with low ambient noise, teachers reduce their voice levels as the room gain increases, at a rate of $-13.5 \, \text{dB/dB}_G$. In an alternative description, the voice levels de-
crease as the voice support increases, at a rate of -0.78 dB/dB_{S}. The variations in voice level with changing room acoustic conditions are referred to as room effect.

Measurements in paper B showed that the room effect is mostly due to the variation of the distance between speaker and listeners, but also found non-zero room effects for speakers addressing listeners at equal distances under different room acoustic conditions. The room effect became stronger as the distance between speaker and listener increased (from -1.6 dB/dB_{G} at 1.5 m to -3.7 dB/dB_{G} at 12 m).

Paper C determined the voice levels that keep the loudness level of one’s own voice (i.e., the autophonic level) constant under different conditions of room gain and voice support. The results in paper C explain, for example, that speakers addressing listeners at short distances adjust their voices to hear themselves equally loud under different room acoustic conditions.

Paper D proposed a prediction model for voice support and validated it through measurements in 30 classrooms, finding average values of voice support between -20 and -5 dB.

Paper E found significant group-wise differences in the way that teachers with and without voice problems react to voice support during teaching, but no differences in their reactions to background noise. Teachers with voice problems lowered their voice with voice support at a rate of -0.72 dB/dB_{S}, whereas teachers without voice problems raised their voices at a rate of +0.84 dB/dB_{S}. This finding suggests that teachers with voice problems are more sensitive to environmental factors than their voice-healthy colleagues.

Paper F described a loudspeaker setup to generate virtual acoustics of rooms in real time, so that speakers have the feeling of being in rooms with acoustic conditions that are different from the physical laboratory room. This setup was used in paper G for different laboratory experiments regarding vocal intensity adjustment and speakers’ comfort under different room acoustic conditions.

The vocal intensity adjustment in laboratory conditions depends critically on the task and the instruction given. For tasks that convey specific requirements of the teaching situation, i.e., addressing a relatively large group of students, the average room effect (≈ -0.9 dB/dB_{S}) is similar to that reported in paper A, supporting the validity of the findings in the latter. However, in other tasks as reading or describing a map, speakers vary much less their voice levels, following a strategy of keeping their autophonic levels constant. It is also observed that the room effect has a large spread across
2. Interaction between room acoustics and the voice of a speaker

individuals, and that some individuals are able to respond to different room acoustic conditions much more strongly than others.

The acoustic comfort for a speaker has an optimum range for values of reverberation time at the ears between 0.45 and 0.55 s. The reverberation time at the ears is the reverberation time derived from the decay between -5 and -35 dB in the backward integrated energy curve of an impulse response measured between the mouth and the ears of a dummy head.
Chapter 3
Implications for classroom acoustics design

Classrooms and educational spaces are places where the learning process takes place, primarily by means of speech communication; i.e., by speaking and listening. The success in communication highly depends on the delivery of a clear and intelligible message throughout the room. At the same time, the acoustic conditions in a classroom have to allow the teacher to speak comfortably and support his voice, so as to avoid the use of excessively high voice levels.

The acoustics of classroom are traditionally designed to optimize speech intelligibility. While this approach is very useful and is commonly applied for the design of classrooms, it lacks parameters that explain how the acoustic design can affect teachers’ voices and their perception of the acoustic environment. Section 3.1 introduces some of the suggested approaches to classroom acoustic design found in the literature, with a focus on the listener.

Section 3.2 uses the knowledge acquired during the project—described in chapter 2—to propose some guidelines for speaker-oriented classroom acoustic design. The recommendations result as a combination of optimizing vocal comfort, minimizing the required vocal effort, and providing high enough speech SPL over the audience area.

Section 3.3 discusses how the requirements for speakers and listeners meet and suggests directions to combine both approaches. Finally, section 3.4 summarizes the recommendations on speaker-oriented classroom acoustic design and the limitations of the approach used to obtain them.
3.1 Listener-oriented classroom acoustics design

Studies that suggest different criteria for classroom acoustic design are commonly based on an optimization of the listening conditions. Bradley [9] suggests that useful-to-detrimental ratios are the best predictors of speech intelligibility in rooms, even better than the Speech Transmission Index (STI) [40]. The useful-to-detrimental ratio, introduced by Lochner and Burger [60] but simplified by Bradley [9], is dependent on three parameters: the speech SPL, the background noise level, and the early/late ratio. Another implicit parameter is the early/late time threshold. Speech intelligibility increases with the signal-to-noise ratio (SNR). It has also been observed that late reverberation degrades speech intelligibility, and that maximum speech intelligibility can be achieved when the reflections in a room arrive mostly in the first 50 ms after the arrival of the direct sound [72].

The listening conditions in schools are more critical than in other rooms for speech because of different factors. First, children are developing their language and cognitive abilities. They need more SNR than adults to achieve the same speech intelligibility scores [98]. There are some indications that the ability of coping with speech in reverberation is not fully developed in children [59]. However, Yang and Bradley [121] argued that the variations in SNR affect negatively children’s ability of understanding speech much more than variations in reverberation time. Second, children acquiring a second language require higher SNR than those who are native speakers of the language [76]. Third, educational spaces serve students who have different disabilities that affect speech perception: ear infections, hearing loss, language learning problems, behavior disorders, reduced cognitive skills. These students have special needs for rooms where speech is clear and intelligible.

Early studies by Bradley [10] and Houtgast [34] found that speech intelligibility improved for A-weighted SNRs up to +15 dB. Later, Bradley and Sato [12] suggested that the SNR that makes 75% of the students achieve 95% of speech intelligibility scores above grade 6 should be +15 dB, but for students in grade 1 the SNR should be +20 dB.

Bradley et al. [13] found that early reflections on the first 50 ms after the arrival of the direct sound amplified the sound without degrading speech intelligibility. Sato and Bradley [93], in a theoretical study, found that the benefit from early reflections increased more than the detrimental effect of late reflections at the lowest values of reverberation time, therefore recommending reverberation times between 0.2 and 0.5 s
3.1 Listener-oriented classroom acoustics design

In a classroom of about 200 m$^3$, in the teaching scenario, a reverberation time of 0 s degrades speech intelligibility [32], because usually the students are the main source of noise.

Bradley [11] suggests that the reverberation in the classroom should be such that early reflections are maximized while keeping the energy of the late reflections as low as possible, reporting optimum reverberation times between 0.5 and 0.7 s under occupied conditions, while acceptable values are between 0.4 and 0.8 s. At the same time, Bradley [11] argues that A-weighted background noise levels should not be higher than 35 dB, as already established in some standards [1, 20].

Recently, Nijs and Rychtarikova [77] proposed a model to calculate useful-to-detrimental ratios as a function of the talker-to-listener distance, from which it is possible to derive optimum reverberation times for given SNRs. As an example, in a classroom of 170 m$^3$, Nijs and Rychtarikova [77] stated that ‘excellent’ speech intelligibility scores could only be obtained for reverberation times below 0.4 s, whereas ‘good’ speech intelligibility could be obtained with reverberation times between 0.4 and 0.6 s.

Whitlock and Dodd [120] found that children between 7 and 9 yr had lower ‘integration times’ for speech than adults (35 ms instead of 50 ms) and suggested that they need lower reverberation times than adults, stating that values of reverberation times not higher than 0.4 s are ideal. Additionally, Whitlock and Dodd [120] argued that low reverberation times help to reduce the “café effect”, as shown by Korn [52].

Some of the current acoustic standards for classroom acoustic design set maximum admissible limits on reverberation time. The American standard ANSI S12-60:2002 (R2009) [1] recommends mid-frequencies reverberation times not higher than 0.6 s in furnished but unoccupied classrooms with volumes up to 283 m$^3$ and 0.7 s in classrooms with larger volumes, up to 566 m$^3$. The British standard Building Bulletin 93 (BB93) [20] sets reverberation time upper limits as a function of the use of the room. For general primary school classrooms, the limit is 0.6 s, whereas it is 0.8 s for secondary school classrooms (for furnished but unoccupied classrooms). The German standard DIN 18041 [21] sets a target mid-frequency reverberation time $T_{soll}$ (soll means target in German) under occupied conditions that depends on the volume $V$ of the room:

$$T_{soll} = 0.32 \log V - 0.17 \text{ [dB]}.$$  \hfill (3.1)

---

$^1$ Even though the reverberation time is not a measure of speech intelligibility, by specifying it together with the volume of the room, it is correlated with speech intelligibility descriptors such as early/late ratios.
3. Implications for classroom acoustics design

If the room is unoccupied, the reverberation time shall not exceed $T_{\text{soil}}$ in more than 0.2 s.

Although there is no explicit study, to the knowledge of the author, which assesses the acoustic conditions for speakers, it has been argued that having to talk at long distances in very damped classrooms might demand from the teachers additional vocal intensity (which is described as overdamping effect [77]). On the other hand, teachers may have to raise their voice levels in more reverberant classrooms to cope with increased noise levels from the students, as a model from Hodgson et al. [33] predicts.

3.2 Speaker-oriented classroom acoustics design

As it was shown in chapter 2, there are two room acoustics-related aspects which appear to be important for a speaker. The first is the room effect (section 2.6), i.e., the variation in voice level motivated by the acoustic conditions of the room, characterized by the voice support (or the room gain). The second is the acoustic comfort offered by the room to a speaker, which depends non-linearly on the reverberation time at the ears (section 2.8).

Prediction models for average voice support $ST_V$ and average reverberation time at the ears $T_{30,\text{ears}}$ in rooms, as a function of the volume $V$ and the diffuse-field reverberation time $T$, have been presented in sections 2.4.4 and 2.5.2, respectively. Figure 3.1 shows the mutual relationship between the two magnitudes $ST_V$ and $T_{30,\text{ears}}$, for equal values of $V$ (dotted lines) and $T$ (solid lines). The curves in the figure have been calculated for a room with proportions 2.8:1.6:1. However, other room proportions result in nearly identical curves.

On the bottom axis of figure 3.1, there is an indication of the relative voice power level $\Delta L_W$ used by a speaker in the presence of low background noise levels. The values on the axis are derived from the findings in paper A and are calculated from Eq. (2.16) for $ST_V < -14.5$ dB and from Eq. (2.17) for $ST_V$ values between -14.5 and -6.5 dB. The values in this axis illustrate how different classroom acoustic designs affect the voice levels of teachers while the audience is silent.

As an example, one can evaluate the effect of reducing the reverberation time in a room of 200 m$^3$ from 1 s to 0.5 s. In figure 3.1, the intersection between $V = 200$ m$^3$ and $T = 1.0$ s occurs for $ST_V \approx -10.5$ dB, which corresponds to $\Delta L_W \approx -5.7$ dB. The intersection between $V = 200$ m$^3$ and $T = 0.5$ s occurs for $ST_V \approx -12.8$ dB,
3.2 Speaker-oriented classroom acoustics design

Figure 3.1: Reverberation time at ears versus voice support for different values of diffuse-field reverberation time (solid lines) and volume (dotted lines). The variations in voice power level as a function of the voice support expected from the results of paper A in the presence of low background noise levels are shown in the horizontal axis at the bottom. The optimum acoustic conditions for a speaker are indicated in hatched areas, as a function of the number of students in the classroom. All the values correspond to occupied classrooms.
which corresponds to $\Delta L_W \approx -3.0$ dB. Therefore, the reduction of reverberation time in a room of 200 m$^3$ from 1.0 s to 0.5 s would result in an increase of voice power level of $-3.0 - (-5.7) = +2.7$ dB.

Although these curves have been obtained for the teaching scenario, they could in principle be used to assess voice power level variations in other rooms for speech, ranging from meeting rooms to drama theaters.

Appendix A describes the derivation of optimum acoustic conditions for a speaker based on the statistical models of Hodgson et al. [33] for student-activity noise, teachers’ voice power levels, and speech SPL throughout the classrooms. As a summary, the criteria used to define optimum acoustic conditions for a speaker are:

1. A reverberation time at the ears between 0.45 and 0.55 s which maximizes the acoustic comfort for a speaker

2. A speech SPL of at least 50 dB throughout the room.

3. A speaker should not have to use average voice power levels higher than a limit value (chosen as 66 dB)

4. The volume of the room has to be appropriate for the number of students in the classroom (a minimum value of 6 m$^3$/student is used)

The rather arbitrary value of 66 dB chosen at the third point is not relevant for the reported results in rooms with more than 15 students, as the limitation on room volume (fourth point) limits the maximum voice power levels derived from the prediction model.

According to these criteria, the optimum acoustic conditions for a speaker are shown in the hatched areas in figure 3.1 and in table 3.1 for classrooms with 10, 20, and 40 students.

Table 3.1: Recommended ranges of values for the parameters voice support $ST_V$, reverberation time $T$ and volume $V$ for a speaker-oriented classroom acoustic design, as a function of the number of students $N$.

<table>
<thead>
<tr>
<th>$N$</th>
<th>$ST_V$ [dB]</th>
<th>$T$ [s]</th>
<th>$V$ [m$^3$]</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>-10.5 to -8.0</td>
<td>0.5 to 0.65</td>
<td>70 to 170</td>
</tr>
<tr>
<td>20</td>
<td>-11.5 to -9.5</td>
<td>0.55 to 0.7</td>
<td>120 to 210</td>
</tr>
<tr>
<td>40</td>
<td>-12.0 to -11.5</td>
<td>0.7 to 0.75</td>
<td>240 to 280</td>
</tr>
</tbody>
</table>

For an optimum acoustic design for a speaker, the average voice support has to be
3.2 Speaker-oriented classroom acoustics design

in the range between -10.5 and -8.0 dB in small classrooms with 10 students, which corresponds approximately to volumes between 70 and 170 m$^3$ and reverberation times between 0.5 and 0.65 s. In classrooms of about 20 students, the average voice support should be between -11.5 and -9.5 dB, which corresponds to volumes between 120 and 210 m$^3$ and reverberation times between 0.55 and 0.7 s. For classrooms with about 40 students, the optimum design area is quite narrow. The voice support should be around -12.0 dB, the reverberation time around 0.7 s, and the volume between 240 and 280 m$^3$.

These optimum conditions are for an average description of a classroom, where the position of the teacher is not defined (because the models of voice support and reverberation time at the ears defined in chapter 2 only characterize average values across the room). Therefore, these guidelines are useful in the case of participative teaching methods where the teacher stands at many position in the classroom and the students themselves need to be heard across the classroom as well.

The analysis in Appendix A states that optimum acoustic conditions for a speaker cannot be achieved globally in rooms for more than 50 students because of practical space requirements. However, in such classrooms, frontal teaching methods (lecturing style) are much more common, which means that the acoustic design can be optimized for a certain position of the speaker.

It is possible to improve the acoustic conditions for a speaker in large rooms by using reflector panels above the speaker, which increase the speech SPL at remote positions in the classroom by adding early reflections, without degrading speech intelligibility. A few early reflections should return to the speaker in order to provide enough voice support and not demand excessively high voice levels from the speaker. Different textbooks (e.g., Templeton [101, pp.72–73]) give examples on how to design the acoustics of lecture rooms.

As stated in the beginning of chapter 2, the mechanism of vocal intensity adjustment is driven by the desire of successful communication (see figure 2.1). Ensuring enough speech SPL in the audience area by adding early reflections, combined with an adequate voice support at the position of the speaker is the right direction to follow in this case. At the same time, the voice support at the student positions should be kept low in order to reduce conversational feedback and achieve low student-activity noise.

Electroacoustic amplification may also be used in classrooms for more than 50 students. Electroacoustic amplification in classrooms helps teachers to reduce their vocal loading because of lower demands in terms of vocal intensity.
3.3 Combined design for listeners and speakers

The optimum acoustic conditions for a speaker found in the previous section have to be assessed in terms of the requirements for listeners. One requirement for listeners was already included for the determination of the optimum range: the speech SPL had to be at least 50 dB across the audience area, in order to provide a SNR of +15 dB with the students being silent.

A descriptor of speech intelligibility with low background noise levels is the early/late ratio $C_{50}$ (where the 50 stands for a threshold of 50 ms between early and late reflections). For purposes of calculation, an exponential decay of the acoustic energy density $E(t)$ in a room—after reaching steady-state—is assumed (see e.g., Kuttruff [53, p.139])

$$E(t) = E_0 e^{-\frac{6 \ln 10}{T} t},$$

where $E_0$ is the initial energy density present in the room. The early-to-late ratio $C_{t_x}$ for a early-late threshold $t_x$ is

$$C_{t_x} = L_{E, \text{early}} - L_{E, \text{late}} = 10 \log \left( e^{\frac{6 \ln 10}{T} t_x} - 1 \right),$$

with $L_{E, \text{early}}$ and $L_{E, \text{late}}$ indicating explicitly the early and late energy levels (with an arbitrary reference).

Finally, a model that represents $C_{50}$ is

$$C_{50} = 10 \log \left( e^{\frac{6 \ln 10}{T}} - 1 \right),$$

which only depends on the reverberation time. The volume would usually be an important indicator of the relative strength of the direct sound and the reverberant tail. The model in Eq. (3.4), however, assumes a position far away from the source which is marginally affected by the direct sound but is representative of the worst case scenario.

Nijs and Rychtarikova [77] expressed the requirements of speech intelligibility in terms of the useful-to-detrimental ratio $U_{50}$, which in the presence of low background noise levels equals $C_{50}$. These requirements are illustrated as different levels of gray in figure 3.2. On the same figure, the optimum ranges of acoustic conditions for a speaker are indicated. They all lay in the area defined as 'good speech intelligibility'. Because the model does not take into account the influence of the direct sound, the
speech intelligibility is underestimated. The speech intelligibility can be improved on the audience area by designing the rooms with specific reflector panels to increase the level of early reflections without increasing too much the late reverberation. However, such room design requires specific locations of teacher and students, which may be incompatible with new teaching styles.

The optimum acoustic conditions for a speaker are shown together with the requirements of different standards for classroom acoustics in figure 3.3. The optimum
acoustic conditions for a speaker are derived from actual teaching performance, therefore the reverberation times correspond to occupied conditions. As can be seen, the optimum conditions for a speaker exceed the values of $T$ in occupied classrooms marked by the German standard DIN 18041 [21]. The optimum conditions for a speaker also exceed the limits of the American standard ANSI S12-60:2002 (R2009) [1] in most of the range. However, a classroom compliant with this standard would have lower reverberation times than shown in figure 3.3 under occupied conditions.

Nevertheless, Bradley [11] suggests that these standards might be a little too restrictive and that the optimum reverberation times under occupied conditions should be between 0.5 and 0.7 s, whereas acceptable values should be between 0.4 and 0.8 s. According to Bradley’s criteria, the optimum acoustic conditions for a speaker (see table 3.1) would almost correspond to the optimum acoustic conditions for listeners.

### 3.4 Summary

The acoustic conditions of classrooms play an important role in determining the success of the communication in the teaching/learning process. Traditionally, classroom acoustic design has been approached in terms of speech intelligibility. However, there were no means in the literature to assess the effects of different designs on the voice of teachers, who often suffer from voice problems.

Considerations of teachers’ subjective preference on rooms for speech, maximum desirable voice levels, speech SPL across the audience, and practical design requirements have been combined to obtain a set of optimum acoustic conditions for a speaker.

If the room is to be used for flexible teaching styles, there is no possible optimum design when there are more than approximately 50 students in the classroom. However, larger rooms may be regarded as lecture rooms geometrically designed to enhance early reflections across the audience for a determined arrangement of audience and position of the speaker. In large rooms, it is desirable that some of the early reflections come back to the speaker to enhance the voice support.

According to Bradley, the optimum acoustic conditions for a speaker fulfill the optimum requirements for listeners, despite requiring higher reverberation times than those indicated in the current standards.

In order to prove the validity of the theoretical derivations in this chapter, the suggested design guidelines of optimum acoustic conditions for a speaker should be ex-
3.4 Summary

Figure 3.3: Optimum acoustic conditions for a speaker as a function of the reverberation time at ears and the voice support (hatched areas). The target diffuse-field reverberation time for occupied classrooms according to German standard DIN 18041 is shown as a black solid line, whereas the limit reverberation time for unoccupied but furnished classrooms is shown as a dashed line. The requirements for unoccupied classrooms according to ANSI S12-60:2002 are shown in orange.
3. Implications for classroom acoustics design

experimentally verified in subjective (e.g., preference, satisfaction) and objective terms (e.g., evolution of student-activity noise, scholar achievement). Another shortcoming of the derived optimum acoustic conditions for a speaker is that the long-term benefits of classroom acoustic design are largely unknown in terms of well-being and reduction of vocal loading of the teacher.

The derivation of optimum acoustic conditions for a speaker has been largely determined by experimentation and measurements in university teachers and students. Student-activity noise generation patterns in primary and secondary school students may differ significantly from the ones at university, which in turn may modify the optimum acoustic conditions for a speaker.
Chapter  4

General discussion

As it has been shown in chapter 2, there is an effect of the acoustics of classrooms on the way that teachers adjust their voices and on the perceived acoustic comfort. When describing this effect, the sources of bias can be in the characterization of the classroom acoustic conditions (section 4.1), in the measurement of the vocal intensity variations (section 4.2), or in the determination of acoustic comfort (section 4.3). Furthermore, there are different factors that can affect directly the relation between classroom acoustics, vocal intensity adjustment, and speakers’ comfort during the experimental design (section 4.4).

Section 4.5 summarizes the statistical analysis methods used in the different papers, which aimed at maximizing the inference from the available measurement data and avoiding bias on the conclusions extracted from the analyses.

The knowledge obtained with the research reported in chapter 2 was used to derive the optimum classroom acoustic conditions for a speaker in chapter 3. The effectiveness of the suggested measures is discussed in section 4.6.

Finally, section 4.7 points possible directions to follow in future research in the area which relates room acoustics and speakers’ voice adjustment.
4. General discussion

4.1 Factors affecting the characterization of classroom acoustic conditions

4.1.1 Measurement equipment dependency

The objective measures developed during the PhD project, namely the room gain, the voice support, and the reverberation time at the ears, have been defined in connection with a very specific model of equipment: the dummy head B&K type 4128 with left ear simulator type 4159 and right ear simulator type 4158. Section 2.4.3 specified the sources of bias potentially affecting the determination of the final result: the airborne direct sound path, the radiation characteristics, and the head-related transfer function (or diffuse-field response).

The recommendation ITU-T P.58 [41] specifies the maximum admissible limits for the measurement equipment in terms of the diffraction from the mouth to the eardrum, radiation characteristics, and diffuse-field response. Although paper A defined the measurement of room gain and voice support with ITU-T P.58 [41] compliant equipment, the uncertainty introduced by the use of different equipments within this recommendation is largely unknown without further investigation.

4.1.2 Body conduction

In all the acoustic parameters for a speaker that have been defined, the effect of the body conduction of one’s own voice has been disregarded. If taken into account, the measured values of voice support would vary by a fixed amount of decibels, equal in every case (remember the definition of the voice support: difference between the energy level of the reflections and the energy level of the direct sound). The measured values of room effect (voice level variation versus voice support variation) in this case would not be affected. However, the influence in the room gain would not be linear, because the total energy level and the energy level of the direct sound would be affected by the body conduction. In this case, the values of room effect reported in chapter 2, expressed as a function of the room gain, would change.
4.1 Factors affecting the characterization of classroom acoustic conditions

4.1.3 Lack of validation

Reverberation time at the ears

A possible source of bias in the recommendations of chapter 3 is the lack of experimental validation of the prediction model for reverberation time at the ears developed in section 2.5.2. However, this prediction model has been built on the same assumptions as the prediction model for voice support, which was an unbiased estimator of the average voice support measured in classrooms (in paper D).

Laboratory setup

The loudspeaker-based system for real-time auralization of one’s own voice described in paper F and used in the experiments of paper G lacked an experimental validation which ensured a matching between the desired acoustic conditions to be simulated and the acoustic conditions actually simulated in the laboratory room.

A proper validation of the system would have required a set of expected values of voice support and reverberation time at the ears, calculated from the computer acoustic simulation program, and the corresponding IRs to generate the acoustic environments in the laboratory room. When the IRs be loaded in the loudspeaker-based system, objective acoustic measurements performed in the laboratory room should have values similar to the expected ones.

The consequence of a deviation from the expected acoustic properties is a mismatch between the perceived room and the room that the simulations aimed at. With mismatching acoustic properties, there is a risk that the simulated acoustic conditions are not feasible in any real room and the sound becomes unnatural. For example, this mismatch would happen if the gain applied to the IR of a concert hall was too high (similar to the gain in a small room). There is no smaller room that can produce the same IR as a concert hall, with the same reverberation time, direct-to-reverberant ratio, and early reflection pattern (with relatively long delays).

Nevertheless, it is unknown how important a mismatch is in terms of perception, or how sensitive the human ear is to deviations from “real” conditions. It would have been ideal to avoid this bias, although measurements suggested that the subjective perception and the reactions of the speakers were correlated to the parameters reverberation time
4. General discussion

at the ears and voice support measured “in-situ” inside the loudspeaker-based system—
independently of the conditions that were aimed at during the computer simulations.

4.2 Sources of bias in vocal intensity

4.2.1 Vocal intensity measures

The aim of the project was to characterize the variations in vocal intensity of speakers
in different acoustic environments. Therefore, the absolute values of the vocal intensity
measures were not given much relevance and were assumed to be similar to the vocal
intensity characterized in extensive and systematic measurements reported in the works
of Pearson et al. [81] and more recently, Cushing et al. [18]. Only the relative values of
the vocal intensity measures were of interest in the present study.

Different magnitudes were used to characterize the vocal intensity:

- In paper A, the relative voice power level, using as a reference for each subject
  the voice power level in free-field condition
- In paper B, the absolute voice power level
- In paper C, the SPL at the microphone position (located on the cheek of the
  speaker, at approximately 6 cm from the edge of the lips), relative to the SPL in
  free-field condition
- In paper E, the 1 s-equivalent A-weighted SPL that was exceeded 50% of the time
  at the position of the microphone, at a distance of approximately 15 cm from the
  mouth of the teacher
- In paper G, the equivalent on-axis, free-field SPL at a distance of 1 m in front of
  the speaker. This value was derived from the measurement of SPL at the head-
  worn microphone—as in paper C—and a measurement of the level difference
  between the SPL at the head-worn microphone and the on-axis SPL at 1 m in
  front of the speaker in free-field.

The voice power level, or sound power level of a speaker’s voice, is the most
appropriate measure of vocal intensity, as it represents the total radiated acoustic energy
per unit of time. However, the determination of voice power level with measurements
4.2 Sources of bias in vocal intensity

in a reverberation chamber can introduce bias. In paper B, the voice power levels were determined from the measurement of SPL at the head-worn microphone, and a calibration measurement in reverberation chamber to determine the relation between the SPL at the head-worn microphone and the voice power level. This last measurement was performed with human speakers as the sound source, which is problematic from a methodological perspective. The determination of sound power in reverberation rooms is limited to long-cycle noise sources which are able to build up a steady-state sound field [42]. Human speech has large dynamic variations, combining phases of silence with phases of sound radiation in time intervals which are much shorter than the reverberation time of the reverberant chamber used for the measurement of sound power.

Theoretically, the sound power can also be determined by integrating the sound intensity on a surface surrounding the sound source [48], but it is difficult from a practical point of view, since it would not be feasible to measure sound intensity around a speaker while teaching in classrooms. Nevertheless, by assuming a certain average radiation characteristics of the speaker [17], the voice power level can be calculated from the on-axis SPL in free-field conditions (i.e., without having an influence of the reflections from the room boundaries). This is the reason why the measurements in papers C, E, and G were focused on the SPL and did not report voice power levels.

Ideally, the same magnitude should have been used throughout the different investigations. Nevertheless, it is believed that the bias introduced by the vocal intensity measure in the room effect is much lower than the intersubject variation of the room effect measure.

4.2.2 Intersubject variation

Apart from the uncertainty introduced by the choice of vocal intensity measure, there is additional uncertainty related to the choice and the characteristics of subjects, mainly due to three aspects. First, different individuals speak at different levels due to gender, cultural, health condition, and physiological differences. Second, individuals may interpret or perform tasks in different ways when asked to. An example of it is the way in which subjects read a text in paper G. Some subjects read as if it was a story for little children, whereas others read in a totally dispassionate way.

The third aspect is the variation in placement of the measurement microphone. The microphone could move slightly (in the order of millimeters) while the subject turned
his head, which could have resulted in level variations due to a change of sound propagation distance. Additionally, the geometry of the head differed across subjects, which could have resulted in different relations between the SPL at the head-worn microphone and the free-field, on-axis SPL at 1 m in front of the speaker. These variations were not quantified in the present study, and the difference between the SPL at the head-worn microphone and the free-field, on-axis SPL at 1 m in front of the speaker, averaged across speakers, was assumed to be equal to that for the HaTS B&K type 4128.

4.3 Sources of bias in the judgment of acoustic comfort

The sensation of acoustic comfort as judged in paper G could have been biased by a number of factors:

1. The slightly different placement of the head-worn microphone in each subject, and the geometrical differences among the heads of the subjects, resulted in different gains applied to the simulated room reflections, therefore producing the sensation of different rooms in different individuals.

2. The interpretation of the questions, or the different ideas on what are good and bad acoustic conditions for a speaker.

3. In connection with the previous factor, the knowledge of room acoustics and the knowledge that longer reverberation times enhance student activity noise may have introduced some bias.

4. The instruction used in the laboratory experiments (reading a text aloud or describing a map) was not representative of real teaching, so there might be a gap between how teachers interact with the acoustics of the classroom in actual teaching and in laboratory.

It is convenient to point out that the answers to the questionnaires were not biased by individual scale and range effects, because they were accounted for by using $z$-scores of the answers during the statistical analysis.
4.4 Factors affecting the room effect

As was observed in the experiments reported in papers B and G, the variation of vocal intensity in different acoustic environments, i.e., the room effect, is a highly individual characteristic.

**Teaching experience**

Initially, the teaching experience seemed to be an important factor in the interaction with the acoustics of classrooms. It may seem reasonable that teachers learn from their own experience to adjust their voices according to the acoustics of the classrooms. However, there was no sign of a different performance between a group of teachers and a group of students in the laboratory experiments of paper G.

**Performance in noise and silent conditions**

There can be a large uncertainty in the measurements of the room effect under actual teaching conditions because student-activity noise depends on the acoustics of the room and then teacher reacts to both room acoustics and student-activity noise. In silence, an increase in voice support would make teachers lower their voice levels (paper A). With the prediction model of Hodgson *et al.* [33] (and appendix A), an increase in voice support would increase student-activity noise and therefore, would also increase the required voice level of the teacher. At the same time, the dependency of student-activity noise with voice support may be influenced by the education level (primary, secondary, tertiary), the teaching/learning methods, and the skill of the teacher for managing classroom noise and engaging students into learning.

**Healthy and non-healthy voices**

Paper E reported a significant difference in the room effect between teachers with healthy voices and teachers with voice problems, which was believed to be due to an increased sensitivity of teachers with voice problems toward their teaching environment. However, this difference was based on a group-wise statistical analysis, and there were no means to determine individual reactions of teachers to different room acoustic conditions. The two group of teachers did not perform differently under laboratory conditions in the experiments reported in paper G. It would have been desirable to perform a more
complete field study, where every teacher had to teach in rooms of different acoustic conditions.

**Laboratory performance and importance of the task**

As pointed out, the laboratory performance was in some cases very different from actual teaching, and paper G reveals the importance of the instruction on the measured room effects. Speakers who were requested to give a lecture to an imaginary group of 30 students performed much more similar to actual teaching conditions than the speakers who were requested to read or to describe a map for a single listener. Speakers who were asked to read or to describe a map adjusted their voices to hear themselves equally loud (on average). Therefore, the task has to be close to real teaching and well-defined, so as to minimize the need of interpretation by the teacher.

### 4.5 About the statistical methods

Different statistical methods have been used in different papers for analyzing data.

In paper A, simple linear regression models were used, which modeled the relative voice power level as a function of voice support or room gain. Mixed models might have been more appropriate, considering the subject as a random effect. However, the simple linear regression model was chosen to describe the variation on the average data, as in the paper of Brunskog *et al.* [16].

In paper B, mixed models were used to describe the voice power level (and other voice parameters) as a function of the talker-to-listener distance and the room. The subject was considered to introduce a random effect on the observations. The use of mixed models provided a more realistic description of the data than fixed-effects models, because the results could be generalized to the overall population of speakers. Furthermore, the assumption of normality of the random effects was fulfilled. The standard deviation of voice power levels across subjects was around 2.7 dB.

In paper C, a four-way ANOVA was used, in order to determine the influence of the acoustic condition, the gender, the vowel used, and the reference signal on the relative voice level used by the speaker to match the loudness of the reference signal. After determining that, among these variables, the acoustic condition and the gender explained almost 90% of the variance, non-linear models were used to fit the relative voice level
4.6 Effectiveness of the suggested measures

With the research done in this PhD thesis, it is not possible to accurately quantify the beneficial effects of optimum classroom acoustic design for a speaker in terms of actual vocal loading, working satisfaction, or student achievement.

The results of the pilot laboratory experiments in paper G show qualitative effects of non-noisy teaching environments: with increasing voice support, voice level significantly decreases (negative room effect in figure 2.21 for PRE1 and PRE2) but the phonation time increases (positive slopes in figure 2.24 for PRE1 and PRE2). An increase in phonation time partially counteracts the contribution of the decreased voice level to vocal loading. Nevertheless, an increase in phonation time may be beneficial to speech intelligibility [62].

The design guidelines of figure 3.1 indicate average voice level variations in rooms in the presence of low background noise levels, showing a decrease in voice level with increasing voice support. However, in the presence of student-activity noise, the voice level variations may be different and voice level may even increase with voice support (see figures A.2 and A.3 in appendix A). Therefore, the suggested design guidelines of optimum acoustic conditions for a speaker described in chapter 3 should be experi-
mentally verified in subjective (e.g., preference, satisfaction) and objective terms (e.g., evolution of student-activity noise, scholar achievement).

The vocal intensity variations shown on the axis at the bottom of figure 3.1 were calculated with an average value of room effect across subjects. Some individuals react to different room acoustics much stronger than the average, softening their voices in environments that provide enough support, which might be beneficial in terms of vocal endurance. It would be interesting to find out whether other speakers could learn to enhance the room effect to obtain long-term benefits on their vocal health.

The judgment of the effectiveness of optimum classroom acoustic design for a speaker should be assessed when more studies of voice loading are available, and when the knowledge about vocal fatigue and recovery processes increases.

Therefore, it is not clear whether actions to improve classroom acoustics would contribute by themselves to reduce the prevalence of voice problems among teachers significantly. However, it would be positive in any case. A good classroom acoustic design may originate a domino effect, in which the teacher feels more confident and is able to teach in a more engaging way, improving the attention of students, reducing student-activity noise and stress levels, and improving students’ performance.

Approaching teachers’ voice problems requires a combination of preventive actions, not only regarding an optimal classroom acoustic design, but also in terms of voice training and instruction at university programs for teacher education that raise awareness among teachers, teaching schedules that facilitate vocal rest, workshops and information regarding voice at teachers’ unions, and providing the means to treat early symptoms of voice problems. In large classrooms with more than 50 students, either an acoustic design to enhance early/late ratios at specific locations of teacher and students or the use of electroacoustic amplification systems is required.

4.7 Future directions

The present work has introduced different objective acoustic parameters for a speaker—room gain, voice support, and reverberation time at the ears—and their relevance in terms of acoustic comfort and voice level variation across classrooms.

Although paper D presented the measurement of voice support and room gain in 30 classrooms, it is necessary to measure these parameters and the reverberation time at the ears in a wider range of classrooms with different characteristics, validating the
4.7 Future directions

proposed model of reverberation time at the ears. It would be desirable to link the measured values of these parameters during a field research to the auditory perception of the environment in response to the voice produced.

Moreover, the findings on paper E about the significant differences between teachers with and without voice problems in their response to voice support should be proved with a factorial experimental design, where every teacher would teach in more than one classroom. In future experiments, the long-term effects of the acoustic environment on voice health should be at focus. A more complete description of voice loading in terms of vocal doses is required. This would require reporting not only vocal intensity, but also fundamental frequency and phonation time.

Student-activity noise has to be taken into account for a more detailed description of the room effect. In these studies, the number of students and the teaching methods should be monitored. Separate analyses of voice support in the teaching area and the student area could bring new ideas on how to design classroom acoustics so as to minimize the conversational noise among students while supporting and providing comfort to the voice of the teacher.

The individual characteristics of the room effect could be further investigated, looking at which are the factors that raise awareness among speakers about the room acoustic conditions, whether it develops with experience, or whether it is possible to learn. In this case, it could be possible to instruct teachers so that they benefit from adapting their voice to the acoustic features of the room.
4. General discussion
Chapter 5
Conclusions

Teachers are one of the professional groups with the highest risk of suffering from voice disorders. Among different causes, they claim classroom acoustics, and not only background noise, to be one of the potential hazards affecting their vocal health. The present project investigated the relationship between classroom acoustics and voice regulation, focusing on the vocal intensity as the main parameter. The main conclusions are

- Teachers adjust their vocal intensity according to the room gain or voice support of the classrooms, which are equivalent objective measures that quantify the amplification of one’s own voice in a room due to the reflections at the boundaries. The variation of vocal intensity with these measures is referred to as *room effect*.

- The magnitude of the room effect is highly dependent on the individual.

- Most of the vocal intensity variation in the room effect is due to the dependency of the room gain and the voice support with room size, and thus with the average talker-to-listener distance. However, there is a significant room effect for equal talker-to-listener distances, which becomes stronger at longer distances.

- For a distance of 1.5 m, the vocal intensity adjustment strategy of a speaker is to keep the autophoncic level (i.e., the loudness level of his own voice) constant. At distances of 6 m or further, the variation in voice level due to the room gain or the voice support is twice the variation required to keep the autophoncic level constant.

- Teachers with and without self-reported voice problems react differently to the voice support of the classrooms where they teach, whereas they react equally to
variations in background noise, increasing their voice levels approximately 0.86 dB per each dB of increase in noise. Teachers with voice problems might be more sensitive to the conditions of the working environment than their healthy colleagues.

- The room effect derived from laboratory experiments in virtual classrooms depends highly on the methodology and the task used. Experiments with simulated teaching to an imaginary group of 30 students replicated and confirmed the existence of the room effect with low background noise levels. Other tasks, as reading or describing a map, resulted in much fainter room effects. In this case, laboratory performance does not correspond to the performance during actual teaching.

- The optimum acoustic comfort for a speaker is obtained for values between 0.45 and 0.55 s of the parameter *reverberation time at the ears*. This parameter is the reverberation time derived from the decay between -5 and -35 dB of the backward integrated energy curve of the impulse response measured between the mouth and the ears of a dummy head.

- Speaker-oriented classroom acoustic design guidelines have been derived from prediction models for voice support and reverberation time at the ears and an empirical model by Hodgson *et al.* [33]. For flexible teaching methods, classrooms shall not have more than 50 students. The voice support shall be in the range between -12 and -8 dB, the reverberation time in occupied conditions shall be between 0.5 and 0.75 s, and the volume shall be between 70 and 300 m$^3$, respecting a series of constraints among these magnitudes imposed by the number of students. The reverberation times derived in these guidelines are higher than those recommended by the current classroom acoustic standards, but are in agreement with recent findings by Bradley [11].

- Classrooms for more than 50 students can benefit from specific acoustical design to enhance early reflections and minimize late reverberation for given positions of the speaker and the audience. In these classrooms, it is advised to use electroacoustic amplification systems, especially if they are to be used for flexible teaching methods.

- There is not enough scientific evidence that speaker-oriented classroom acoustic design can reduce the prevalence of voice problems among teachers. It is there-
fore necessary to combine this approach with other preventive actions that reduce vocal loading and improve the capacity of coping with vocal loading, such as voice training programs, adaptation of teaching methods, and teaching schedules that allow for vocal recovery. Moreover, teachers should be more aware of the early symptoms of voice problems and have an efficient access to their treatment.
5. Conclusions
Appendix A

Optimum acoustic conditions for a speaker

In order to obtain optimum acoustic conditions for a speaker, the requirements for a speaker have to be specified in terms of $T_{30,\text{ears}}$ and $ST_V$, or in terms of $T$ and $V$. The first design goal is the maximization of acoustic comfort for a speaker in terms of $T_{30,\text{ears}}$ as reported in section 2.8:

- $0.45s < T_{30,\text{ears}} < 0.55s$.

With the knowledge obtained in papers A to G, it is not possible to set requirements for the voice support. First, it is not known how the small voice intensity variations caused by the room effect would affect the long-term performance of a teacher. Second, the room effect has been mostly analyzed with low background noise levels, which is not the case during real teaching, as the field study measurements showed statistical A-weighted background noise levels $L_{50}$ in the range between 40 and 75 dB (section 2.6.3 and paper E).

Hodgson et al. [33] made measurements of noise levels and speech SPL in classrooms, together with several physical parameters of the room—including width and volume—and teaching scenario (e.g., number of students and distance between instructor and students) in a total of 11 university classrooms during 18 lectures. These measurements were used to make prediction models for:

- Ventilation noise level
- Student-activity noise level
Appendix A: Optimum acoustic conditions for a speaker

- Instructor speech SPL
- Instructor sound power level

A.1 Student-activity noise level

From the models proposed by Hodgson et al. [33], only the student-activity noise level and the instructor sound power level are used below. The student-activity noise level SA (using the alternative approach suggested by Hodgson et al. [33]) is described as

\[
SA = 83.0 + 10.0 \log N - 34.4 \log A_0 + 0.081 A_0 \text{[dB]} \tag{A.1}
\]

where \( N \) is the number of students in the room and \( A_0 \) is the total absorption area in the room (which can be calculated from Sabine’s formula as \( A_0 = 0.161V/T \) in international system units).

The student-activity noise levels are shown in figure A.1 as a function of \( STV \) and \( T_{30,ears} \), for different values of the number of students \( N \). As can be seen, the lines corresponding to equal SA are almost vertical, except for the lowest range of reverberation time, where Sabine’s formula does not hold. The prediction model shows that, when the voice support increases, SA increases due to conversational feedback or café effect [120] (even though the university students were not particularly noisy). When increasing the number of students, the same value of voice support results in higher SA values.

A.2 Voice power level of a speaker

The prediction model for the voice power level \( L_W \), averaged for male and female instructors, is

\[
L_W = 53.5 + 0.5SA + 0.016V - 9.6 \log A_0 \text{[dB]} \tag{A.2}
\]

As can be seen, the voice power level depends on the student-activity noise level, where the coefficient 0.5 is sometimes called the Lombard slope and is in good agreement with other studies [57]. It also depends on the volume \( V \) of the room and the total absorption area \( A_0 \).
A.2 Voice power level of a speaker

Figure A.1: Student activity noise level iso-contours for different number of students in a classroom (top-left N=10, top-right N=20, bottom-left N=40, bottom-right N=80), according to Hodgson et al. [33], expressed as a function of voice support and reverberation time at ears. The reverberation time and volume guidelines are shown in gray.

The equal voice power level contours derived from Eq. (A.2) are presented as a function of $ST_V$ and $T_{30,ears}$ in figure A.2 for N=10 and N=20 students and figure A.3 for N=40 and N=80 students. These curves illustrate different vocal behaviors of the teacher that occur above and below a volume of approximately 500 m$^3$. For large volumes, increasing voice support values result in a reduction of $L_W$, as would be expected from the room effect under conditions of low background noise level. However, for small room volumes, an increase in voice support means an increase in the $L_W$ of the teacher, who has to raise the voice to overcome the student-activity noise. Another
Appendix A: Optimum acoustic conditions for a speaker

A general observation in these figures is that, as the number of students increases, the predicted $L_W$ for fixed objective parameters in the room increases. For example, in a room of $V = 400 \text{ m}^3$ and $T = 1.0 \text{ s}$, an instructor would need to use a voice power level of 61 dB for 10 students, whereas he would need to use 65.5 dB for 80 students.

A.3 Speech SPL in a room

The speech SPL (SL) at the student position (away from the source) is calculated in a slightly different way as done in Hodgson et al. [33]. Here, one approximation is made: the speech SPL far away from the source can be calculated from the sound power level according to the diffuse-field theory (disregarding the attenuation of SPL with the distance to the source, as measured by Barron and Lee [3] or Sato and Bradley [93]),

$$SL = L_W + 10 \log \left( \frac{4}{R} \right),$$

(A.3)

where $L_W$ is the voice power level of the model in Eq. (A.2) and $4/R$ is

$$\frac{4}{R} = \frac{cT}{6V \ln 10} - \frac{4}{S},$$

(A.4)

with $c$ being the speed of sound in the air and $S$ the total surface area of the room. The equal speech SPL contours predicted with the model in Eq. (A.3) are shown in figure A.4 for $N=10$ and $N=20$ students, and in figure A.5 for $N=40$ and $N=80$ students. As happens in the curves for $L_W$, an increase in voice support results in increased speech SPLs due to two facts: the increase in $L_W$ and the increase of the reverberant energy.

A.4 Criteria and recommendations for optimum acoustic design for a speaker

In order to determine good conditions for a speaker, two more requirements are established:

- The average voice power level $L_W$ should be limited to a maximum value. For convenience, this value is arbitrarily set to 66 dB, which is 1.5 dB higher than the average 64.5 dB found by Hodgson et al. [33].
A.4 Criteria and recommendations for optimum acoustic design for a speaker

Figure A.2: Voice power level iso-contours of a speaker for different number of students in a classroom (left N=10, right N=20), according to Hodgson et al. [33], expressed as a function of voice support and reverberation time at ears. The reverberation time and volume guide lines are shown in gray. The shaded area indicates the optimum design parameters for a speaker. The arrows show the maximum and minimum limits of voice support in the shaded area.
Figure A.3: Voice power level iso-contours of a speaker for different number of students in a classroom (left N=40, right N=80), according to Hodgson et al. [33], expressed as a function of voice support and reverberation time at ears. The reverberation time and volume guide lines are shown in gray. The shaded gray area indicates the optimum design parameters for a speaker. The arrows show the maximum and minimum limits of voice support of the shaded area.
A.4 Criteria and recommendations for optimum acoustic design for a speaker

Figure A.4: Speech SPL iso-contours of a speaker for different number of students in a classroom (left N=10, right N=20), according to Hodgson et al. [33], expressed as a function of voice support and reverberation time at ears. The reverberation time and volume guidelines are shown in gray. The shaded gray area indicates the optimum design parameters for a speaker. The arrows show the maximum and minimum limits of voice support of the shaded area.
Appendix A: Optimum acoustic conditions for a speaker

Figure A.5: Speech SPL iso-contours of a speaker for different number of students in a classroom (left N=40, right N=80), according to Hodgson et al. [33], expressed as a function of voice support and reverberation time at ears. The reverberation time and volume guide lines are shown in gray. The shaded gray area indicates the optimum design parameters for a speaker. The arrows show the maximum and minimum limits of voice support of the shaded area.
A.4 Criteria and recommendations for optimum acoustic design for a speaker

- The average speech SPL at the listener position should be higher than 50 dB, in order to ensure a SNR of at least 15 dB when the students are silent, provided that the classroom meets the current classroom acoustics standards (e.g., [1, 20]) in terms of unoccupied background noise levels, which should not exceed 35 dB.

These two requirements are combined with the optimization of the acoustic comfort for a speaker (i.e., a reverberation time at ears between 0.45 and 0.55 s) and are shown in gray shaded areas in the figures A.2 to A.5. These shaded areas reveal approximate ranges of voice support for classrooms with different number of students that, combined with reverberation times at the ears between 0.45 and 0.55 s, result in the best possible acoustic conditions for a speaker. These values are summarized in table A.1. Alternatively, the acoustic conditions are expressed in terms of reverberation time and volume of the room in an approximate way in table A.1. To ensure the fulfillment of the optimal conditions for a speaker, particular combinations of reverberation time and volume must lay inside the shaded areas in figures A.2 to A.5 for a particular number of students.

Table A.1: Recommended ranges of values for the parameters voice support $ST_V$, reverberation time $T$ and volume $V$ for a speaker-oriented classroom acoustic design, as a function of the number of students $N$. The two last columns in the table show the recommended areas $S_{rec}$ according to classroom space planning guidelines (e.g., [102]) and the recommended volume $V_{rec}$ for a 3 m high classroom.

<table>
<thead>
<tr>
<th>$N$</th>
<th>$ST_V$ [dB]</th>
<th>$T$ [s]</th>
<th>$V$ [m$^3$]</th>
<th>$S_{min}$ [m$^2$]</th>
<th>$V_{min}$ [m$^3$]</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>-10.5 to -8.0</td>
<td>0.5 to 0.65</td>
<td>70 to 170</td>
<td>20</td>
<td>60</td>
</tr>
<tr>
<td>20</td>
<td>-11.5 to -9.0</td>
<td>0.55 to 0.7</td>
<td>80 to 210</td>
<td>40</td>
<td>120</td>
</tr>
<tr>
<td>40</td>
<td>-12.0 to -9.5</td>
<td>0.6 to 0.75</td>
<td>120 to 280</td>
<td>80</td>
<td>240</td>
</tr>
<tr>
<td>80</td>
<td>-13.0 to -10.5</td>
<td>0.6 to 0.8</td>
<td>140 to 350</td>
<td>160</td>
<td>480</td>
</tr>
</tbody>
</table>

The range of recommended voice support decreases as the number of students increases. For a classroom of 10 students, the voice support should be between -10.5 and -8 dB, whereas for 80 students, it should be between -13 and -10.5 dB. At the same time, the recommended reverberation times and volumes increase with the number of students. For a classroom of 10 students, optimum acoustic conditions for a speaker can be achieved with volumes between 70 and 170 m$^3$ and reverberation times between 0.5 and 0.65 s. For a classroom of 80 students, the recommended volumes are between 140 and 350 m$^3$ and the reverberation times between 0.6 and 0.8 s.

Table A.1 also shows the minimum floor area $S_{min}$ that the classrooms must have for a given number of students, considering that each student needs a floor area of 2 m$^2$. 
Appendix A: Optimum acoustic conditions for a speaker

[102], and the minimum volume of the classroom $V_{\text{min}}$, considering an average ceiling height of 3 m. As can be seen, the recommended volumes for classrooms of 80 students are lower than the minimum required volume for such a number of students. Therefore, optimum acoustic conditions for a speaker cannot be achieved with such a number of listeners. According to the evolution of the recommended volumes and minimum required volumes for different number of students, optimum acoustic conditions for a speaker can be achieved only when the number of students is less than approximately 50.

A.5 Signal-to-noise ratio

The equal signal-to-noise ratio (SNR) contours observed under the presence of student-activity noise according to the prediction models of Eqs. (A.1) and (A.3) are shown in figure A.6 for N=10 and N=20 students, and in figure A.7 for N=40 and N=80 students. The predicted SNR under student-activity noise for the optimum acoustic conditions for a speaker (shown in shaded areas) decreases with the number of students. It is between 8 and 9 dB for 10 students, around 7.5 dB for 20 students, between 6 and 7 dB for 40 students, and between 4 and 5 dB for 80 students. Nevertheless, in classrooms for 80 students, optimum acoustic conditions for a speaker are not possible without optimizing early reflections for particular locations of speaker and listeners or without the use of electroacoustic amplification.
Figure A.6: Classroom SNR iso-contours for different number of students in a classroom (left N=10, right N=20), according to Hodgson et al. [33]. The reverberation time and volume guide lines are shown in gray. The shaded gray area indicates the optimum design parameters for a speaker.
Appendix A: Optimum acoustic conditions for a speaker

Figure A.7: Classroom SNR iso-contours for different number of students in a classroom (left N=40, right N=80), according to Hodgson et al. [33], expressed as a function of voice support and reverberation time at ears. The reverberation time and volume guide lines are shown in gray. The shaded gray area indicates the optimum design parameters for a speaker.
Bibliography


Bibliography


98 Bibliography


Bibliography


Bibliography


Bibliography


Papers A-G
I. INTRODUCTION

Brunskog et al., published “Increase in voice level and speaker comfort in lecture rooms” previously in this journal.1 Their work showed a possible influence of room acoustics (through a new parameter named room gain) on the vocal intensity used by teachers for talking in rooms. In addition, different subjective aspects regarding the perceived acoustic conditions while talking were studied by means of questionnaires. The work extended its relevance to the areas of ergonomics and occupational health, as it described an interaction between man and environment with possible consequences for voice health originated from working conditions. A recent epidemiological study has shown that teachers with voice problems rate classroom acoustics as an element affecting their voice much more often than those without voice problems.2 In this context, the work of Brunskog et al., could offer a reference dataset to compare the vocal performance of teachers’ with and without voice problems under different acoustic conditions. However, it has been impossible to replicate the room gain measurements of Brunskog et al., in the original rooms of their study. The aim of this paper is to provide a more accurate and replicable dataset relating the voice power levels measured by Brunskog et al., to the objective parameters room gain and voice support derived with an alternative method. The first section presents the definition of room gain according to the method of Brunskog et al., pointing out some potential limitations, and it is followed by the definition of room gain and voice support according to an alternative method. The second section compares the objective measurements in the rooms of Brunskog et al., as they appear in the original study and with the alternative method. The last section describes two empirical models relating the voice power levels to the room gain and the voice support.

1Author to whom correspondence should be addressed. Electronic mail: dpg@elektro.dtu.dk

II. ROOM ACOUSTIC PARAMETERS FOR A TALKER

Two equivalent metrics that characterize the effect of room acoustics as perceived by a talker are used: room gain (\(G_{RG}\)) and voice support (\(STV\)).

Brunskog et al., defined the room gain as the degree of amplification produced by the room on the talker’s voice, as perceived by the talker himself. The calculation of room gain proposed in Brunskog et al., requires the measurement of two impulse responses (IRs) corresponding to the sound transmission path between the mouth and the ears of a dummy head: one at the room of interest \(h(t)\) and another one at an anechoic chamber \(h_{ach}(t)\). From these two measurements, the energy levels of the IR at the position of interest, \(L_E\), and at the anechoic room, \(L_{E,ach}\) are calculated

\[
L_E = 10 \log \frac{\int_0^\infty h^2(t)dt}{E_0},
\]

(1)

\[
L_{E,ach} = 10 \log \frac{\int_0^\infty h_{ach}^2(t)dt}{E_0},
\]

(2)

where \(E_0\) is an arbitrary energy reference. The room gain is calculated as the difference between these two energy levels

\[
G_{RG} = L_E - L_{E,ach}.
\]

(3)

The room gain is conceptually related to Gade’s objective support,\(^1\) which is widely used in stage acoustics to compare the energy of early sound reflection patterns from a music instrument to the player’s ears among different rooms.
for music performance. Gade’s objective support is used to characterize many different kinds of instruments, with different distances from the source to the ears of the musicians and different directivity patterns. In the case of voice, the path between mouth and ears is rather well defined.

The method of Brunskog et al. for calculating room gain is conceptually and theoretically correct and can be used to calculate the room gain at positions with very close reflecting surfaces. However, an important limitation of the method is the required measure of an IR in anechoic conditions, which can be an obstacle for many professionals. Additionally, in practice, the IR in anechoic conditions might differ from the direct sound in the measuring conditions due to changes in temperature, humidity, background noise, and distortion artifacts when measuring. The practical limitations lead to measurement error, which is illustrated in the following example.

Nine IRs in a small room, corresponding to the acoustic path between the mouth and the left ear of a Head and Torso Simulator (HATS) B&K type 4128 (Nørum DK-2850, Denmark) with left ear simulator B&K type 4159, were measured with the 01dB Symphonie system (Limonest Cedex F-69578, France). The measurements corresponded to three repetitions at three different reproduction gains, keeping the HATS position fixed. The signal-to-noise ratio (SNR), calculated from the peak level to the noise floor level, was at least 60 dB in all IRs. The IRs were trimmed to the intersection of the exponential decay curve with the noise floor of the measurement with the lowest SNR (the intersection time was noted as \( t_{\text{min}} \)). The IRs were normalized to a peak amplitude of 1, and the energy levels \( L_E \) in the interval \( (0-t_{\text{min}}) \) were calculated. The estimated standard deviation of \( L_E \) was 0.02 dB, whereas the maximum difference between two measurements of \( L_E \) was 0.06 dB. This error is not usually regarded as important, but as defined in Eq. (3), the room gain can be significantly biased by such an amount, since typical values lie between 0 and 0.6 dB.

It would be beneficial to derive the room gain from a single IR measurement and increase the sensitivity of the method. For this, the author proposes the measurement of the IR using a HATS with a mouth simulator, according to recommendation ITU-T P.58,4 and an ear simulator with ear canal, according to recommendation ITU-T P.579 type 3. The source should be at least 1 m away from all boundaries, including the floor, using a stand to appropriately place the HATS at the height of the head of an average standing person. The distance gap of 1 m allows for a time gap free of reflections of approximately 5.8 ms. The direct sound \( h_d(t) \) is obtained by applying a window \( w(t) \) to the measured IR \( h(t) \) (see Fig. 1)

\[
h_d(t) = h(t) \times w(t),
\]

where

\[
w(t) = \begin{cases} 
1 & t < 4.5 \text{ ms} \\
0.5 + 0.5 \cos(2\pi(t-t_0)/T) & 4.5 \text{ ms} < t < 5.5 \text{ ms} \\
0 & t > 5.5 \text{ ms}
\end{cases}
\]

with \( t_0 = 4.5 \text{ ms} \) and \( T = 2 \text{ ms} \). The reflected sound \( h_r(t) \) is the complementary signal

\[
h_r(t) = h(t) - h_d(t).
\]

The energy levels corresponding to the direct sound \( (L_{E,d}) \) and the reflected sound \( (L_{E,r}) \) are calculated as

\[
L_{E,d} = 10 \log \frac{\int_0^\infty h_d^2(t) \, dt}{E_0},
\]

\[
L_{E,r} = 10 \log \frac{\int_0^\infty h_r^2(t) \, dt}{E_0}.
\]

The voice support \( ST_V \), in analogy to Gade’s objective support, is defined as the difference between the reflected sound and the direct sound from the HATS’ mouth to ears IR

\[
ST_V = L_{E,r} - L_{E,d},
\]

which is related to the room gain through the formula

\[
G_{RG} \approx 10 \log \left( 10^{ST_V/10} + 1 \right).
\]

This formula is obtained under the assumption that the total energy is approximately the sum of the energies corresponding to the direct and the reflected sound after windowing

\[
L_E \approx 10 \log \left( 10^{L_{E,d}/10} + 10^{L_{E,r}/10} \right).
\]

Gade’s objective support is intended for big rooms, so the early reflections are counted from 20 ms, and the first 10 ms in the IR are regarded as direct sound. This parameter cannot be used in small rooms (e.g., rooms for speech), as the early reflections are much closer to the direct sound than in large halls, and may fall in the direct sound interval or in the interval from 10 to 20 ms, which is ignored by the definition. With the present definition of direct and reflected paths, it is possible to calculate room gain and voice support in many rooms. The only limitation is that all boundaries of the room should be 1 m away from the measurement equipment.
The indirect calculation of room gain after measuring the voice support with Eq. (10) reduces the deviation in the results. Using the same IRs of the previous example, the standard deviation in the measured room gain was reduced from 0.02 to 0.004 dB, and the maximum differences between two measurements did not exceed 0.01 dB.

III. ABOUT THE MEASURED PARAMETERS

Table I shows the six rooms used in the study of Brunskog et al., with their volume and the original measurements of reverberation time $T_{30}$ and room gain, noted as $G_{RG}$.

<table>
<thead>
<tr>
<th>Name</th>
<th>Abbreviation</th>
<th>$V$ (m$^3$)</th>
<th>$T_{30}$ (s)</th>
<th>$G_{RG}$ (dB)</th>
<th>$G_{RG}$ (dB)</th>
<th>$\Delta G_{RG}$ (dB)</th>
<th>$ST_V$ (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Auditorium 81</td>
<td>A81</td>
<td>1900</td>
<td>1.06</td>
<td>0.28</td>
<td>0.14</td>
<td>0.14</td>
<td>−14.9</td>
</tr>
<tr>
<td>Auditorium 21</td>
<td>A21</td>
<td>1220</td>
<td>1.53</td>
<td>0.29</td>
<td>0.16</td>
<td>0.13</td>
<td>−14.2</td>
</tr>
<tr>
<td>Lecture room 019</td>
<td>LR</td>
<td>190</td>
<td>0.46</td>
<td>0.42</td>
<td>0.32</td>
<td>0.10</td>
<td>−11.1</td>
</tr>
<tr>
<td>Meeting room 112</td>
<td>MR</td>
<td>94</td>
<td>0.42</td>
<td>0.58</td>
<td>0.43</td>
<td>0.15</td>
<td>−9.8</td>
</tr>
<tr>
<td>Large anechoic chamber</td>
<td>ACH</td>
<td>1000</td>
<td>0.06</td>
<td>0</td>
<td>0.01</td>
<td>0.01</td>
<td>−27.3</td>
</tr>
<tr>
<td>IEC listening room</td>
<td>IEC</td>
<td>100</td>
<td>0.34</td>
<td>1.12</td>
<td>0.39</td>
<td>0.73</td>
<td>−10.3</td>
</tr>
</tbody>
</table>

The room IRs in the six rooms of the study were measured again, following the procedure described in the previous section. No filtering, other than the intrinsic response of the loudspeaker, was applied to the signals for deriving the objective parameters. The values of voice support $ST_V$ and room gain $G_{RG}$, measured for each room as the average of six repetitions, are shown in Table I. The differences between old and new room gain values are indicated as $\Delta G_{RG}$.

The new measurements confirm the initial suspicions. The room gain in the IEC listening room is indeed lower than in the meeting room. The room gain in the anechoic chamber was 0 dB in the original study by definition and it is 0.01 dB by the present method described here. In general, the room gain values are lower than in the original study ($\Delta G_{RG} > 0$ in all cases), a fact that has been already reported. None of the room gain values was higher than 0.5 dB. The voice support has a greater dynamic range and might be more suitable for use in architectural acoustics. However, in anechoic rooms, $ST_V \rightarrow −\infty$, and the finite values measured under these conditions must be treated carefully.

IV. REVISED EMPIRICAL MODELS

The new room gain values differ considerably from the original values. In order to enable reliable comparison with future studies, the empirical model relating voice power level as a function of the room gain has to be recomputed. The relative voice power level ($\Delta L_W$) is defined as the difference between the overall $L_W$ in a certain room and the overall $L_W$ measured in the anechoic room. A simplified linear model of only one explanatory variable is preferred

$$\Delta L_W [\text{dB}] = 0.5 − 13.5 \times G_{RG}.$$  

The model predicts a decrease in the expected voice power level with increasing room gain ($R^2 = 0.83, p = 0.01$). Alternatively, rooms with low room gain demand higher vocal intensity from talkers. The measured values, and the regression model (12), are shown in Fig. 2. A two-variable model, similar to the one proposed in Brunskog et al., which describes the relative voice power level as a function of the room gain and the logarithm of the volume, is not significant at the 5% level ($R^2 = 0.83, p = 0.07$) and shows marginal or no influence of the logarithm of the volume on the voice levels.

Figure 3 shows the relative values of voice power level measured by Brunskog et al., versus the voice support. The critical dependence of $ST_V$ value on the measurement SNR in the anechoic chamber suggests that voice level does not change much for very negative values of $ST_V$, also shown with the transformed regression model using the room gain (dotted curve in Fig. 3). A linear dependence of $\Delta L_W$ and $ST_V$ for all the conditions studied is not a good approximation. This approximation does not exclude the possibility of modeling a linear dependence between $L_W$ and $ST_V$ in a limited range of

FIG. 2. Relative $L_W$ produced by talkers in the study by Brunskog et al., as a function of the room gain. The reference $L_W$ is the average overall $L_W$ measured by Brunskog et al., in the anechoic chamber.
STV, as has been done in recent studies, while approaching an asymptotic LW value for very negative STV (dashed line in Fig. 3). Excluding the measurement in the anechoic chamber, the best linear model (solid line in Fig. 3) is

\[ \Delta L_W [\text{dB}] = -13 - 0.78 \times STV. \]  

The accuracy of the predictions decreases with this parameter ($R^2 = 0.66$, $p = 0.09$). It would not be wise to conclude that the voice support is less valid than the room gain to describe the changes in voice level due to the acoustic conditions perceived by the talker. More conditions are needed to assess the robustness of room gain and voice support as explanatory variables of voice level variations due to changes in the auditory perception of one’s own voice elicited by the room.

Vocal effort with changing talker-to-listener distance in different acoustic environments

David Pelegrín-García, \textsuperscript{a}) Bertrand Smits, Jonas Brunskog, and Cheol-Ho Jeong
Acoustic Technology, Department of Electrical Engineering, Technical University of Denmark,
Kongens Lyngby DK-2800, Denmark

(Received 11 February 2010; revised 16 October 2010; accepted 14 January 2011)

Talkers adjust their vocal effort to communicate at different distances, aiming to compensate for the sound propagation losses. The present paper studies the influence of four acoustically different rooms on the speech produced by 13 male talkers addressing a listener at four distances. Talkers raised their vocal intensity by between 1.3 and 2.2 dB per double distance to the listener and lowered it as a linear function of the quantity “room gain” at a rate of −3.6 dB/db. There were also significant variations in the mean fundamental frequency, both across distance (3.8 Hz per double distance) and among environments (4.3 Hz), and in the long-term standard deviation of the fundamental frequency among rooms (4 Hz). In the most uncomfortable rooms to speak in, talkers prolonged the voiced segments of the speech they produced, either as a side-effect of increased vocal intensity or in order to compensate for a decrease in speech intelligibility. © 2011 Acoustical Society of America. [DOI: 10.1121/1.3552881]

PACS number(s): 43.55.Hy, 43.70.Mn [NX] Pages: 1981–1990

I. INTRODUCTION

In face-to-face communication, a talker makes a decision about the desired vocal output based on the given communication scenario. Some factors affecting this decision are the intention of the talker (dialog, discipline, rebuke…), the distance between talker and listener, and special requirements of the listener, due to hearing impairment or language disorders. Once the decision is made, the talker starts to speak and uses a series of feedback mechanisms (auditory, tactile, proprioceptive, and internal) to grant that the actual vocal output matches the desired vocal output.

Speaking in various rooms leads to different experiences or sensations for a talker, due to changes in auditory feedback. The vocal effort required for communicating with a listener at different distances changes with room acoustic conditions, as does also the feeling of vocal comfort. One should differentiate between the concepts of vocal effort and vocal comfort. Vocal effort, according to Traunmüller and Eriksson,\textsuperscript{2} is a physiological magnitude different from vocal intensity, which accounts for the changes in voice production required for the communication at different distances. This definition of vocal effort can be extended to also include the changes in voice production induced by noise or the physical environment. These changes include vocal intensity, fundamental frequency (F\textsubscript{0}), vowel duration, and the spectral distribution of speech. Vocal comfort, according to Titze,\textsuperscript{3} is a psychological magnitude determined by those aspects that reduce the vocal effort. Vocal comfort reflects the self-perception of the vocal effort by the feedback mechanisms listed above.

The maximization of vocal comfort should be a priority in situations of very high vocal demands, which are hazardous for the vocal health, such as teaching environments. A recent study revealed that around 13\% of teachers suffer from voice problems.\textsuperscript{4} Indeed, the prevalence of voice problems among teachers is much higher than it should be, compared to their representation in overall population.\textsuperscript{5–7} Vilkman\textsuperscript{8} points out “bad classroom acoustics” as one of the hazards for voice health from the testimonies of teachers who had suffered from voice disorders. These disorders are related, in many cases, to the intensive use of the voice as an occupational tool.

To characterize the amount of voice use, and to estimate the risk of suffering from voice problems, Titze \textit{et al.}\textsuperscript{9} introduced a set of measures of the accumulated exposure of vocal fold vibration, called vocal doses. The vocal doses are calculated from the phonation time, F\textsubscript{0}, and the vocal fold vibration amplitude. In the present work, the variations of vocal intensity (as a rough estimate of the vocal fold vibration amplitude), F\textsubscript{0}, and the phonation time are reported without going further into a detailed risk analysis, leaving this task to future studies and more advanced analytical models. As in the study by Rantala \textit{et al.},\textsuperscript{10} both the mean and the standard deviation of F\textsubscript{0} are measured as indicators of vocal effort.

Although bad classroom acoustics might be hazardous for voice health, only a few works have attempted to relate classroom acoustics to voice production. Hodgson \textit{et al.}\textsuperscript{11} suggested a simple empirical prediction model to calculate average voice levels used by teachers in university lecture rooms, depending on individual factors, acoustical characteristics of the room, and student activity noise. Brunskog \textit{et al.}\textsuperscript{12} found that the average vocal intensity used by teachers in different classrooms is closely related to the amplification of the room on the talker’s perceived own voice (defined as “room gain”). From this study, it appears that teachers speak louder in rooms with a low room gain and softer in rooms with a high room gain, at a rate of −13.5 dB/dB (decibels of voice level per decibels of room gain).\textsuperscript{13} However, none of the two previous studies took into account the distance

\textsuperscript{a)Author to whom correspondence should be addressed. Electronic mail: dp@elektro.dtu.dk}
between teachers and students, which could explain by itself some of the changes in voice level. From a different perspective, Kob et al. found that teachers with voice disorders were more affected by unfavorable classroom acoustics than their healthy colleagues.

In a more general communication context, several investigations have analyzed the vocal intensity used by a talker to address a listener located at different distances. One general finding is that the vocal intensity is approximately proportional to the logarithm of the distance. The slope of this relationship is in this paper referred to as the compensation rate (in decibels/double distance), meaning the variation in voice level (in decibels) each time that the distance to the listener is doubled (double distance). Warren found compensation rates of 6 dB/dd when talkers produced a sustained vocalization (/a/) addressing listeners at different distances, suggesting that talkers had a tacit knowledge of the attenuation of sound with the distance. However, a sound attenuation of 6 dB/dd is only found in free-field or very close to the source. Warren did not provide information on the experimental acoustic surroundings. Michael et al. showed that the speech material (natural speech or bare vocalizations) influenced the compensation rates and found lower values than Warren. 2.4 dB/dd for vocalizations and 1.3 dB/dd for natural speech. Healey et al. obtained compensation rates in a range between 4.5 and 5 dB/dd when the task was to read a text aloud to a listener at different distances. Liénard and Di Benedetto found an average compensation rate of 2.6 dB/dd in a distance range from 0.4 to 6 m using vocalizations. Traunmüller and Eriksson carried out their experiments with distances ranging from 0.3 to 187.5 m to elicit larger changes in vocal effort, finding a compensation rate of 3.7 dB/dd with spoken sentences. In general, there is a substantial disagreement among the results of different studies.

Each of the previous experiments analyzing voice production with different communication distances was carried out in only one acoustic environment. Michael et al. pointed out that unexplained differences among experimental results might be ascribed to the effect of different acoustic environments, because the attenuation of sound pressure level (SPL) with distance depends on the room acoustic conditions. Zahorik and Kelly investigated how talkers varied their vocal intensity to compensate for the attenuation of sound with distance in two acoustically different environments (one indoor and one outdoor), when they were instructed to provide a constant SPL at the listener position. When uttering a sustained /a/, the talkers provided an almost uniform SPL at each of the listener positions, which indicated that talkers had a sophisticated knowledge of physical sound propagation properties. The measured compensation rates laid between 1.8 dB/dd for an indoor environment and 6.4 dB/dd for an outdoor environment.

In addition, some of the studies investigated further indicators of vocal effort at different communication distances. Liénard and Di Benedetto also found a positive correlation between vocal intensity and F0 and significant spectral changes in vowels. Traunmüller and Eriksson observed that the duration of vocalic segments increased with communication distance, and thus, with vocal effort.

In summary, there have been many studies reporting vocal intensity at different communication distances, as well as other descriptors of vocal effort: F0 and vowel duration. Only one study analyzed the additional effect of the acoustic environment on the vocal intensity, although the instruction—provide a constant SPL at the listener position—and the speech material—vocalizations—were not representative of a normal communication scenario. The aim of the present study is to analyze the effect of the acoustical environment on the natural speech produced by talkers at different communication distances in the absence of background noise, reporting the parameters which might be relevant for the vocal comfort and for assessing the risks for vocal health.

II. EXPERIMENTAL METHOD

The speech from 13 talkers speaking to one listener at four different distances in four different rooms was recorded. The speech signals were processed to calculate measures of vocal intensity, F0, and the relative duration of the phonated segments.

A. Subjects

Thirteen male talkers participated in the experiment as talkers. Two of the talkers were acting as listeners and experimenters at different times. All 13 subjects had ages between 23 and 40 yr and had neither hearing and visual impairments nor vocal disorder. None of the subjects were native English speaker, but nevertheless all of them used English as the spoken language during the tests.

B. Instruction

Before the start of the tests, the listener/experimenter explained the instructions verbally to each talker at a close distance. The talkers were given a map that contained roughly a dozen of labeled items (e.g., “diamond mine,” “fast flowing river,” and “desert”), starting and ending point marks, and a path connecting these two points. They were instructed to describe the route between the starting point and the finish point, indicating the items along the path (e.g., “go to the west until you find the harbor”), while trying to enable eye-contact with the talker. There were 16 maps in total, and a different map was used at each condition. The order of the maps was randomized differently for each subject. These maps have been used extensively in previous research to obtain a dialog-based speech corpus. The object of using maps was evoking natural speech from the talkers in a very specific context and mode of communication. An alternative method for obtaining natural speech could have been instructing talkers to speak freely. However, there would have been different modes of communication and contexts among subjects, which would have introduced higher variability in the data.

After explaining the task to the talker, the listener stood at different positions and indicated the talker non-verbally when to start talking. The listener gave no feedback to the talker, either verbally or non-verbally, about the voice level perceived at his position.
The reverberation time $T_{30}$ was measured according to ISO-3382,\textsuperscript{21} using a dodecahedron loudspeaker as an omnidirectional sound source and a 1/2 in. microphone, Bruel & Kjær (B&K) type 4192 (Bruel & Kjær Sound & Vibration Measurement A/S, Nærum, Denmark). The measurements were carried out with DIRAC,\textsuperscript{22} using an exponential sweep as the excitation signal. The $T_{30}$ obtained from the impulse response using Schroeder’s method\textsuperscript{23} and averaging the measurements in the one-octave bands with a sound level meter, B&K type 2250, at each environment of four distances (1.5, 3, 6, and 12 m) and four different environments (an anechoic chamber, a lecture hall, a long, narrow corridor, and a reverberation room). The environments were chosen so as to represent a wide range of room acoustic conditions, while being large enough to allow distances between talker and listener of up to 12 m. However, not all of these rooms were representative of everyday environments. The order of the rooms was randomized for each subject, but the distances from talker-to-listener were always chosen from closest to furthest. Talker and listener stood further than 1 m from the walls and faced each other.

The volume $V$, reverberation time $T_{30}$, room gain $G_{RG}$, speech transmission index (STI) between talker’s mouth and ears, and A-weighted background noise levels $L_{N,Aeq}$ measured in the rooms are shown in Table I. Possible noise sources contributing to the reported levels are ventilation systems, traffic, and the activity in neighboring areas. All the measured background noise levels were below 45 dB(A) so, according to Lazarus,\textsuperscript{25} the produced voice levels were not affected by the noise.

The STI was derived with the AURORA software suite\textsuperscript{24} from the same mouth-to-ears impulse responses used for the $G_{RG}$ measurements and ignoring the effect of background noise. The values resulting from averaging three repetitions and the two channels (left and right) at each environment are shown on Table I. One should note that the STI parameter was not originally intended to explain the transmission of speech between the mouth and the ears of a talker, as in this case, but to characterize the transmission channel between talker and listener. The STI values presented here are used only as rough indicators of the perceived degradation in one’s own voice due to reverberation and ignoring completely the bone-conducted component of one’s own voice.

### 3. Background noise level

The A-weighted, 20-s equivalent background noise levels ($L_{N,Aeq}$) were measured in the empty rooms using a sound level meter, B&K type 2250. The results from averaging the measurements across four positions in each room are shown in Table I. Possible noise sources contributing to the reported levels are ventilation systems, traffic, and the activity in neighboring areas.

### 5. Speech sound level

The speech sound level\textsuperscript{26} $S$ is defined as the difference between the SPL $L_p$ produced by a source with human voice radiation characteristics at a certain position and the level $L_{ref}$ produced by the same source at 10 m in free-field, averaged over all directions in space,

$$S = L_p - L_{ref}.$$  

A directive loudspeaker JBL Control One (JBL Professional, Northridge CA) was used as the sound source and was placed at the talker position, with the edge of the low frequency driver at a height of 165 cm above the floor and pointing toward the listener. The SPL $L_p$ produced by the loudspeaker reproducing pink noise was analyzed in one-octave bands with a sound level meter, B&K type 2250, at the listener position for each of the four distances in each room.

The reference SPL $L_{ref}$ was calculated as the average of 13 measurements in an anechoic chamber with a distance of 10 m between the sound level meter and the loudspeaker. For each measurement, the loudspeaker was turned at steps of $15^\circ$ from $0^\circ$ to $180^\circ$ and reproduced the same pink noise signal with the same gain settings as used for the measurement of $L_p$.

The resulting $S$, as a function of distance, averaged across the one-octave mid-frequency bands of $500$ Hz and $1$ kHz, is presented in Fig. 1.

### Table I. Physical volume, reverberation time, room gain, STI (mouth-to-ears), and A-weighted background noise level measured in the four environments: anechoic chamber, lecture hall, corridor, and reverberation room.

<table>
<thead>
<tr>
<th></th>
<th>$V$ [m$^3$]</th>
<th>$T_{30}$ [s]</th>
<th>$G_{RG}$ [dB]</th>
<th>STI</th>
<th>$L_{N,Aeq}$ [dB]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anechoic room</td>
<td>1000</td>
<td>0.04</td>
<td>0.01</td>
<td>1.00</td>
<td>&lt;20</td>
</tr>
<tr>
<td>Lecture hall</td>
<td>1174</td>
<td>1.88</td>
<td>0.16</td>
<td>0.93</td>
<td>28.2</td>
</tr>
<tr>
<td>Corridor</td>
<td>410</td>
<td>2.34</td>
<td>0.65</td>
<td>0.83</td>
<td>37.7</td>
</tr>
<tr>
<td>Reverberation room</td>
<td>500</td>
<td>5.38</td>
<td>0.77</td>
<td>0.67</td>
<td>20.6</td>
</tr>
</tbody>
</table>
D. Processing of the voice recordings

The acoustic speech signal was picked up with a DPA 4066 headworn microphone (DPA Microphones A/S, Allerød, Denmark), placed on the talker’s cheek at a distance of 6 cm from the lips’ edge. The signal was recorded with a Sound Devices 722 digital recorder (Sound Devices, LLC, Reedsburg, WI) in 24 bits/44.1 kHz pulsed-code modulation (PCM) format and later processed with MATLAB. The length of the recordings varied between 1 and 2 min, depending on the map and the talker.

1. Voice power level

Vocal intensity is related to the strength of the speech sounds. There are many ways to represent this magnitude, e.g., on-axis SPL at different distances in free-field, sound power level \( (L_W) \), or vibration amplitude of the vocal folds. Among these parameters, the sound power level appears to be the most appropriate one to characterize the total sound radiation from a source. Indeed, it is possible to determine the sound power level if the on-axis SPL in free-field conditions and the directivity of the speaker are known. Following the works of Hodgson et al.\(^{11}\) and Brunskog et al.\(^{12}\), the sound power level was chosen as the main index of vocal intensity and is also referred to as voice power level.

To determine the voice power level of the recordings, the equivalent SPL in the one-octave bands between 125 Hz and 4 kHz was first calculated. A correction factor due to the increase of SPL at the headworn microphone in the different rooms was applied (see values in Table II). The correction factor was measured by analyzing the SPL produced by the HATS, reproducing pink noise with a constant sound power level in the different rooms, at the headworn microphone, which was placed on the HATS. The SPL readings from the anechoic chamber were subtracted to the readings in each room. The difference between the corrected SPL at the headworn microphone and the voice power level was determined by performing sound power measurements in a reverberation room in a similar way as described by Brunskog et al.\(^{12}\). However, instead of using a dummy head (as in Brunskog et al.), the speech of six different talkers, one by one, was recorded simultaneously using a headworn microphone DPA 4066 and a 1/2 in. microphone, B&K type 4192, positioned in the far field, where the sound field is assumed to be diffuse. The difference between the mean corrected SPL measured at the headworn microphone and the voice power level as a function of frequency is shown in Fig. 2.

2. Fundamental frequency

\( F_0 \) was extracted from the recordings with the application WAVESURFER\(^{27}\) using the entropic signal processing system method at intervals of 10 ms. Taking a sequence with the \( F_0 \) values of the voiced segments (the only segments for which the algorithm gave an estimation of \( F_0 \)), the mean (noted as \( \bar{F}_0 \)) and the standard deviation (noted as \( \sigma_{F_0} \)) were calculated.

3. Phonation time ratio (PTR)

Due to the large variations in the length of speech material among subjects and conditions, the absolute phonation time is not reported, but the ratio of the phonation time \( t_P \) to the total duration of running speech \( t_S \) in each recording, referred to as PTR. The calculation procedure is shown in Fig. 3. First, the original speech signal [Fig. 3(a)] is processed to obtain the running speech signal [Fig. 3(b)]. Then, this signal is split into \( N \) non-overlapping frames or segments of a duration \( t_F = 10 \) ms [Fig. 3(c)]. In the \( i \)th frame, the logical variable \( k_i \) (\( k_i = 0 \) if the segment is unvoiced; \( k_i = 1 \) if the segment is voiced) is calculated.
The actual models were built as simplifications of the “full model.” First, the significance of the interaction (room-dependent slope $b_k$) was tested by means of likelihood ratio tests (using the function `anova` in R), comparing the outcomes of the full model and a reduced model without the interaction (constant slope $b$). If the full model was significantly better than the reduced model, the first one was kept. Otherwise, the reduced model was used. Another test for the suitability of random slopes was made by comparing the full model to another one with fixed slopes by means of a likelihood ratio test. In the same way, if the model with random slopes was significantly better than the one with fixed slopes, the first one was chosen. The suitability of including the basic variables (room and distance) was assessed by comparing the chosen model from the previous tests to a reduced version that only contained one variable (room or distance) with likelihood ratio tests. However, all the parameters showed dependence on the room and the distance. The models did not include a random effect for the room due to the subject.

The $p$-values for the overall models were calculated by means of likelihood ratio tests comparing the fit of the chosen model to the fit of a reduced model which only contained the random intercept due to the effect of the room $k$ on the response variable. The random part is also referred to as intersubject variability. The residual or unexplained variation $e_{ijk}$ is also regarded as a random effect. The standard deviations of the random effects $\sigma_z, \sigma_\beta,$ and $\sigma_e$ are respectively.

The $p$-values associated to each predictor and the standard deviations of the random effects were obtained with the function `pvals.fnc` (with `MCMC = T`) of the library `languageR` (Ref. 31) in R, which makes use of the Markov Chain Monte Carlo (MCMC) sampling method.

**E. Statistical method**

For each parameter ($L_0$, $\bar{F}_0$, $\sigma_{F_P}$, and PTR), a linear mixed model was built from a total of 208 observations (13 subjects $\times$ 4 distances $\times$ 4 rooms), using the `lme4` method in the library `lme4` of the statistical software R. The “full model” included the logarithm of the distance as a covariate and the acoustic environment (or room) as a factor and the interaction between the distance and the room. In the present paper, the mixed model for a response variable which depends on the $i$th subject, the $j$th distance $d_j$, and the $k$th room is presented in the form

$$y_{ijk} = a_k + b_k + (b_k + \beta_i) \times \log_2(d_j/1.5) + e_{ijk}.$$  

The fixed-effects are written on roman characters ($a_k$ and $b_k$) and the random effects are written on greek characters ($\alpha_i$, $\beta_i$, and $e_{ijk}$). The random effects are stochastic variables normally distributed with zero mean. The distance dependence is contained in the parameters $b_k$ and $\beta_i$ (fixed slope and random slope, respectively). On the fixed part, the subscript $k$ indicates an interaction between room and distance. If there is no interaction, $b_k$ becomes a constant $b$. The presence of $\beta_i$ indicates that the dependence of the response variable $y$ on the distance $d$ is different for each subject. The intercept ($a_k + \alpha_i$) adjusts the overall value of $y$, and it has a fixed part $a_k$ and a random part $\alpha_i$. The fixed intercept contains the effect of the room $k$ on the response variable. The random part is also referred to as intersubject variability. The residual or unexplained variation $e_{ijk}$ is also regarded as a random effect. The standard deviations of the random effects $\sigma_z, \sigma_\beta,$ and $\sigma_e$ are respectively.

The statistical method included the logarithm of the distance as a covariate and the acoustic environment (or room) as a factor and the interaction between the distance and the room. In the present paper, the mixed model for a response variable which depends on the $i$th subject, the $j$th distance $d_j$, and the $k$th room is presented in the form

$$PT = \frac{\text{Phonation time}}{\text{Running speech time}} = \frac{\sum_{i=1}^{N} t_F}{t_S}, \quad N = \left\lfloor \frac{t_S}{t_F} \right\rfloor. $$

The floor operator $\lfloor \cdot \rfloor$ results in the closest integer not larger than the operand.
TABLE III. Fixed and random effects included in the mixed models. The fixed-effects are characterized for the intercepts $a$ and slopes $b$, whereas the random effects have zero mean and only their standard deviation is shown. Abbreviations are used instead of the complete name of the rooms: ACH for the anechoic room, LH for the lecture hall, COR for the corridor, and REV for the reverberation room. Note that the $b$ values for $F_0$, $\sigma_{F_0}$, and PTR are independent of the room.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>ACH</th>
<th>LH</th>
<th>COR</th>
<th>REV</th>
<th>ACH</th>
<th>LH</th>
<th>COR</th>
<th>REV</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L_W$ [dB]</td>
<td>56.8</td>
<td>56.0</td>
<td>54.8</td>
<td>56.2</td>
<td>2.2</td>
<td>2.0</td>
<td>1.9</td>
<td>1.3</td>
</tr>
<tr>
<td>$F_0$ [Hz]</td>
<td>123.6</td>
<td>120.1</td>
<td>119.8</td>
<td>119.3</td>
<td>3.8</td>
<td>0.63</td>
<td>0.026</td>
<td>1.33</td>
</tr>
<tr>
<td>$\sigma_{F_0}$ [Hz]</td>
<td>23.2</td>
<td>22.0</td>
<td>20.6</td>
<td>19.2</td>
<td>5.22</td>
<td>1.29</td>
<td>2.77</td>
<td></td>
</tr>
<tr>
<td>PTR</td>
<td>0.65</td>
<td>0.55</td>
<td>0.56</td>
<td>0.67</td>
<td>0.059</td>
<td>—</td>
<td>0.062</td>
<td></td>
</tr>
</tbody>
</table>

The choice of mixed models has the following basis: a considerable amount of the variance in the observations is due to the intersubject differences (which could be revealed with an analysis of variance table), so the subject is regarded as a random effect. Conceptually, it is similar to applying a normalization for each subject or regarding the subject as a factor in traditional statistical modeling.

III. RESULTS AND ANALYSIS

The measurements of $L_W$, $F_0$, $\sigma_{F_0}$, and PTR were used to build four different linear mixed models according to Eq. (3). The coefficients for the intercepts and slopes corresponding to the fixed-effects of the models, together with the standard deviations of the random effects, are presented in Table III. The statistical significance ($p$-value) of the fixed-effects and interactions included in each model, along with the overall significance levels, is shown in Table IV.

A. Voice power level

The measured $L_W$, as a function of the distance and for each of the rooms, averaged across all subjects, is shown in Fig. 4. In the same figure, the lines show the fixed-effects part of the empirical model described in Eq. (3) and Table III. $L_W$ depends almost linearly on the logarithm of the distance (with slopes between 1.3 and 2.2 dB per doubling distance) and changed significantly among rooms (intercepts between 54.8 and 56.8 dB). At each distance, the highest $L_W$ was always measured in the anechoic room. A significant interaction was found between the room and the logarithm of the distance, because the variation of $L_W$ with distance in the reverberation room (1.3 dB per doubling distance) was lower than the variation in the other rooms (1.9 to 2.2 dB per doubling distance). The standard deviation of the intersubject variation was estimated to be 2.7 dB, whereas the individual differences in the variation of $L_W$ with distance had a standard deviation of 0.76 dB per doubling distance.

B. Fundamental frequency

Figure 5 shows the subject-averaged measured $F_0$ (data points) and the corresponding empirical model (lines) described in Eq. (3) and Table III, for the different distances and rooms. $F_0$ changed significantly among rooms (intercepts between 119.3 and 123.6 Hz) and had an almost linear dependence on the logarithm of the distance, with a slope of 3.8 Hz per doubling distance, identical for all the rooms. However, by visual inspection of Fig. 5, in the anechoic and reverberant rooms, there was less variation between the distances of 1.5 and 3 m than at further distances. $F_0$ in the anechoic room was about 4 Hz higher than in the other rooms for all distances. The standard deviation of the intersubject variation was estimated in 16.3 Hz, whereas the individual differences in the variation of $F_0$ with distance had a standard deviation of 2.95 Hz per doubling distance.

The measured $\sigma_{F_0}$, as a function of the distance and for each of the rooms, averaged across all subjects, is shown in Fig. 6. The lines in the figure show the fixed-effects part of the empirical model described in Eq. (3) and Table III. $\sigma_{F_0}$ changed significantly among rooms (intercepts between 19.2

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Intercept</th>
<th>Slope</th>
<th>Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a$ (Intercept)</td>
<td>56.8</td>
<td>56.0</td>
<td>54.8</td>
</tr>
<tr>
<td>$b$ (Slope)</td>
<td>2.2</td>
<td>2.0</td>
<td>1.9</td>
</tr>
<tr>
<td>$\sigma_x$</td>
<td>2.74</td>
<td>0.76</td>
<td>1.33</td>
</tr>
<tr>
<td>$\sigma_y$</td>
<td>16.3</td>
<td>2.95</td>
<td>3.6</td>
</tr>
<tr>
<td>$\sigma_z$</td>
<td>5.22</td>
<td>1.29</td>
<td>2.77</td>
</tr>
<tr>
<td>PTR</td>
<td>0.059</td>
<td>—</td>
<td>0.062</td>
</tr>
</tbody>
</table>

TABLE IV. Statistical significance and $p$-values of the fixed-effects and interactions considered in the empirical models and overall significance of the models. NS: Non-significant.
and 23.2 Hz) and had a weak linear dependence on the logarithm of the distance, with a slope of 0.63 Hz per doubling distance, equal among the rooms. The standard deviation of the intersubject variation was estimated in 5.22 Hz, whereas the individual differences in the variation of $\sigma_{F_0}$ with distance had a standard deviation of 1.29 Hz per doubling distance. The latter value is larger than the fixed-effect slope (0.63 Hz) which means that, for a number of subjects, $\sigma_{F_0}$ decreased with distance. This is the reason for the low statistical significance of the $\sigma_{F_0}$ dependence with the logarithm of the distance shown on Table IV. Therefore, the amount of $\sigma_{F_0}$ change as a function of distance was mainly an individual factor.

### C. PTR

The measured PTR, as a function of the distance and for each of the rooms, averaged across all subjects, is shown in Fig. 7. In the same figure, the lines show the fixed-effects part of the empirical model described in Eq. (3) and Table III. PTR had a weak linear dependence on the logarithm of the distance (with a slope of 0.026 per doubling distance, equal for all rooms) and changed significantly among rooms, especially between two groups: one formed by the anechoic room and the reverberation room (intercepts 0.65 and 0.67) and a second group formed by the lecture hall and the corridor (intercepts 0.55 and 0.56). The standard deviation of the intersubject variation was estimated in 0.059. The change in PTR with distance was not significantly different among subjects, so the model does not include a random slope.

### D. Subjective impressions

The talkers expressed their opinions verbally about the experience of talking in the different rooms. One general comment was that the anechoic chamber was an unnatural place to speak in, due to the lack of sound reflections, and that they felt moved to raise their vocal intensity to make themselves heard at the listener location, and for this reason, it was not a comfortable environment for talking. The reverberation room was very unpleasant for speaking, due to the excessive reverberation. Talkers admitted that they had to modify their speech strategy to compensate for the poor acoustic conditions. A few of the subjects preferred overall the corridor, due to the sensation of support or being helped by the room to reach longer distances without having to increase their voice level too much, although they pointed out some acoustical deficiencies like a noticeable echo. Most of the subjects preferred the lecture hall for speaking. However, they admitted that it was demanding to talk at the longest distance (12 m). Many subjects commented that the acoustic conditions of the experimental rooms were not the desirable ones in rooms for speech.

### IV. DISCUSSION

Figures 4 to 7 show the variation of the measured parameters ($L_W$, $F_0$, $\sigma_{F_0}$, and PTR) with distance and across environments. As all of the measured parameters indeed have variation with distance and acoustic environment, they are potential indicators of vocal effort.

The measurements shown in Fig. 4 reveal that the average variations of $L_W$ when the distance increases from 1.5 to 12 m are in the range between 3.9 dB in the reverberation room and 6.6 dB in the anechoic room. These variations are mainly the consequence of a conscious decision of the talker to raise the voice level as a response to a change in communication distance. However, the fact that the compensation

![FIG. 5. Average mean fundamental frequency used by talkers at different distances to the listener. The lines show the predictions of the empirical model.](image1)

![FIG. 6. Average long-term standard deviation of the fundamental frequency used by talkers at different distances to the listener. The lines show the predictions of the empirical model.](image2)

![FIG. 7. Average PTR (relative appearance of voiced segments in running speech) used by talkers at different distances to the listener. The lines show the predictions of the empirical model.](image3)
rates differ among rooms shows the influence of auditory feedback in voice level adjustment. Furthermore, the effect of room on $L_W$ varies between 2 dB at 1.5 m and 3.3 dB at 12 m. These values are smaller but comparable to the effect of distance on $L_W$. Thus, the perception of one’s own voice via reflections in the room boundaries is important for voice level regulation, together with the direct air transmission and the bone-conducted components, as Siegel and Pick\cite{32} stated.

Brunskog et al. used $G_{RG}$ as a metric to quantify the importance of the reflected sound from one’s own voice. This measure is indeed a measure of sidetone (one’s own voice reaching the ears) amplification. Taking the subject-averaged $L_W$ values measured at 6 m, a distance which is representative of a lecturing scenario, the least squares regression model using $G_{RG}$ as a predictor is

$$L_{W,6} = 61.5 - 3.56 \times G_{RG}. \quad (4)$$

The $R^2$ for this regression model is 0.82, whereas the $p$-value is 0.09. The $L_W$ values, with the regression line (4), are compared to the results of Brunskog et al.\cite{12,13} in Fig. 8. The slope of the regression line in the current measurements is much lower than the slope obtained by Brunskog et al. (−3.6 dB/db vs −13.5 dB/db). The difference between slopes might be explained by the fact that the distance was not taken into account by Brunskog et al. In their study, the rooms with high $G_{RG}$ values were small rooms where the listeners stood close to the talker whereas the rooms with low room gain were larger and the listeners stood far from the talker. Thus, there is an unwanted correlation between the room gain and the distance, due to the experimental design, but which is found in typical real rooms. The model from Brunskog et al. predicts $L_W$ in a general situation with varying distance to the listeners, but the model (4) accounts for the variation due exclusively to changes in auditory feedback.

As in some studies of sidetone amplification,\cite{33} $L_W$ decreases with increasing sidetone amplification (estimated by $G_{RG}$). However, there are two differences between these studies and the present study. One is the range of $L_W$ variation and the second is the magnitude of the effect. In the present study, talkers raised $L_W$ by 3.2 dB on average while speaking in the anechoic room at a distance of 12 m, compared to the reverberant room. In other studies of voice production with altered sidetone, variations in voice level of up to 20 dB were reported. In these studies, the sidetone was altered by inducing temporary hearing loss on the subjects, thus decreasing all components of sidetone (direct, reflected, and bone-conducted sound) or attenuating the airborne sound while bone conduction is preserved. The significantly different ranges of voice level variation obtained in the previous studies (up to 20 dB) and in this study (approximately 3.2 dB by the effect of room) might be due to the fact that only the reflected component was changed in this study, while the direct and bone-conducted components of the talker’s own voice were kept unchanged. Therefore, the overall sidetone variations were much smaller than in the other studies. The magnitude of the effect on traditional sidetone compensation was in the range between −0.25 and −0.57 dB/db, whereas in the present study the magnitude of the effect was −3.6 dB/db, as can be seen in Eq. (4). These differences could be explained by two alternative hypotheses. The first is that the changes in $L_W$ are purely due to the Lombard effect and that the room reflections alter the loudness of one’s own voice to a greater extent than indicated by the single figure $G_{RG}$. The second is that there are additional psychological attributes related to room perception affecting the voice regulation at a cognitive level, through internal feedback mechanisms.

The measured compensation rates for $L_W$ due to changes in distance between talker and listener were between 1.3 db/dd in the reverberation room and 2.2 db/dd in the anechoic chamber. These compensation rates are much lower than the ones obtained by Warren,\cite{15} Healey et al.,\cite{17} and Traunmüller and Eriksson.\cite{2} However, they are closer to other studies\cite{16,18} and especially close to the 1.8 db/dd measured indoor by Zahorik and Kelly.\cite{19} Differences from the previous studies might arise from the selection of subjects or different instruction. In the present study, there were significant differences in vocal behavior among subjects, indicated by the random slope effect in Table III, which predicts a standard deviation of 0.76 db/dd over the fixed slopes 1.3 to 2.2 db/dd. In any case, the individual compensation rates were not as large as 6 db/dd.\cite{15,19} In addition, natural speech was evoked in the present experiment by means of the map task, which resulted in a different vocal behavior compared to the experimental setup in previous studies.

In this study, the compensation rate of $L_W$ was 0.09. The $p$-value for this regression model is 0.25, whereas in the present study the magnitude of the effect was −3.6 dB/db, as can be seen in Eq. (4). These differences could be explained by two alternative hypotheses. The first is that the changes in $L_W$ are purely due to the Lombard effect and that the room reflections alter the loudness of one’s own voice to a greater extent than indicated by the single figure $G_{RG}$. The second is that there are additional psychological attributes related to room perception affecting the voice regulation at a cognitive level, through internal feedback mechanisms.

The measured compensation rates for $L_W$ due to changes in distance between talker and listener were between 1.3 db/dd in the reverberation room and 2.2 db/dd in the anechoic chamber. These compensation rates are much lower than the ones obtained by Warren,\cite{15} Healey et al.,\cite{17} and Traunmüller and Eriksson.\cite{2} However, they are closer to other studies\cite{16,18} and especially close to the 1.8 db/dd measured indoor by Zahorik and Kelly.\cite{19} Differences from the previous studies might arise from the selection of subjects or different instruction. In the present study, there were significant differences in vocal behavior among subjects, indicated by the random slope effect in Table III, which predicts a standard deviation of 0.76 db/dd over the fixed slopes 1.3 to 2.2 db/dd. In any case, the individual compensation rates were not as large as 6 db/dd.\cite{15,19} In addition, natural speech was evoked in the present experiment by means of the map task, which resulted in a different vocal behavior compared to the experimental setup in previous studies.
in lower compensation rates than would be obtained by using short vocalizations, as Michael et al.\textsuperscript{16} stated.

Figure 9 shows the relationship between the $L_W$ produced by the talkers and the sound speech level $S$ at the listener position, which is an alternative representation of the data in Fig. 4. The dashed line in Fig. 9 represents the theoretical $L_W$ values that would keep the SPL constant at the listener position. According to Zahorik and Kelly,\textsuperscript{19} if talkers accurately compensated for the sound propagation losses—providing an almost constant average SPL at the listener position—the expected $L_W$ would lay exactly on top of a line with the same slope as the dashed line, meaning that a talker would lower $L_W$ by 1 dB whenever $S$ increases by 1 dB. The $L_W$ data points in Fig. 9 follow approximately straight lines with different slopes for each room: $-0.4$ dB/dB in the anechoic chamber, $-0.8$ dB/dB in the lecture hall, $-1.1$ dB in the reverberation room, and $-3.8$ dB/dB in the corridor. In the lecture hall and the reverberation room, talkers approximately compensated for sound propagation losses. However, there was an undercompensation in the anechoic chamber, meaning that the SPL produced at the listener position decreased with distance, and an overcompensation in the corridor, where the SPL increased with the distance. Undercompensation appears to take place in rooms with big differences of $S$ between short and long distances, i.e., rooms with dominating direct sound. Overcompensation takes place in rooms where differences in $S$ at short and long distances were small, i.e., rooms with strong reverberant field. Undercompensation and overcompensation were present because the talkers were not explicitly asked to compensate for sound propagation losses, and many of the talkers were not used to talk in the environments of the study. It is presumed that talkers would be able to compensate for sound propagation losses with an explicit instruction and training to get acquainted with the acoustical properties of each room.

Compensation rates have a meaning when the distance between talker and listener is well defined, such as in a face-to-face conversation. In the case of a distributed audience, as in the usual teaching context, the situation is more complex and it is not clear what is the distance estimation of the talker. In that case, according to Brunskog et al.,
\textsuperscript{12,13} talkers apparently adjust their voice levels guided by the room gain or degree of amplification provided by the room at their ears (Fig. 8).

The changes in $F_0$ were similar to those in $L_W$, as both parameters increased linearly with the logarithm of the distance, and it was in the anechoic room where the highest $F_0$ were obtained at each distance. Table III shows that $F_0$ changed 3.8 Hz by doubling the distance and was 4 Hz higher in the anechoic room than in the other rooms. In simplified terms, the extra vocal effort demanded to speak in the anechoic room is comparable to the effect of doubling the distance to the listener in other rooms. However, the changes among other rooms (maximum of 0.8 Hz) were not as important so as to ascribe a significant effect to the room. It seems more likely that the unfamiliarity of talkers with the anechoic room accentuated some changes in speech production too much, which are not observed in everyday rooms. Nevertheless, $F_0$ is an important measure of vocal effort to show that, at long communication distances, the number of vocal fold vibrations (or collisions) increases, which leads to higher vocal doses that might eventually result in vocal fold trauma.

The talkers had the general remark that the anechoic room and the reverberation room were the most uncomfortable environments to speak in. Both environments were the two most extreme rooms in terms of $T_{30}$, STI, and $G_{RG}$, as shown in Table I. The anechoic chamber demanded an increased vocal effort due to lacking support, with a $G_{RG}$ value of 0.01 dB. On the other hand, it was very unpleasant and stressing to speak in the reverberation room, which could be explained by the remarkably lower STI value (only 0.67) corresponding to the transmission between mouth and ears. Talkers’ comments suggest that there is a compromise between STI and $G_{RG}$, in order for rooms to be comfortable. The poor vocal comfort rating for the reverberation room cannot be explained by the measured $L_W$ or $F_0$, as the $L_W$ and $F_0$ in this room were not higher than the values measured in the lecture hall and the corridor, the most preferred rooms. This observation supports the idea that the concepts of vocal effort and comfort are not exactly opposite.

As shown in Fig. 6 and Table III, the model predicted significant differences in $\sigma_{F_0}$ among the environments for all distances. The highest $\sigma_{F_0}$ was found in the anechoic room, followed by those in the lecture hall, the corridor, and the reverberation room, in reverse order to the reverberation times: the reverberation room, the corridor, the lecture hall, and the anechoic chamber (in decreasing order), or in the same order as the STI. According to this observation, speech produced in acoustically live rooms is more monotonous (meaning low variability in $F_0$) than in acoustically dry rooms. The extreme values of $\sigma_{F_0}$ were obtained in the least preferred rooms. The highest $\sigma_{F_0}$ in the anechoic room might be an indication of increased vocal demands (increased $L_W$ and $F_0$), whereas the low $\sigma_{F_0}$ in the reverberant room might be an observable feature of the speech produced under low STI conditions. However, this assertion needs to be proved in a broader range of acoustic conditions.

In Fig. 7, the average PTR was remarkably different between two groups of environments and correlated well with the subjective impressions of talkers regarding vocal comfort. The highest PTR values were measured in the most uncomfortable rooms (0.67 in the reverberation room and 0.65 in the anechoic room), whereas the PTR in the other two rooms was significantly lower (0.55 in the lecture hall and 0.56 in the corridor). The increased voice levels or vocal efforts explain the high values obtained for the anechoic chamber, as Lienard and Di Benedetto\textsuperscript{18} also reported. However, the high PTR obtained in the reverberation room might be due to the adaptation of the talker to the environment. It seems that talkers tried to improve the speech intelligibility in such a reverberant environment by separating the consonant segments of their speech, resulting in longer vocalic segments.

\section*{V. CONCLUSIONS}

The present paper studies the changes in different speech parameters (voice power level, fundamental frequency, PTR) describing vocal effort when talkers addressed a single listener at different distances under various room
acoustic conditions in the absence of background noise. The main conclusions are as follows:

1. The decision of using a certain voice level depends on the visually perceived distance to the listener and varies between 1.3 and 2.2 dB per double distance to the listener.

2. The room acoustic conditions modify the auditory feedback of the talker’s own voice, inducing significant changes in voice level with an approximately linear dependence on the amplification of the room to one’s own voice, given by the magnitude “room gain,” at a rate of −3.6 dB/dB.

3. The mean fundamental frequency increases with distance at a rate of 3.8 Hz per double distance to the listener and is 4 Hz higher in anechoic conditions.

4. A room that provides vocal comfort requires a compromise between room gain and STI, supporting the voice from a talker but not degrading the perceived speech quality.

5. The standard deviation of the fundamental frequency and the relative duration of voiced segments in a running speech signal might be symptomatic indicators of vocal comfort in a room.

ACKNOWLEDGMENTS

The authors wish to express their gratitude to all the participants in the measurements, to one anonymous reviewer that enabled significant improvements in this work, and to Finn Jacobsen and Torben Poulsen for their positive feedback toward improving the quality of the paper. This research has been partially funded by the Swedish company AFA Försäkring as a part of the project “Speaker’s comfort and voice health in classrooms.”


Equal autophonic level curves under different room acoustics conditions

David Pelegrín-García,a) Oier Fuentes-Mendizabal, Jonas Brunskog, and Cheol-Ho Jeong
Acoustic Technology, Department of Electrical Engineering, Technical University of Denmark,
Kongens Lyngby DK-2800, Denmark

(Received 16 February 2011; revised 13 May 2011; accepted 13 May 2011)

The indirect auditory feedback from one’s own voice arises from sound reflections at the room boundaries or from sound reinforcement systems. The relative variations of indirect auditory feedback are quantified through room acoustic parameters such as the room gain and the voice support, rather than the reverberation time. Fourteen subjects matched the loudness level of their own voice (the autophonic level) to that of a constant and external reference sound, under different synthesized room acoustics conditions. The matching voice levels are used to build a set of equal autophonic level curves. These curves give an indication of the amount of variation in voice level induced by the acoustic environment as a consequence of the sidetone compensation or Lombard effect. In the range of typical rooms for speech, the variations in overall voice level that result in a constant autophonic level are on the order of 2 dB, and more than 3 dB in the 4 kHz octave band. By comparison of these curves with previous studies, it is shown that talkers use acoustic cues other than loudness to adjust their voices when speaking in different rooms. © 2011 Acoustical Society of America.

PACS number(s): 43.55.Hy, 43.70.Mn [NX]

I. INTRODUCTION

The sound that a talker perceives from his own voice-auditory feedback or sidetone-consists of two main components: direct and indirect auditory feedback. The direct auditory feedback can be separated into two other components: airborne sound and bone-conducted sound. These two last components are of the same order of magnitude and are always present for building up the sound of talkers’ own voice, as long as the acoustic path between the mouth and the ears is undisturbed and the talker has normal hearing. However, the bone-conducted component is not constant in level and frequency distribution, but varies with different vocalizations. The indirect auditory feedback is essentially airborne and is generated by the reflections of talkers’ own voice at the room boundaries, or by a sound reinforcement system when it is used to amplify the voice of the talkers.

The loudness with which talkers perceive their own voice is called the autophonic rating. The autophonic rating grows at almost twice the rate of the loudness of external sounds, meaning that the change in voice level (in dB) required to double the autophonic rating is half of the amount required for external sounds in order to double the loudness sensation. The differences between the autophonic scale and the loudness (sone) scale are most likely due to the different sensing mechanisms in hearing one’s own voice and external sounds. The sensation for external sounds is essentially auditory, whereas for one’s own voice, it is also dependent on tactile, proprioceptive, and internal mechanisms.

According to Lane and Tranel, speakers adjust their voices to maintain a speech-to-noise ratio suitable for communication. Some factors affecting the speech-to-noise ratio are linked to the auditory perception, such as noise or alterations in sidetone. Other factors are not linked to the auditory perception, but have a clear influence on the voice levels used, as, for example, the distance between the talker and the listener.

The variation in voice level due to the presence of noise is known as the Lombard effect (see a review in Lane and Tranel). Lane et al. showed that talkers accounted for variations of ambient noise level by varying their voice level at a rate of 0.5 dB/dB (voice/noise). In the same study, Lane et al. found an equivalent rate for the so-called sidetone compensation: talkers lowered their voice by 0.5 dB for each additional dB of gain applied to the sidetone, while talking over an interphone. The variations of sidetone can also be due to a temporary hearing loss; Black found a compensation rate of 0.57 dB/dB hearing level (HL).

In the previous cases, the sidetone was altered by damping the direct auditory feedback, or by reproducing an amplified replica of one’s own voice through a monitoring device which had the effect of a single sound reflection with a level high enough to mask the direct auditory feedback components. In rooms, the sidetone is altered in a substantially different way, because the indirect auditory feedback is built up by a number of reflections arriving at different delays, with different amplitudes, and spectral weightings. These reflections may interact with the direct auditory feedback in a different way from a single delay. There are two room acoustic parameters to measure the sidetone variations caused by a room. The voice support ($ST_V$) is defined as the energy ratio of the indirect ($E_I$) to the airborne-direct ($E_D$) auditory feedback. The room gain ($GRG$) is defined as the ratio of the total airborne auditory feedback ($E_I + E_D$) to the airborne-direct auditory feedback, $^{12}$

$$ST_V = 10 \log \frac{E_I}{E_D},$$

where $E_I$ and $E_D$ are the energies of indirect and direct auditory feedback.

---

a)Author to whom correspondence should be addressed. Electronic mail: dpg@elektro.dtu.dk
II. METHOD

Fourteen subjects (ten men and four women) with ages between 20 and 30 yr, without any known problems with hearing or voice and without previous instruction in vocal training, took part in the experiment. A reference sound-either a tone or a vocalization-at a constant sound pressure level (SPL) was presented, and the test subjects were asked to produce a vocalization (either /a/, /i/, or /u/) with the same loudness as the reference. Each subject produced a total of 60 vocalizations that were stored and analyzed to extract the results.
resonance of the ear canal toward higher frequencies, attenuating the resonance peak due to viscous losses. The IL between 63 Hz and 2 kHz was lower than 1 dB, and the maximum attenuation at higher frequencies was 6 dB. These values were assumed to be acceptable for the present application.

With the custom earphones, the frequency response deviated from a flat response (see Fig. 4). They had a poor low and mid frequency response, with a roll-off below 2 kHz, and remarkable resonance peaks at high frequencies, between 3 kHz and 8 kHz. A minimum phase finite impulse response (FIR) filter of 128 samples was used in order to compensate for the frequency response and achieve a relatively flat frequency response, corresponding to the frequency response of the electrostatic headphones STAX (STAX Ltd.; Miyoshi-machi, Japan) model Lambda. This target frequency response was chosen instead of an ideal flat frequency response after realizing—by means of subjective assessment—that the overall sound quality was better in the first case. The FIR filter was preconvolved with the synthetic impulse responses generated for each experimental condition.

A MATLAB program controlled the experiment, changing the synthetic impulse response loaded by jconvolver and reproducing different messages to the talker, indicating beginning and the end of vocalization periods, and which vowel should be produced.

B. Acoustic conditions

There were nine different synthetic impulse responses or conditions C1 to C9 (plus an additional condition C10, namely, the absence of simulated reflections), which added the indirect auditory feedback of talkers’ own voice to the direct sound and the bone conduction. The acoustic properties of the different conditions are summarized in Table I. The synthetic impulse responses were generated artificially, and it was not their goal to replicate the acoustic conditions of actual environments, but to provide well-defined and adjustable experimental conditions. Each synthetic impulse response was obtained in the following manner. First, a white Gaussian noise signal (of 66150 samples at 44.1 kHz), common to all impulse responses, was generated. This was done in order to have the same reflection pattern or “fine structure” in all responses. An exponential decay was applied to the noise signal. The decay constants were chosen so that the reverberation time of the conditions fell into one of three groups: low (C1 to C3, 0.45 s ≤ T ≤ 0.55 s), medium (C4 to C6, 0.93 s ≤ T ≤ 1.12 s), and high (C7 to C9, 1.40 s ≤ T ≤ 1.65 s). Finally, different gains were applied so that the room gain entered in the categories of low (C1, C4, and C7, 0.07 dB ≤ G_{RG} ≤ 0.19 dB), medium (C2, C5, and C8, 0.31 dB ≤ G_{RG} ≤ 1.68 dB), and high (C3, C6, and C9, 2.95 dB ≤ G_{RG} ≤ 8.63 dB).

<table>
<thead>
<tr>
<th>Condition</th>
<th>T (s)</th>
<th>G_{RG} (dB)</th>
<th>S_{TV} (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>0.55</td>
<td>0.07</td>
<td>-17.9</td>
</tr>
<tr>
<td>C2</td>
<td>0.50</td>
<td>0.31</td>
<td>-11.3</td>
</tr>
<tr>
<td>C3</td>
<td>0.45</td>
<td>2.95</td>
<td>-0.12</td>
</tr>
<tr>
<td>C4</td>
<td>1.12</td>
<td>0.13</td>
<td>-15.2</td>
</tr>
<tr>
<td>C5</td>
<td>1.00</td>
<td>1.03</td>
<td>-5.7</td>
</tr>
<tr>
<td>C6</td>
<td>0.93</td>
<td>6.57</td>
<td>5.5</td>
</tr>
<tr>
<td>C7</td>
<td>1.65</td>
<td>0.19</td>
<td>-13.5</td>
</tr>
<tr>
<td>C8</td>
<td>1.50</td>
<td>1.68</td>
<td>-3.3</td>
</tr>
<tr>
<td>C9</td>
<td>1.40</td>
<td>8.63</td>
<td>8.0</td>
</tr>
<tr>
<td>C10</td>
<td>0.01</td>
<td>0.04</td>
<td>-20.3</td>
</tr>
</tbody>
</table>
The reverberation times were chosen to correspond to usual reverberation times found in rooms for speech (low $T$: classrooms, medium $T$: drama theaters, high $T$: opera houses). The room gain/voice support values were chosen to be representative of real rooms without amplification ($-20 \, \text{dB} \leq S^V \leq -5 \, \text{dB}$), although higher values were also chosen to explore the possible effects of electroacoustic amplification on the voice production and perception.

For the objective measurements, a HATS B&K type 4128 with right ear simulator B&K type 4158 and left ear simulator B&K type 4159 was placed at the talker position in the setup in Fig. 1. The headworn microphone and the earphones were attached to the dummy head as explained in the experimental setup section. The HATS had a mouth simulator and microphones at the ears, so it was possible to measure the impulse response corresponding to the path between the mouth and the ears. The direct sound was generated by direct radiation from the mouth to the ears, whereas the reflections were generated artificially by convolution with a synthetic impulse response and reproduction through the earphones. The mouth-to-ears impulse responses were measured with the MLS module in the 01dB (01dB-Metravib; Limonest Cedex, France) Symphonie system. The backward-integrated energy-time curves of the measured responses C1 to C9, averaged between the left and the right ears, are shown in Fig. 5. The reverberation time was calculated from the slope of these curves, in a decay of at least 10 dB neither influenced by the noise floor nor the direct sound. The room gain and the voice support were calculated in the way proposed by Pelegrin-García. The corresponding gain introduced by each response on the direct sound, in one-third octave frequency bands between 100 Hz and 4 kHz, is shown in Fig. 6.

C. Vocalizations

Each acoustic condition was repeated three times but using different vowels every time. The three vowels /a/, /i/, and /u/ were chosen because they are known to be the so-called corner vowels with the widest spread of the formants. The bone conducted acoustic feedback paths for these vowels are different among them. In this way, the contributions from different bone conduction paths to the autophonic ratings are averaged, and the results are more representative of average speech.

D. Procedure

The experiment was carried out using two different signals as the loudness reference. The first one uses recordings from subjects’ own vocalizations as a reference, and the second one uses a 1 kHz tone as a reference. The reason for this decision was twofold. First, having a human vocalization as the reference could lead to an imitation of the vocal effort and not only to a replication of loudness. Second, using a pure tone could have made the task more difficult because of the mismatch in the perceived sound quality of the reference and the vocalization.

The measurements in the first test, using subjects’ vocalizations as reference sounds, required two steps: (a) recording of references and (b) voice matching test.

1. Recording of references

In the beginning of the test, every subject recorded the three vowels /a/, /i/, and /u/ with the following protocol [Fig. 7(a)]:

(1) A voice played back through the earphones the vowel to utter.

(2) After 1.5 s, a beep indicated the beginning of the reference vocalization.

![FIG. 5. Backward-integrated energy-time-curves for the acoustic conditions C1 to C9 presented in the test. The condition C10 (no additional impulse response) is not shown in the figure.](image)

![FIG. 6. Gain of the impulse response of each condition C1 to C9 relative to the energy of the impulse response in the anechoic chamber (condition C10), analyzed in one-third octave bands.](image)

![FIG. 7. Procedure followed in the test. Note: The duration of the events and its separation is only approximate.](image)
(3) The subjects were instructed to produce a steady vocalization after the beep signal, using a comfortable voice level. The voice was recorded.

(4) Another beep, four seconds later, indicated the end of the utterance.

(5) The recordings were analyzed to check its steadiness, and they were repeated (from step 1) until the deviation of 200-ms equivalent overall SPL in consecutive, non-overlapping periods, was in a 3 dB range for at least 2 s. The 2 s segment with the lowest deviation was chosen as the reference for the given vowel and subject.

(6) An equalizer filter was applied to the references recorded with the headworn microphone, so as to later reproduce by the earphones the levels and spectral distributions present at the ears during the original vocalizations.

2. Voice matching test

This phase is shown in Fig. 7(b).

(1) The three vowels were selected in random order. The 2-s reference containing the chosen vowel was played back.

(2) After 1.5 s, a beep indicated the beginning of the vocalization and, at the same time, the convolver was activated with one of the ten conditions C1 to C10 (in random order).

(3) The subjects had been instructed to produce a steady vocalization after the beep signal, with the same vowel and the same loudness as the reference. The voice was recorded.

(4) Another beep, three seconds later, indicated the end of the utterance and the deactivation of the convolver.

The measurements with the tone as a reference—called "tone matching test" [Fig. 7(c)]—were very similar to the voice matching test, but the reference in step 1 was substituted with an audible message of the vowel to produce followed by a 1 kHz sinusoid signal of 2 s duration and played back at a level of 75 dB SPL measured at the eardrum of a dummy head. The subjects were explicitly instructed to match the loudness of the pure tone.

At the beginning of the experiment, the subjects made a training run with five conditions and one vowel from the voice matching test to get acquainted to the procedure. The results of the training measurements were not used for the posterior analysis. In total, each subject produced 60 vocalizations (10 acoustic conditions, 3 vowels, and 2 references) that were used for the analysis.

E. Postprocessing

Each recording was analyzed for a stability criterion, looking for a one-second interval in which the deviation of 200 ms equivalent overall SPL in consecutive, non-overlapping periods, was in a 3 dB range. The one-second interval with the lowest deviation was used in the analysis. The SPL in the one-octave frequency bands between 125 Hz and 4 kHz ($L_i$), together with the overall unweighted ($L_Z$) and A-weighted SPL ($L_A$), were extracted from each recording for building the statistical model. The SPL in condition C10 (anechoic) was used as the reference factor to normalize all the other levels. The relative level $\Delta L_i$ is defined as $\Delta L_{ij} = L_{ij} - L_{i,C10}$. (3a)

$\Delta L_{Zj} = L_{Zj} - L_{Z,C10}$, (3b)

$\Delta L_{Aj} = L_{Aj} - L_{A,C10}$, (3c)

where $i$ is the frequency band and $j$ is one of the conditions C1 to C9.

The spread in SPL among conditions was studied in the frequency domain. For the spectral analysis of the signals, one-third octave band filters were used. Two descriptors were used, one for low frequencies and another one for high frequencies. These were the average rms deviation in the eight one-third octave frequency bands between 100 Hz and 500 Hz, $s_{100–500}$, and the average rms deviation in the nine one-third octave frequency bands between 630 Hz and 4 kHz, $s_{630–4k}$.

\[
s_{100–500} = \frac{1}{8} \sum_{i=1}^{8} \frac{1}{9} \sum_{j=1}^{9} (\Delta L_{ij} - \overline{\Delta L_{ij}})^2, \quad (4a)
\]

\[
s_{630–4k} = \frac{1}{17} \sum_{i=9}^{17} \frac{1}{9} \sum_{j=1}^{9} (\Delta L_{ij} - \overline{\Delta L_{ij}})^2, \quad (4b)
\]

where

$$\overline{\Delta L_{ij}} = \frac{1}{9} \sum_{j=1}^{9} \Delta L_{ij} \quad i = 1, \ldots, 17. \quad (5)$$

The subindex $i$ refers to the third-octave band center frequency ($f_{i=1} = 100$ Hz to $f_{i=17} = 4$ kHz), whereas the subindex $j$ refers to one of the acoustic conditions C1 to C9.

F. Statistical analysis

An analysis of variance (ANOVA) table, including main effects and interactions among the acoustic condition (C1 to C9), the gender (male/female), the vowel (/a/, /i/, or /u/), and the reference (tone or voice), was obtained to calculate their relative contribution to the variations of $\Delta L_Z$ and $\Delta L_A$. For the derivation of this table, an additive, fixed-effects model was assumed. $\Delta L_Z$ was the variable of interest in the study, comparable to other sidetone studies, although $\Delta L_A$ was reported too for being a closer indicator of the loudness perception.

From the inspection of the data, the mean values of $\Delta L_Z$, $\Delta L_A$, or all the $\Delta L_i$ did not change linearly with the room gain or the voice support. Instead, they followed a non-linear trend of the form

$$\Delta L = A(e^{-B \times GRG} - 1) - C \quad (6)$$

as a function of the room gain, or

$$\Delta L = A \left[ \left( 10^{\Delta LV/10} + 1 \right)^{\text{10}^{\Delta LV/10}} - 1 \right] - C \quad (7)$$

where $\Delta L$ is the loudness deviation in decibels, $A$ and $B$ are constants, $GRG$ is the gain ratio gain, and $\Delta LV$ is the loudness deviation in subjective loudness units.
as a function of the voice support. A, B, and C are the parameters of the model (identical in the two previous equations) and the relation

\[ G_{RG} = 10 \log \left( 10^{3Fv/10} + 1 \right) \]  \hspace{1cm} (8)

was used.\(^{12}\)

The fitting of the nonlinear function to the measured data, in order to obtain the A, B, and C parameters, was performed with the routine nls of the library stats of the statistical software R.\(^{21}\)

III. RESULTS

Table II shows the results of the four-way ANOVA for \( \Delta L_Z \), considering a fixed-effects, additive model, with the main effects and all possible interactions. It reveals that there is a significant effect of the acoustic condition \( [F(8,594) = 90.5, \ p < 0.0001] \), responsible for almost the 85% of the explained variance. Gender has also a significant effect \( [F(1,594) = 42.3, \ p < 0.0001] \), and is responsible for another 5% of the explained variance. The variables reference and vowel do not report significant effects. However, there is a significant three-way interaction among reference, vowel, and gender \( [F(2,594) = 11.3, \ p < 0.0001] \) explaining 2.6% of the variance. Two-way interactions including these variables are also significant: reference-vowel interaction \( [F(2,594) = 5.44, \ p = 0.005] \) and vowel-gender interaction \( [F(2,594) = 5.13, \ p = 0.006] \), responsible however, for less than 1.5% of the explained variance. There are no significant interactions between the acoustic condition and any other variable. In the additive model, the average \( \Delta L_Z \) is \(-3.3\) dB for females, whereas it is \(-2.2\) dB for males. The average \( \Delta L_Z \) for the different combinations of gender, vowel, and reference signal are shown in Table III.

Table II also shows the results of the four-way ANOVA for \( \Delta L_A \). As with \( \Delta L_Z \), the most important effect is due to the acoustic condition \( [F(8,594) = 98.4, \ p < 0.0001] \) which accounts for 87.7% of the explained variance. This increase in the explained variance is probably due to the closer relationship of the A-weighting to the loudness perception. The gender has also a significant effect \( [F(1,594) = 19.1, \ p < 0.0001] \) and accounts for 2.1% of the explained variance. In the additive model, the average \( \Delta L_A \) is \(-3.8\) dB for females and \(-2.9\) dB for males. The effect of the reference is at the limit of significance \( [F(1,594) = 4.1, \ p = 0.042] \) and it accounts for barely a 0.5% of the explained variance. However, a one-way ANOVA model with reference as the only explanatory variable does not pass a significance test. The vowel has no significant effect on \( \Delta L_A \). There is a significant three-way interaction among reference, vowel, and gender \( [F(2,594) = 10.8, \ p < 0.0001] \) accounting for 2.4% of

### Table II. Four-way analysis of variance table with main effects and interactions applied to the relative overall SPL, unweighted (\( \Delta L_Z \)) and A-weighted (\( \Delta L_A \)). Interactions between factors are indicated by an asterisk.

<table>
<thead>
<tr>
<th></th>
<th>( \Delta L_Z )</th>
<th>( \Delta L_A )</th>
<th>p-value</th>
<th>% Expl. variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-value</td>
<td>p-value</td>
<td>% Expl. variance</td>
<td>F-value</td>
<td>p-value</td>
</tr>
<tr>
<td>Reference</td>
<td>1.95</td>
<td>&lt; 0.001</td>
<td>4.14</td>
<td>0.042</td>
</tr>
<tr>
<td>Vowel</td>
<td>1.43</td>
<td>&lt; 0.001</td>
<td>0.23</td>
<td>NS</td>
</tr>
<tr>
<td>Gender</td>
<td>42.3</td>
<td>5.0</td>
<td>19.1</td>
<td>2 \times 10^{-5}</td>
</tr>
<tr>
<td>Acoustic condition</td>
<td>90.5</td>
<td>84.9</td>
<td>98.4</td>
<td>&lt; 10^{-6}</td>
</tr>
<tr>
<td>Two-way interactions</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reference*vowel</td>
<td>5.44</td>
<td>0.005</td>
<td>4.7</td>
<td>0.01</td>
</tr>
<tr>
<td>Reference*gender</td>
<td>2.99</td>
<td>0.08</td>
<td>4.00</td>
<td>0.046</td>
</tr>
<tr>
<td>Vowel*gender</td>
<td>5.02</td>
<td>0.007</td>
<td>4.81</td>
<td>0.008</td>
</tr>
<tr>
<td>Reference*acoustic condition</td>
<td>0.47</td>
<td>NS</td>
<td>0.63</td>
<td>NS</td>
</tr>
<tr>
<td>Vowel*acoustic condition</td>
<td>0.32</td>
<td>NS</td>
<td>0.39</td>
<td>NS</td>
</tr>
<tr>
<td>Gender*acoustic condition</td>
<td>0.58</td>
<td>NS</td>
<td>0.40</td>
<td>NS</td>
</tr>
<tr>
<td>Three-way interactions</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reference<em>vowel</em>gender</td>
<td>11.3</td>
<td>2 \times 10^{-5}</td>
<td>10.8</td>
<td>3 \times 10^{-5}</td>
</tr>
<tr>
<td>Reference<em>vowel</em>acoustic condition</td>
<td>0.46</td>
<td>NS</td>
<td>0.59</td>
<td>NS</td>
</tr>
<tr>
<td>Reference<em>gender</em>acoustic condition</td>
<td>0.26</td>
<td>NS</td>
<td>0.44</td>
<td>NS</td>
</tr>
<tr>
<td>Vowel<em>gender</em>acoustic condition</td>
<td>0.37</td>
<td>NS</td>
<td>0.50</td>
<td>NS</td>
</tr>
<tr>
<td>Four-way interaction</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reference<em>vowel</em>gender*acoustic condition</td>
<td>0.41</td>
<td>NS</td>
<td>0.45</td>
<td>NS</td>
</tr>
</tbody>
</table>

### Table III. Average relative overall SPL, unweighted (\( \Delta L_Z \)) and A-weighted (\( \Delta L_A \)), for the different combinations of genders, vowels, and reference signals.

<table>
<thead>
<tr>
<th></th>
<th>( \Delta L_Z ), dB</th>
<th>( \Delta L_A ), dB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Female</td>
<td>Male</td>
</tr>
<tr>
<td>Tone as reference</td>
<td></td>
<td></td>
</tr>
<tr>
<td>/a/</td>
<td>-3.34</td>
<td>-2.48</td>
</tr>
<tr>
<td>/i/</td>
<td>-3.76</td>
<td>-2.66</td>
</tr>
<tr>
<td>/u/</td>
<td>-2.51</td>
<td>-2.18</td>
</tr>
<tr>
<td>Voice as reference</td>
<td></td>
<td></td>
</tr>
<tr>
<td>/a/</td>
<td>-3.81</td>
<td>-1.56</td>
</tr>
<tr>
<td>/i/</td>
<td>-2.06</td>
<td>-2.60</td>
</tr>
<tr>
<td>/u/</td>
<td>-4.34</td>
<td>-2.08</td>
</tr>
</tbody>
</table>
the explained variance. The three two-way interactions resulting from pairs of these variables are also significant: between reference and vowel \( F(2,594) = 4.7, p = 0.01 \), accounting for 1.0% of the explained variance, between reference and gender \( F(1,594) = 4.0, p = 0.046 \), accounting for 0.5% of the explained variance, and between vowel and gender \( F(1,594) = 4.8, p < 0.008 \), accounting for 1.1% of the explained variance. The average \( DLA \) for the different combinations of gender, vowel, and reference signal are shown in Table III.

The values of \( \Delta L_Z \) are plotted as a function of \( T \) in Fig. 8. No trend relating the two variables can be observed from the measurements, because the \( \Delta L_Z \) are scattered homogeneously.

The average results of \( \Delta L_i \) in the frequency bands from 125 Hz to 4 kHz, along with the overall unweighted and A-weighted relative SPL values (\( \Delta L_Z \) and \( \Delta L_A \), respectively) are shown in Fig. 9. In the top row, the results are shown for males and females separately. The abscissa shows the room gain parameter. In the bottom row, the same results are shown, but plotted against the voice support. Each data point corresponds to the average of all subjects of one gender, vowels and reference for the same condition. Different symbols correspond to different measures. The bars around the data points indicate ±1 standard error.

It can be seen that the \( \Delta L_i \) values are arranged in a nonlinear fashion. Observing the data in the room gain plots, each level \( \Delta L \) falls close to a curve given in Eq. (6). This nonlinear model indicates that all points converge to a constant level \(-C\) for \( GRG \to 0 \) and that they tend to a limit value \(-A-C\) as \( GRG \) approaches \( 1 \). The parameter \( B \) defines the slope of the curve, together with \( A \). The best fitting curves are overlaid on Fig. 9, and the \( A, B, \) and \( C \) parameters for all \( \Delta L \), separately for males and females, are shown in Table IV.

An average model for males and females together, for \( \Delta L_Z \) and \( \Delta L_A \) is given by
as a function of the room gain, or alternatively, using the voice support,

\[ \Delta L_Z = 8.4 \times e^{-0.24G_{60}} - 8.9 \text{ [dB]}, \quad (9a) \]

\[ \Delta L_A = 6.4 \times e^{-0.25G_{60}} - 6.9 \text{ [dB]} \quad (9b) \]

Figure 10(a) shows the measured spectra in one-third octave bands for the different vowels (/a/ on the top row, /i/ on the middle row, and /u/ on the bottom row), under the different conditions (different line styles), for the female (left column) and male subjects (right column), averaged for the two reference signals and the different subjects for each gender. As shown in Fig. 9, the differences among conditions are greater at high frequencies. This is also reflected in the average rms deviation \( s \) in Table VI, which is higher in the frequency bands between 630 Hz and 4 kHz \((s_{630-4k})\) in the range from 2.57 to 3.75 dB than in the frequency bands between 100 and 500 Hz \((s_{100-500})\) in the range from 1.47 to 2.09 dB.

Figure 10(b) results from adding the gains of each condition in Fig. 6 to the spectra of the vocalizations on those conditions [plotted in Fig. 10(a)]. As can be seen, the deviations among spectra is greatly reduced, in particular, at high frequencies, where the average rms deviation \( s \) is now in the range of 1.05 to 1.52 dB, as shown in Table VI. By applying the gain of the IR, the average rms deviation in the low frequency range, \( s_{100-500} \), is lower for the vowels /i/ and /u/, but not for /a/, and it ranges from 1.28 to 1.68 dB in all cases. These numbers reflect a uniform spread of the spectra in a broader frequency range for the corrected recordings, which are a closer approximation to the levels perceived by the subjects.

### IV. DISCUSSION

From the observation of the measured relative voice levels in Figs. 8 and 9 and 10(a), it is possible to state that

<table>
<thead>
<tr>
<th>Gender</th>
<th>Parameter</th>
<th>( \Delta L_{125} )</th>
<th>( \Delta L_{250} )</th>
<th>( \Delta L_{500} )</th>
<th>( \Delta L_{1k} )</th>
<th>( \Delta L_{2k} )</th>
<th>( \Delta L_{4k} )</th>
<th>( \Delta L_Z )</th>
<th>( \Delta L_A )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Females</td>
<td>A</td>
<td>2.87</td>
<td>4.83</td>
<td>8.73</td>
<td>8.82</td>
<td>11.11</td>
<td>11.12</td>
<td>6.71</td>
<td>8.18</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>0.35</td>
<td>0.22</td>
<td>0.23</td>
<td>0.29</td>
<td>0.27</td>
<td>0.36</td>
<td>0.26</td>
<td>0.26</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>0.65</td>
<td>0.87</td>
<td>0.92</td>
<td>0.8</td>
<td>1.22</td>
<td>1.99</td>
<td>1.05</td>
<td>1.11</td>
</tr>
<tr>
<td>Males</td>
<td>A</td>
<td>3.11</td>
<td>6.14</td>
<td>8.89</td>
<td>9.70</td>
<td>11.95</td>
<td>10.49</td>
<td>6.33</td>
<td>8.52</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>0.23</td>
<td>0.30</td>
<td>0.20</td>
<td>0.26</td>
<td>0.21</td>
<td>0.24</td>
<td>0.25</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>0.17</td>
<td>0.07</td>
<td>0.27</td>
<td>0.4</td>
<td>0.58</td>
<td>1.07</td>
<td>0.18</td>
<td>0.22</td>
</tr>
</tbody>
</table>

\[
\Delta L_Z = 8.4 \times e^{-0.24G_{60}} - 8.9 \text{ [dB]}, \quad (9a) \\
\Delta L_A = 6.4 \times e^{-0.25G_{60}} - 6.9 \text{ [dB]} \quad (9b)
\]
different acoustic environments alter the autophonic level for a talker. However, the reverberation time is not a good descriptor of the changes in voice level, as seen in Fig. 8, since it is not directly related to the energy of the indirect auditory feedback. Figure 9 describes the changes in voice level that make the voice of the talkers sound equally loud at their ears when the indirect acoustic feedback is changed. The curves for $\Delta L_Z$ show a constant autophonic level under different room gain conditions (top row), or voice support conditions (bottom row). The A-weighted and the one-octave band values follow the same general trend of the nonlinear model in Eq. (6), but with different model parameters. In normal rooms for speech without amplification ($G_{RG} < 1.0$ dB), the variations in voice level to keep a constant autophonic level are within 2 dB, according to model Eq. (9a). In the frequency band of 4 kHz, this range increases to more than 3 dB using the parameters of Table IV.

For the three lowest values of voice support ($-18.0$ dB $\leq S_{TV} \leq -13.5$ dB), excluding the anechoic chamber, the range of $\Delta L_Z$ is about 0.3 dB, calculated from the model in Eq. (10a). There are consistent voice level variations in a range of less than 0.5 dB, which is considered to be the just noticeable level difference for broadband noise signals. These observations agree with recent findings, which suggest that an auditory motor system controls voice intensity in a non-conscious way and is able to react to level variations below the conscious detectability threshold.

The model in Eq. (9a) shows a varying slope in the dependence of voice level with room gain. It is most negative (or maximum in absolute value) for $G_{RG}$ $\rightarrow$ 0 with a value of $-2.0$ dB/dB. In the range observed, the least negative slope is obtained for the highest room gain value ($G_{RG}$ = 8.6 dB). In this case, the slope is $-0.26$ dB/dB. The same equation indicates a saturation effect (zero slope) as $G_{RG}$ $\rightarrow$ $\infty$. This could be an indication that the voice levels approach the phonation threshold with the given experimental setup. However, no generalization of the model is intended for values of $G_{RG}$ higher than the studied range.

In a review of different studies of sidetone, Lane et al. showed that the sidetone compensation function is linear with slopes varying between $-0.4$ and $-0.6$ dB/dB. With the model in Eq. (9a), these slopes are obtained in the range of 5 dB $\leq G_{RG} \leq 6.7$ dB. Using Eq. (2), a $G_{RG}$ of 5 dB is equivalent to a ratio of indirect to direct airborne sound of approximately 2. Several studies have stated that the direct airborne sound and bone conducted sound of one’s own voice are of a comparable magnitude. A $G_{RG}$ of 5 dB indicates that the reflected sound is of the same importance as the combination of the direct airborne sound and the bone conducted sound of one’s own voice. For values of $G_{RG}$ higher than 5 dB, the indirect auditory feedback component is dominating, and the slopes are comparable to those found in traditional sidetone studies.

Lane and Tranell pointed out that the Lombard reflex and the sidetone compensation are two sides of the same coin. In later experiments, Pick et al. showed that the Lombard reflex is very difficult to inhibit. Consequently, it is natural that the sidetone compensation is also difficult to inhibit. In the absence of background noise, large values of room gain would make a talker speak softer, as it could happen when using an electroacoustic reinforcement system. From a different perspective, it could be possible to consider that a good room for speech has a certain value of room gain. A room of drier acoustics and with a lower room gain would make the talker speak louder. However, in rooms without electroacoustic amplification, the range of room gain is bounded between 0 and approximately 1 dB, which would induce changes in voice level of less than 2 dB. At the first glance, this value seems not to be very significant compared to the dynamic range of the human voice (roughly 30 dB, depending on the person and the fundamental frequency).

The equal autophonic level curve for $\Delta L_Z$, described in Eq. (9a), is compared to the results of other two studies (Refs. 12 and 8) in Fig. 11 (Note: the two studies show variations in voice power level, whereas the equal autophonic level curves are indicated as variations in SPL, so the comparison is approximate). The dataset of Ref. 12 shows the variations in voice level of teachers lecturing in classrooms of different sizes and room gains. The slope of the line that relates voice levels with room gain is $-13.5$ dB/dB. However, the changes in voice level are not purely due to the perception of room acoustics, but to other aspects of the

![FIG. 11. Comparison of the voice power levels used by teachers in different classrooms]
communication scenario, such as the variation in distance between talker and listeners that occurs naturally in different rooms of different size. At the same time, the smallest room is the one with the largest room gain. Therefore, the dataset of Ref. 12 is representative of typical voice level variations in rooms without background noise. The dataset of Ref. 8 presents data of a talker addressing a listener at a distance of 6 m in front of him in four different rooms with different room gain. The average voice level varies with the room gain at a rate of $-3.6 \text{ dB/dB}$. In the same range of $G_{RG}$, the equal autophonic level curve approximates a straight line with a slope of $-1.8 \text{ dB/dB}$. The talkers in these two experiments did not follow a communication strategy based on keeping the autophonic level constant. In case they did, the voice measurements would have lain on top of the equal autophonic level curve. Talkers apparently “amplify” the effect of the Lombard reflex. This suggests that they make use of attributes present in the room impulse response other than loudness for the adjustment of their voice, probably in combination with other sensory inputs. One explanation for the difference in slope is that the talkers in Ref. 8 adjust their voice level according to some tacit knowledge of sound attenuation with distance, as suggested by Zahorik and Kelly, although do not completely compensate for that. In the experiment of Ref. 8, the sound attenuation at 6 m from the talker differed by more than 15 dB in the two most extreme cases (with $G_{RG} \approx 0 \text{ dB}$ and $G_{RG} \approx 0.8 \text{ dB}$), whereas the voice level variation was only about 3 dB at the source.

The amount of voice level variation to achieve a constant autophonic level is different for the two genders and for different frequency bands. As shown in Fig. 11, female vocalizations have an important energy peak at 200 Hz. At the same time, the gain applied by the acoustic conditions to the voice, shown in Fig. 6, had a dip at 200 Hz in the present experiment. Therefore, female voices were slightly less amplified than male voices, and females had to use more intensity than males to match the loudness of the reference sound. Therefore, the reported effects of acoustic conditions on female voice level variation may be overestimated (or underestimated for male talkers). The amount of voice level variation is less important at low frequencies and more important at higher frequencies. This can be observed in both Figs. 9 and 10(a). When applying the frequency-dependent gain introduced by the synthetic IR in Fig. 6 to the voice recordings, they seem to fall on similar curves, as shown in Fig. 10(b) and in the reduced average rms deviations in Table V. This means that the subjects kept the resulting sound from their vocalizations constant at their ears, in overall level and in spectral balance of different frequency bands. As a consequence, the parameters $A$, $B$, and $C$ of Table IV can be used in connection with the models in Eqs. (6) and (7) to describe the amount of compensation expected for the different frequency bands. It may be possible that the compensation at high frequencies is a side-effect of the change in vocalization level, because the spectral slope decreases with increasing vocal effort. Another possible explanation is that subjects try to keep the sound quality (loudness and spectral balance) of the vocalizations constant. This hypothesis is reasonable when using a vocalization as a reference, but not when using a tone.

The three-way interaction among reference, vowel, and gender, and the two-way interactions between pairs of the same variables, shown in Table II, can be understood as a result of combining different speech spectra with the frequency-dependent gain of the acoustic condition. The interaction between gender and reference can be due to the different amplification applied to male voices during the playback of the reference (vowel). The equalized response of the earphones has a slight boost at around 100 Hz which affects differently male and female voices. The interaction between gender and vowel might arise from the different frequency characteristic of the vowels for males and females, which have similar formant structure but differ in the region of the fundamental frequency. The interaction between vowel and reference can be due to different amplifications applied to the vowel when reproducing the reference sound. The three-way interaction might result from the vocalizations of particular combinations of gender and vowel that receive more or less amplification than other combinations when a vocalization is used as the reference sound, due to the non-flat equalization of the earphones.

The models in Eqs. (9a) and (10a) can be used to predict the variations in vocal intensity that happen with the use of electroacoustic amplification. As an example, Sapienza et al. found that teachers talked on average 2.4 dB softer in classrooms when using a sound reinforcement system. The gain of the system was tuned so that it increased the SPL at a distant listener position by 10 dB. At these positions, the reflected energy dominates over the direct sound energy. Making this consideration, and considering that the amplification system produces a uniform SPL in the room, the amount of nondirect energy $E_i$ increases by 10 dB when the system is turned on, also at the talker position. By Eq. (1), $ST_i$ would increase about 10 dB when the system is turned on. A representative value of $ST_i$ in nonamplified classrooms is $-13 \text{ dB}$. By using Eq. (10a), talkers would speak 2.5 dB softer when the system is on ($ST_i = -3 \text{ dB}$), compared with what they would do when the system is off ($ST_i = -13 \text{ dB}$). The good agreement of the measured and predicted variations (2.4 dB and 2.5 dB) are probably due to the fact that the only variable that was changed in the study of Sapienza et al. was the sidetone, and not any other variables like the room or the distance to the listeners, and therefore the subjects reacted sympathetically according to the Lombard reflex.

The level of the voice reference recordings was not monitored, and the test subjects received the instruction to produce a vocalization at a “comfortable” level. Since the equal loudness level contours as a function of the frequency in ISO-226:2003 (Ref. 28) are not parallel, it may be possible that the amount of compensation was different at different voice levels. This could have been studied by repeating the test with reference tones at different levels, but this was done only at one level. Since the comfortable and most used voice level changes from subject to subject, the measured equal autophonic level curves are an average indicator of this “most comfortable level.” Because the results of the
tests using the two references (voice and tone) are similar, as shown by the low significance of the variable “reference” in the ANOVA of Table II, significant differences are not to be expected among different reference levels.

V. CONCLUSIONS

An experiment was conducted to obtain the relative voice levels that kept the autophonic level constant under different room acoustics conditions described by the parameters room gain and voice support. Analyzing the voice levels in one-octave bands and with different frequency weightings, a set of equal autophonic level curves was generated. These curves allow us to determine the expected voice level differences in different rooms which are purely related to the Lombard-effect or sidetone compensation. The main conclusions of the study are as follows.

(1) Voice level variations under different room acoustics conditions are primarily related to the room gain or the voice support, rather than to the reverberation time.

(2) Typical voice level variations in rooms for speech $(G_{RG} < 1.0 \, \text{dB})$ to keep a constant autophonic level are not higher than 2 dB.

(3) By comparison with other studies, talkers use cues other than loudness to adjust their voice level in rooms, resulting in larger voice variations than barely keeping the autophonic level constant.

ACKNOWLEDGMENTS

This research has been partially funded by the Swedish organization AFA Försäkring as a part of the project “Speaker’s comfort and voice health in classrooms.” The authors would like to express their gratitude to all the subjects who took part in the tests, and to Sören Laugesen, at Oticon A/S for his unconditional help providing the authors with key elements in the setup.


Measurement and prediction of voice support and room gain in school classrooms

David Pelegrín-García and Jonas Brunskog
Acoustic Technology, Department of Electrical Engineering, Technical University of Denmark, Kongens Lyngby DK-2800, Denmark

Viveka Lyberg-Åhlander and Anders Löfqvist
Dept. Logopedics, Phoniatrics & Audiology, Lund University, Lund, Sweden

(Dated: August 15, 2011)

Objective acoustic parameters have been measured in 30 school classrooms. These parameters include usual descriptors of the acoustic quality from the listeners’ standpoint, such as reverberation time, speech transmission index, and background noise levels, and two descriptors of the acoustic properties for a speaker: voice support and room gain. The paper describes the measurement method for these two parameters and presents a prediction model for voice support and room gain derived from the diffuse field theory. The voice support for medium-sized classrooms with volumes between 100 and 250 m$^3$ and good acoustical quality lies in the range between -14 and -9 dB, whereas the room gain is in the range between 0.2 and 0.5 dB. The prediction model for voice support describes the measurements in the classrooms with a coefficient of determination of 0.84 and a standard deviation of 1.2 dB.

PACS numbers: 43.55.Gx, 43.55.Fw

I. INTRODUCTION

Learning spaces or classrooms are environments where people spend a large amount of their lifetime, mainly dedicated to acoustic communication tasks. Students spend the early part of their lives listening to the teacher in classrooms in order to learn, and also need to communicate efficiently with fellow students and the teacher. The success in communication is fundamental to develop the full potential of every student. At the same time, school classrooms are the working place of teachers, who represent an important percentage of the working population. Acoustical conditions have to be evaluated both for teachers and students.

Most of past research in classroom acoustics has been devoted to the acoustic design for students. The negative effects of noise on children perception and performance have been observed, the effect of reverberation on speech intelligibility has been quantified, and the combination of noise and reverberation has been the object of a number of studies. Different quantities are used to predict speech intelligibility: signal-to-noise ratios, useful-to-detrimental ratios, and speech transmission index (STI).

Acoustic conditions are also important for teachers. Teachers suffer from voice disorders in a higher proportion than in the rest of the population (around 13% in Sweden and a similar proportion in the US), which is most likely due to the high vocal requirements that the teaching occupation demands. Noise and bad classroom acoustics are often reported risk factors for voice disorders. Talking in the presence of high noise levels results in the use of higher voice power levels than required to talk in soft noise conditions. This is known as the Lombard effect, and it is estimated that for each decibel of noise, a speaker raises his voice power level between 0.5 and 0.7 dB. In the presence of low background noise, speakers still modify their voice power under different room acoustic conditions, even when the distance between speaker and listener is kept constant.

The quantity room gain has a negative correlation with the voice power levels used by speakers in different rooms. The room gain is defined as the gain applied by the room to the voice of the speaker at his own ears, relative to free-field. However, this magnitude has a low dynamic range, and the use of voice support seems more appropriate in room acoustics. The voice support is conceptually equivalent to Gade’s objective support used in the assessment of the acoustic conditions for musicians in concert halls.

A number of surveys have analyzed the acoustic conditions of school classrooms based on measurements of reverberation time and background noise, many times reporting measures of speech intelligibility. Despite the importance of assessing the acoustic conditions for a speaker, there are no studies that report room gain, voice support or other speaker-related parameters in school classrooms. Some studies advise about possible detrimental effects of poor acoustic conditions on vocal effort. The present paper aims at providing some information in this respect, giving reference values for voice support and room gain in typical school classrooms. In addition, the two parameters are explained in more detail than previously reported, and a prediction model based on the diffuse-field theory is presented.
II. THEORY

A. Definition and calculation of voice support and room gain

Brunskog et al.\textsuperscript{16} introduced the parameters room gain and voice support, and Pelegrin-Garcia\textsuperscript{18} suggested an alternative method for the calculation of these two parameters from a single impulse response. The procedure followed in the present paper is based on the latter approach, although it is refined regarding the frequency weighting. Given the impulse response (IR) measured with a dummy head between the mouth and the ears, $h_{ME}(t)$, the room gain $G_{RG}$ is defined as the difference between the total energy level of the IR $L_t$ and the energy level of the direct sound $L_d$. The voice support $ST_V$ is defined as the difference between the energy level of the reflections coming back from the boundaries $L_r$ and the energy level of the direct sound,

$$ G_{RG} = L_t - L_d, \quad (1) $$

$$ ST_V = L_r - L_d. \quad (2) $$

Assuming that the total energy is the sum of the direct and reflected energies,

$$ G_{RG} = 10 \log \left( 10^{\frac{ST_V}{20}} + 1 \right). \quad (3) $$

The practical calculation of the voice support from the IR is illustrated in the diagram of Fig. 1.

The IR measured between mouth and ears is split into two branches: the top one is multiplied by a window $w_d(t)$ to extract the direct sound $h_d(t)$, and the low branch is multiplied by a window $w_r(t)$ to extract the reflected sound $h_r(t)$. The two window functions are defined as

$$ w_d(t) = \begin{cases} 1 & \text{if } t < 4.5 \text{ ms} \\ 0.5 + 0.5 \cos(2\pi(t-t_0)/T_W) & \text{if } 4.5 \text{ ms} < t < 5.5 \text{ ms} \\ 0 & \text{if } t > 5.5 \text{ ms} \end{cases} $$

$$ w_r(t) = 1 - w_d(t) \quad (4) $$

with $t_0 = 4.5$ ms and $T_W = 2$ ms. To separate the direct and the reflected components, the mouth/source and the ears/receivers must be located at least 1 m away from reflecting surfaces or scattering objects other than the dummy head and the mounting elements. The time window for the reflected sound is intended to include all the decaying energy of the IR, because all of it contributes to increase the loudness of one’s own voice.

The next stage in the diagram of Fig. 1 is the spectral analysis. The direct sound IR $h_r(t)$ is decomposed into narrow band components $h_{r,i}(t)$ by using a filterbank composed of six one-octave band filters with the standardised center frequencies between 125 Hz ($i = 1$) and 4 kHz ($i = 6$). The energies $E_{d,i}$ and energy levels $L_{d,i}$ are calculated for each band. The same spectral analysis is applied to the reflected sound. The energy levels for the direct sound $L_{d,i}$ are subtracted from the reflected sound $L_r$,

$$ \Delta L_{HRTF} \quad (5) $$

B. Prediction model

Using the definition of $ST_V$ in Eq. (2), a prediction model must account for the relation between the direct and the reflected sound at the ears, when the mouth acts as a source. To build this model, it is assumed that the measurement equipment is a HATS (head and torso

<table>
<thead>
<tr>
<th>Band $i$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency [Hz]</td>
<td>125</td>
<td>250</td>
<td>500</td>
<td>1000</td>
<td>2000</td>
<td>4000</td>
</tr>
<tr>
<td>Typical speech SPL on-axis at 1 m $L_{d,1m}$ [dB]</td>
<td>44.9</td>
<td>57.3</td>
<td>61.8</td>
<td>58.2</td>
<td>53.7</td>
<td>48.9</td>
</tr>
<tr>
<td>Difference with SPL at eardrum $L_d - L_{d,1m}$ [dB]</td>
<td>13.1</td>
<td>11.8</td>
<td>11.7</td>
<td>13.5</td>
<td>15.3</td>
<td>14.1</td>
</tr>
<tr>
<td>Typical speech levels at the eardrum $L_{ref,ears}$ [dB]</td>
<td>58.0</td>
<td>69.1</td>
<td>73.5</td>
<td>71.7</td>
<td>69.0</td>
<td>63.0</td>
</tr>
<tr>
<td>Relation between $L_W$ and on-axis SPL at 1 m $L_d - L_{d,1m} - L_W$ [dB]</td>
<td>-9.5</td>
<td>-8.1</td>
<td>-9.2</td>
<td>-9.5</td>
<td>-7.0</td>
<td>-6.0</td>
</tr>
<tr>
<td>Constant $K$ for model Eq. (19) $K$ [dB]</td>
<td>3.6</td>
<td>3.7</td>
<td>2.5</td>
<td>4.0</td>
<td>8.3</td>
<td>8.1</td>
</tr>
<tr>
<td>Directivity of human speech on downward direction $Q^*$</td>
<td>0.95</td>
<td>0.78</td>
<td>0.79</td>
<td>0.60</td>
<td>0.21</td>
<td>0.25</td>
</tr>
<tr>
<td>Diffuse field HRTF $\Delta L_{HRTF}$ [dB]</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>11</td>
<td>13</td>
</tr>
</tbody>
</table>

$\Delta L_{HRTF}$, obtaining the values of voice support in one-octave bands $ST_{V,i}$. These values are weighted with a typical speech spectrum at the ears, $L_{ref,ears}$, shown in Table I. These levels have been determined from typical anechoic speech levels on-axis at 1 m\textsuperscript{24} and the relation between the SPL on-axis at 1 m $L_{d,1m}$ and the SPL at the eardrum measured in anechoic chamber $L_d$. The overall weighted reference direct sound level $\tilde{L}_d$ and reflected sound level $\tilde{L}_r$ are

$$ \tilde{L}_d = 10 \log \left( \sum_{i=1}^{6} \frac{L_{ref,ears,i}}{10} \right) \quad (6) $$

$$ \tilde{L}_r = 10 \log \left( \sum_{i=1}^{6} \frac{L_{ref,ears,i + ST_V,i}}{10} \right) \quad (7) $$

from which the overall speech-weighted voice support $\tilde{ST}_V$ (or simply, voice support) is finally calculated as

$$ \tilde{ST}_V = \tilde{L}_r - \tilde{L}_d = 10 \log \left( \sum_{i=1}^{6} \frac{L_{ref,ears,i + ST_V,i}}{10} \right). \quad (8) $$
The reference pressure, $\rho$, where $Q$ alone is L for the reflected sound level, $S$ and $\overline{\alpha}$ is the mean absorption coefficient, which is derived from the volume $V$ and the reverberation time $T$ measurements through Sabine’s formula $\overline{\alpha} = 4 \ln(10^6)V/(cST)$. Therefore, the reflected sound level, $L_r$, due to the reflections alone is

$$L_r = L_W + 10 \log \left( \frac{4}{R} \overline{S}_{\text{ref}} \right).$$

For predicting $ST_V$, it would be enough to substitute Eqs. (9) and (12) into (2). However, there are three factors that make the calculation of $ST_V$ slightly different:

1. **Modeling of the direct sound**

To account for the special propagation between mouth and ears due to the diffraction of sound around the head and the filtering of the external ear, instead of using Eq. (9), $L_d$ is related to $L_W$ through

$$L_d = L_W + K,$$

where

$$K = (L_d - L_{d,1m}) + (L_{d,1m} - L_W).$$

By introducing this pair of terms, the value of $K$ is decomposed into two quantities. The first quantity, $(L_d - L_{d,1m})$, is determined by the simultaneous SPL measurement at the ears and one meter in front of the mouth of a HATS B&K 4128 reproducing pink noise in an anechoic chamber. The second quantity, $(L_{d,1m} - L_W)$, is determined from the speech directivity patterns measured by Chu and Warnock. The values of the two quantities and $K$ in the different frequency bands are shown in Table I.

2. **Ground reflection**

The level of a sound reflection from the ground $L_{\text{refl}}$ would be

$$L_{\text{refl}} = L_W + 10 \log \left( \frac{Q^*}{4\pi(2d)^2} \overline{S}_{\text{ref}} \right)$$

at the position of the source, which is at a height $d$ from the ground. $Q^*$ is the directivity factor of speech in the downward direction (derived from Chu and Warnock) and its frequency-dependent values are shown in Table I. The height $d$ can be regarded as 1.5 m, which corresponds to the mouth position of a standing female speaker.

Under these conditions, the expected reflected SPL at the position of the dummy head (without it disturbing the sound field) would be

$$L_r = L_W + 10 \log \left( \frac{4}{R} - \frac{Q^*}{4\pi(2d)^2} \overline{S}_{\text{ref}} \right).$$

3. **HRTF correction**

Actually, the artificial head used for measurements disturbs the sound field. Therefore, it is necessary to apply a simulator) B&K (Brüel & Kjær Sound & Vibration Measurement A/S; Nærum, Denmark) type 4128.

In general, the sound pressure level (SPL) caused by a point source with sound power level $L_W$ at a distance $r$ in free-field (direct sound level, $L_d$) is

$$L_d = L_W + 10 \log \left( \frac{Q}{4\pi r^2} \overline{S}_{\text{ref}} \right).$$

where $Q$ is the directivity of the source and $\overline{S}_{\text{ref}}$ is the reference area

$$\overline{S}_{\text{ref}} = \frac{W_{\text{ref}} \rho_0 c}{\rho_{\text{ref}}^2}.$$  

$W_{\text{ref}} = 1$ pW is the reference power, $\rho_{\text{ref}} = 20$ $\mu$Pa is the reference pressure, $\rho_0 \approx 1.204$ kg · m$^{-3}$, $c \approx 343$ m/s, and $\overline{S}_{\text{ref}} \approx 1$ m$^2$. If the source is radiating into half-space (e.g., due to the presence of a reflective plane, like a typical floor) $Q$ becomes 2. When this source is placed in a room, the SPL increases due to sound reflections at the boundaries. Assuming a diffuse sound field, the SPL in a room $L_p$ becomes

$$L_p = L_W + 10 \log \left( \frac{Q}{4\pi r^2} + \frac{4}{R} \overline{S}_{\text{ref}} \right),$$

where $R = S\overline{\alpha}/(1 - \overline{\alpha})$ is sometimes called “room constant”, $S$ is the total surface area of the room and $\overline{\alpha}$ is the mean absorption coefficient, which is derived from the volume $V$ and the reverberation time $T$ measurements through Sabine’s formula $\overline{\alpha} = 4 \ln(10^6)V/(cST)$. Therefore, the reflected sound level, $L_r$, due to the reflections alone is

$$L_r = L_W + 10 \log \left( \frac{4}{R} \overline{S}_{\text{ref}} \right).$$

FIG. 1. Block diagram for the calculation of voice support.
correction term that relates the SPL at the measurement position when the equipment is present to the SPL at the same position in the absence of the equipment. In the case of the HATS, this correction corresponds to the definition of the head related transfer function (HRTF) and is notated as $\Delta L_{HRTF}$. This magnitude is usually direction dependent. As the reflected sound can arrive from many different directions, a direction averaged quantity—the diffuse field $\Delta L_{HRTF}$ given by the manufacturer\(^{26}\) is used (see Table 1).

Therefore, the reflected sound measured with the HATS is

$$L_r = L_W + 10 \log \left[ \left( \frac{4}{R} + \frac{Q^*}{4\pi(2d)^2} \right) S_{ref} \right] + \Delta L_{HRTF}. \quad (17)$$

Finally, combining Eqs. (13) and (17) into (2), the frequency-dependent model for voice support is

$$ST_V = 10 \log \left[ \left( \frac{4}{R} + \frac{Q^*}{4\pi(2d)^2} \right) S_{ref} \right] + \Delta L_{HRTF} - K, \quad (18)$$

or in terms of directly measurable variables

$$ST_V = 10 \log \left[ \left( \frac{cT}{\ln(10^6)V} - \frac{4}{S} + \frac{Q^*}{4\pi(2d)^2} \right) S_{ref} \right]$$

$$+ \Delta L_{HRTF} - K. \quad (19)$$

The results from the individual bands should be weighted to obtain a single value by means of Eq. (8). Figure 2 shows an example set of curves for calculating $ST_V$ from $V$ and $T$, assuming that the room has proportions 2.8:1.6:1 and the reverberation time has a flat frequency characteristic.

III. MATERIAL AND METHOD

Acoustic measurements of the objective parameters background noise level, $T$, STI, $ST_V$, and $G_{RG}$ have been performed in 30 unoccupied but totally furnished school classrooms. The physical dimensions of the rooms are shown in Table II. According to the volume, the rooms were classified into three groups: small ($V < 100\text{m}^3$), medium ($100 < V < 500\text{m}^3$), and large ($V > 3500\text{m}^3$) classrooms. The rooms in the last group were sports halls where gymnastic lessons took place.

A. Background noise level measurements

The A-weighted, 10-second equivalent background noise levels ($L_{N,Aeq}$) were measured in the empty classrooms using the 01dB (01dB-Metravib; Limonest Cedex, France) Symphonie system with two microphones B&K type 4192 at a height of 1.2 m. For each classroom, the measurements across four points at representative student seats were averaged.

B. Measurements with an omnidirectional sound source

The reverberation time and STI were derived from the measurements of the room IR $h_{IR}(t)$ using an omnidirectional sound source B&K type 4295 “Omnisource”. The source was placed at two different teaching positions and with the radiating opening at a height of 1.6 m pointing upwards. Two 1/2” pressure-field microphones B&K type 4192 were used as receivers and were placed close to student seats at a height of 1.2 m. The 01dB Symphonie system, incorporating the MLS software module, was used to produce the measurement signal and send it to the loudspeaker via a power amplifier, acquire the signal from the microphones, calculate the IR, and derive the parameters $T$ and STI. The reverberation time was obtained by evaluating the backwards integrated curve\(^{27}\) of the room IR in the decay interval from -5 to -25 dB. A single value descriptor corresponding to the average of the frequency bands between 500 Hz and 2 kHz $T_{500-2k}$ is given. The average (SD) values of the signal-to-noise ratio of the IR measurements in the different classroom groups were 52 dB (4.1 dB) in small classrooms, 46 dB (5.4 dB) in mid-size classrooms, and 34 dB (5.4 dB) in large classrooms.

C. Measurements with a dummy head

The voice support was determined from the measurement of an IR corresponding to the airborne sound transmission path between the mouth and the ears in the empty classrooms. For this purpose, a HATS B&K type 4128 was used. The HATS included a loudspeaker at its mouth, and microphones at its ears. The HATS was placed at a representative teaching position, with the mouth at a height of 1.5 m, and more than 1 m away from reflecting surfaces. The 01dB Symphonie system was used to produce the excitation signal and determine the mouth-to-ears impulse response from the measured signal at the microphones. For each classroom, the $ST_V$ values of the two ears at two different positions were averaged. The room gain was calculated by applying Eq. (3) on the $ST_V$ values.
D. Prediction model for voice support

The prediction model for $ST_V$ in Eq. (19) was evaluated in octave bands by using the frequency-dependent measured values of $T$, along with the volume and total surface area of the classrooms. In addition, a broadband value (speech-weighted $ST_V$) was calculated from the frequency-band values using Eq. (8).

The prediction model was assessed by comparing the measured and the predicted $ST_V$ values. In each frequency band (or overall speech-weighted), a regression line of the type $ST_{V,\text{pred}} = a \cdot ST_{V,\text{meas}} + b$ was calculated, where $ST_{V,\text{pred}}$ is the regressor for the predicted values of voice support (notated as $ST_{V,\text{pred}}$), $ST_{V,\text{meas}}$ are the measured values, and $a$ and $b$ are the coefficients of the regression line. Ideally, a perfect model would result if the predicted and the measured values were equal ($ST_{V,\text{pred}} = ST_{V,\text{meas}}$). An unbiased model would result if $a = 1$ and $b = 0$, i.e., $ST_{V,\text{pred}} = ST_{V,\text{meas}}$.

The goodness of fit of the prediction model was evaluated with three parameters: a) the coefficient of determination $R^2$ of the linear regression model for the measured versus predicted values, b) the residual deviation $\sigma_r$ of the predicted values from this regression line, and c) the deviation $\sigma_T$ of the predicted values from an unbiased prediction, which is a measure of the bias in the prediction.

\[ \sigma_r^2 = \frac{1}{N-2} \sum_{i=1}^{N} (ST_{V,\text{pred}} - ST_{V,\text{meas}})^2 \]  
\[ \sigma_T^2 = \frac{1}{N} \sum_{i=1}^{N} (ST_{V,\text{pred}} - ST_{V,\text{meas}})^2 \]

IV. RESULTS

A. Correlation between parameters

The correlation coefficients between the measured values $V$, $\log(V)$, $L_{N,A_{eq}}$, $T_{500-2k}$, $STI$, $G_{RG}$, and $ST_V$ are shown in Table III. $L_{N,A_{eq}}$ has very low correlation with all the other parameters, because it is not determined from physical properties of the room, but depends on different noise sources from installations inside the room, and from other external noise sources (traffic noise, students in neighboring classrooms, corridors, or playground). The reverberation time is correlated to the volume and negatively correlated to the STI. The prediction model for voice support is strongly correlated to the logarithm of the volume, as expected from the prediction model in Eq. (19). The presence of some high correlation coefficients is largely caused by the large measured range of volumes and most of the other parameters of the classrooms. The correlation between $V$ and $T_{500-2k}$ is 0.97, due to the large difference between the volumes in the sports halls and the rest of the classrooms (see Table II) and the similarity of reverberation times and absorption characteristics of the materials within each group.

B. Background noise levels

The mean and maximum background noise levels (A-weighted and in one octave frequency bands) are shown in Table IV. Although it is not explicitly shown, 73.3% of the classrooms had $L_{N,A_{eq}}$ lower than 35 dB, another 13.3% between 35 and 40 dB, and the remaining 13.3% of the measurements were between 40 and 45 dB. In most of the cases, the noise sources corresponded to the ventilation systems, although in a few cases, the background noise was affected by external sources, such as neighboring activities, playground, and traffic. The background noise levels were similar for all room sizes, although the overall level in the large rooms was slightly higher than in smaller rooms. In all cases, low frequency noise was markedly dominating. This is an indication that the sources, in most of the cases, were in fact the machinery of the ventilation systems, or external noise that leaks into the room due to the usually low insulation performance of walls, doors, and windows at low frequencies.

C. Reverberation time

The mean reverberation times (in octave bands and 500 Hz-2 kHz average) and their standard deviation are shown in Table V. 81.5% (22 out of 27) of the small and medium classrooms had reverberation times lower than 0.5 s, and the remaining 18.5% were between 0.5 and 0.6 s. In the sports halls, $T$ was between 1.4 s and 1.8 s.

D. Speech transmission index

The average (standard deviation) measured STI with a negligible effect of background noise was 0.80 (0.02) in small classrooms, 0.75 (0.03) in medium classrooms and 0.63 (0.02) in large classrooms. The spread of STI among rooms, indicated by the standard deviation, was similar.

### Table II. Average (standard deviation) dimensions and volumes of the measured school classrooms

<table>
<thead>
<tr>
<th>Group size</th>
<th>Number of rooms</th>
<th>W (m)</th>
<th>L (m)</th>
<th>H (m)</th>
<th>Volume (m$^3$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small</td>
<td>3</td>
<td>4.5 (0.8)</td>
<td>3.3 (0.8)</td>
<td>2.7 (0.2)</td>
<td>40.6 (18.3)</td>
</tr>
<tr>
<td>Medium</td>
<td>24</td>
<td>8.9 (1.7)</td>
<td>7.0 (1.0)</td>
<td>2.8 (0.2)</td>
<td>180.2 (61.6)</td>
</tr>
<tr>
<td>Large</td>
<td>3</td>
<td>23.6 (2.4)</td>
<td>20.8 (0.2)</td>
<td>7.4 (0.7)</td>
<td>3614.3 (77.0)</td>
</tr>
</tbody>
</table>
TABLE III. Correlation coefficient indicating the strength of the linear dependence between pairs of variables. Only coefficients with absolute value of at least 0.5 are shown. Correlation coefficients larger than 0.80 in absolute value are marked as bold.

<table>
<thead>
<tr>
<th></th>
<th>V</th>
<th>log V</th>
<th>( L_{N,A_{eq}} )</th>
<th>( T_{500–2k} )</th>
<th>STI</th>
<th>( G_{RG} )</th>
<th>( ST_V )</th>
</tr>
</thead>
<tbody>
<tr>
<td>V</td>
<td>1.00</td>
<td>0.91</td>
<td>—</td>
<td>0.97</td>
<td>-0.81</td>
<td>-0.50</td>
<td>-0.75</td>
</tr>
<tr>
<td>log V</td>
<td>1.00</td>
<td>—</td>
<td>—</td>
<td>0.91</td>
<td>-0.57</td>
<td>-0.75</td>
<td>-0.87</td>
</tr>
<tr>
<td>( L_{N,A_{eq}} )</td>
<td>1.00</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>( T_{500–2k} )</td>
<td>1.00</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>STI</td>
<td>1.00</td>
<td>0.91</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>( G_{RG} )</td>
<td>1.00</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>( ST_V )</td>
<td>1.00</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

TABLE IV. Frequency band values and overall A-weighted background noise levels \((L_N)\) measured in the classrooms.

<table>
<thead>
<tr>
<th>Octave band center frequency (Hz)</th>
<th>125</th>
<th>250</th>
<th>500</th>
<th>1000</th>
<th>2000</th>
<th>4000</th>
<th>A-weighted</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Small classrooms</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean ( L_N ) (dB)</td>
<td>38.3</td>
<td>32.4</td>
<td>28.2</td>
<td>26.1</td>
<td>22.3</td>
<td>19.4</td>
<td>32.3</td>
</tr>
<tr>
<td>Maximum ( L_N ) (dB)</td>
<td>48.8</td>
<td>39.3</td>
<td>34.5</td>
<td>32.6</td>
<td>27.5</td>
<td>21.3</td>
<td>38.5</td>
</tr>
<tr>
<td><strong>Medium classrooms</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean ( L_N ) (dB)</td>
<td>40.2</td>
<td>33.7</td>
<td>27.8</td>
<td>24.4</td>
<td>22.7</td>
<td>19.9</td>
<td>32.7</td>
</tr>
<tr>
<td>Maximum ( L_N ) (dB)</td>
<td>53.4</td>
<td>43.6</td>
<td>43.7</td>
<td>40.1</td>
<td>37.3</td>
<td>32.4</td>
<td>43.5</td>
</tr>
<tr>
<td><strong>Large classrooms</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean ( L_N ) (dB)</td>
<td>45.1</td>
<td>37.9</td>
<td>33.5</td>
<td>32.0</td>
<td>28.3</td>
<td>21.9</td>
<td>37.6</td>
</tr>
<tr>
<td>Maximum ( L_N ) (dB)</td>
<td>51.5</td>
<td>46.2</td>
<td>41.1</td>
<td>37.4</td>
<td>30.1</td>
<td>23.2</td>
<td>43.5</td>
</tr>
</tbody>
</table>

in all of the three classroom groups. The small classrooms had the highest STI, which falls in the category of “excellent”. The medium classrooms had an average STI rating which is between “good” and “excellent”, and the sports halls had an STI rating of “good”, which is most likely to decrease in the presence of activity noise.

### E. Voice support and room gain

#### 1. Measurements

The mean and standard deviation of \( ST_V \) and \( G_{RG} \) in the octave bands between 125 Hz and 4 kHz measured in the classrooms are shown in Table VI. The frequency characteristics of \( ST_V \) and \( G_{RG} \) are similar for small and medium classrooms, with an increase of the values at high frequencies. The only difference between the two classroom groups is that the small classrooms have a slightly higher overall value. The large classrooms (sports halls) have an overall lower value and, in addition, the frequency characteristic is qualitatively different, because the low frequencies are predominant. This indicates that these large rooms do not reflect efficiently the high frequencies of a speaker. The spread of \( ST_V \) among rooms does not depend on the frequency band, because the standard deviation does not present a frequency-dependent pattern in the different classroom groups. However, the standard deviation of \( G_{RG} \) is proportional to its absolute value.

#### 2. Prediction model

The values of \( V \) and \( S \) of each classroom, together with the frequency-dependent average measurements of \( T \), were used in connection with Eq. (19) to predict the \( ST_V \) values. The comparison between the measured and the predicted values of \( ST_V \) in the octave bands between 125 Hz and 4 kHz is shown in Fig. 3. The most accurate predictions are found in the most important bands for speech (between 500 Hz and 2 kHz). In these bands, \( R^2 \) was at least 0.8, the residual deviation was not higher than 1.2 dB, and the bias or deviation from the unbiased prediction was lower than 2 dB. The prediction for the 125 Hz band had a large uncertainty, shown by the low value of \( R^2 \) (0.18), and large residual deviation (3.3 dB) and bias (4.3 dB).

The speech-weighted \( ST_V \) predictions are plotted in Fig. 4 as a function of the measured \( ST_V \) values. The regression line relating measurements and predictions had a slope of 1 and an offset of 0.36 dB. The \( R^2 \) was 0.84, the residual error was 1.1 dB and the bias was 1.2 dB.

### V. DISCUSSION

The acoustic properties of school classrooms described in the results section correspond to typical primary and secondary schools in southern Sweden built during the 1970s. The background noise levels in almost three fourths of the small and medium sized classrooms were below 35 dBA—which is the maximum acceptable value of different guidelines, e.g., the standard ANSI S12.60-2002 in the US, the Building Bulletin 93 in the UK, or the guidelines from the World Health Organisation.
TABLE V. Reverberation times ($T$) measured in the classrooms.

<table>
<thead>
<tr>
<th>Octave band center frequency (Hz)</th>
<th>125</th>
<th>250</th>
<th>500</th>
<th>1000</th>
<th>2000</th>
<th>4000</th>
<th>Average 500–2000</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Small classrooms</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean $T$ (s)</td>
<td>0.59</td>
<td>0.39</td>
<td>0.32</td>
<td>0.34</td>
<td>0.35</td>
<td>0.34</td>
<td>0.33</td>
</tr>
<tr>
<td>s.d.</td>
<td>0.42</td>
<td>0.14</td>
<td>0.04</td>
<td>0.05</td>
<td>0.05</td>
<td>0.02</td>
<td>0.05</td>
</tr>
<tr>
<td><strong>Medium classrooms</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean $T$ (s)</td>
<td>0.72</td>
<td>0.53</td>
<td>0.45</td>
<td>0.47</td>
<td>0.47</td>
<td>0.44</td>
<td>0.46</td>
</tr>
<tr>
<td>s.d.</td>
<td>0.33</td>
<td>0.17</td>
<td>0.08</td>
<td>0.08</td>
<td>0.07</td>
<td>0.07</td>
<td>0.08</td>
</tr>
<tr>
<td><strong>Large classrooms</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean $T$ (s)</td>
<td>1.46</td>
<td>1.58</td>
<td>1.59</td>
<td>1.55</td>
<td>1.35</td>
<td>1.04</td>
<td>1.57</td>
</tr>
<tr>
<td>s.d.</td>
<td>0.24</td>
<td>0.35</td>
<td>0.29</td>
<td>0.18</td>
<td>0.07</td>
<td>0.07</td>
<td>0.23</td>
</tr>
</tbody>
</table>

TABLE VI. Frequency band values and overall speech-weighted voice support ($ST_V$) and room gain ($G_{RG}$) measured in the classrooms.

<table>
<thead>
<tr>
<th>Octave band center frequency (Hz)</th>
<th>125</th>
<th>250</th>
<th>500</th>
<th>1000</th>
<th>2000</th>
<th>4000</th>
<th>Speech-weighted</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Small classrooms</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean $ST_V$ (dB)</td>
<td>-9.4</td>
<td>-11.1</td>
<td>-9.5</td>
<td>-7.6</td>
<td>-6.4</td>
<td>-4.6</td>
<td>-5.6</td>
</tr>
<tr>
<td>s.d.</td>
<td>0.46</td>
<td>0.81</td>
<td>0.91</td>
<td>0.38</td>
<td>0.72</td>
<td>1.04</td>
<td>0.78</td>
</tr>
<tr>
<td>Mean $G_{RG}$ (dB)</td>
<td>0.50</td>
<td>0.33</td>
<td>0.47</td>
<td>0.70</td>
<td>0.91</td>
<td>1.31</td>
<td>1.06</td>
</tr>
<tr>
<td>s.d.</td>
<td>0.02</td>
<td>0.06</td>
<td>0.09</td>
<td>0.06</td>
<td>0.13</td>
<td>0.25</td>
<td>0.16</td>
</tr>
<tr>
<td><strong>Medium classrooms</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean $ST_V$ (dB)</td>
<td>-12.1</td>
<td>-13.9</td>
<td>-13.5</td>
<td>-11.6</td>
<td>-10.9</td>
<td>-9.1</td>
<td>-10.2</td>
</tr>
<tr>
<td>s.d.</td>
<td>1.46</td>
<td>1.27</td>
<td>1.43</td>
<td>1.68</td>
<td>1.75</td>
<td>1.52</td>
<td>1.58</td>
</tr>
<tr>
<td>Mean $G_{RG}$ (dB)</td>
<td>0.28</td>
<td>0.18</td>
<td>0.20</td>
<td>0.32</td>
<td>0.37</td>
<td>0.54</td>
<td>0.42</td>
</tr>
<tr>
<td>s.d.</td>
<td>0.10</td>
<td>0.06</td>
<td>0.07</td>
<td>0.13</td>
<td>0.16</td>
<td>0.19</td>
<td>0.16</td>
</tr>
<tr>
<td><strong>Large classrooms</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean $ST_V$ (dB)</td>
<td>-10.8</td>
<td>-16.0</td>
<td>-18.2</td>
<td>-19.1</td>
<td>-19.5</td>
<td>-19.4</td>
<td>-18.8</td>
</tr>
<tr>
<td>s.d.</td>
<td>1.56</td>
<td>1.91</td>
<td>0.92</td>
<td>1.31</td>
<td>1.40</td>
<td>1.31</td>
<td>1.01</td>
</tr>
<tr>
<td>Mean $G_{RG}$ (dB)</td>
<td>0.36</td>
<td>0.12</td>
<td>0.07</td>
<td>0.06</td>
<td>0.05</td>
<td>0.06</td>
<td>0.058</td>
</tr>
<tr>
<td>s.d.</td>
<td>0.14</td>
<td>0.06</td>
<td>0.01</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.01</td>
</tr>
</tbody>
</table>

FIG. 3. Expected versus measured values of voice support in frequency bands. The solid lines show the regression lines for the predictions and the dotted lines indicate the ideal and unbiased prediction lines.
The background noise levels in the remaining fourth of classrooms were half below 40 dB and half between 40 and 45 dB. The average value was 32.6 dB, which is lower than the 45 to 48 dB reported by Shield and Dockrell\(^1\) in their review from several surveys on empty classrooms (without acoustical treatment).

In small and medium-sized classrooms, \(T\) did not exceed 0.6 s, in fulfilment of different guidelines of classroom acoustic design.\(^{29-31}\) Reverberation times and background noise levels are within the recommended values in most of the cases. This seems to be reflected in the non-problematic perception of classroom acoustics by teachers without voice problems in schools of the same region in Sweden.\(^{13}\) The Swedish standard for acoustic conditions in classrooms\(^32\) is more strict, requiring reverberation times below 0.5 s for the octave frequency bands above 250 Hz and below 0.6 s at 125 Hz, which only a few of the classrooms fulfill.

The STI measured in classrooms with high signal-to-noise ratios was higher than 0.6 in all cases, even in the sports halls. However, the subjective speech intelligibility with ongoing activity, specially in the sports halls, will be lower than predicted, due to an actual lower signal-to-noise ratio under these conditions. Unfortunately, none of the current guidelines specify the signal-to-noise ratio that should be used for the assessment of STI.

The prediction model for \(ST_V\) has been derived theoretically and it has been assessed by comparing its predictions with actual \(ST_V\) measurements. There is a slight bias in the prediction, as the regression line of measured versus predicted \(ST_V\) is not \(\overline{ST_V}_{\text{pred}} = 1.0 \cdot \overline{ST_V}_{\text{meas}} - 0.36\) (see Fig. 4). This bias results in a deviation of 1.2 dB from the actual values, slightly higher than the residual deviation (1.1 dB). Taking into account that the measurement dataset has not been used to derive the model, the predictions are reasonably accurate.

In the range of medium-sized classrooms (with volumes \(100 < V < 250\) m\(^3\)), \(G_{\text{RHC}}\) is in the range between 0.2 and 0.5 dB, whereas \(ST_V\) is in the range between -14 and -9 dB. There is some spread of data in this range, as seen in Fig. 4. Measured \(ST_V\) values can deviate as much as 3 dB from the predicted value. \(ST_V\) is influenced by the early reflections which can not be accurately represented with a statistical model such as the one in Eq. (19).

The voice support, analogously to the objective support in concert halls, is not a stand-alone parameter to design classroom acoustics. It is a magnitude related to the additional vocal load that teachers experience while speaking in a classroom due to the acoustic conditions. Other magnitudes, like \(T\), STI, sound strength, and background noise levels, should be taken into account as well. There is not enough scientific evidence to establish a definite range of recommended values of \(ST_V\), but the range between -14 and -9 dB obtained in most of the medium-sized classrooms seems adequate, since \(T\) and STI fulfilled the recommendations without the rooms being too damped. Using the graph in Fig. 2, for a room of 100 m\(^3\), the range of \(-14 < ST_V < -9\) dB corresponds to reverberation times in the range 0.25 < \(T\) < 0.6 s. For a room of 300 m\(^3\), the same range of \(ST_V\) corresponds to the range 0.55 < \(T\) < 1.4 s. In this last case, the design criteria should be to aim at the highest reverberation time that does not compromise speech intelligibility, because too high values of reverberation are detrimental to speech intelligibility. For the same reason, it is not advisable to aim at values of \(ST_V\) higher than -9 dB. However, in very small classrooms, \(ST_V\) may be higher than -9 dB without compromising speech intelligibility.

### VI. CONCLUSIONS

The present paper has measured and provided a reference set for voice support and room gain values, which are important parameters to assess the vocal effort required to speak in a room. The voice support in classrooms of good acoustical quality, with volumes between 100 and 250 m\(^3\), has been found to be in the range between -14 and -9 dB, and the room gain in the range between 0.2 and 0.5 dB.

A model, derived from the diffuse field theory, has been developed to predict average values of voice support in classrooms. The model is based on geometrical room properties of volume, total surface area, and reverberation time. It points out necessary geometrical restrictions in rooms to obtain good acoustic conditions both for a listener (in terms of reverberation time) and for a speaker (in terms of voice support). The model describes the present voice support measurements in classrooms with a coefficient of determination of 0.84 and a standard deviation of 1.2 dB.

### Acknowledgments

This research has been funded by the Swedish organization AFA Försäkring as a part of the project “Speaker’s comfort and voice health in classrooms”. Tobias Olesen from DTU has given some important feedback to improve the content of the paper. The authors would like to ex-
press their gratitude to all the schools and staff who made these measurements possible.

5aSC5. Influence of Classroom Acoustics on the Voice Levels of Teachers With and Without Voice Problems: A Field Study

David Pelegrin-Garcia*, Viveka Lyberg-Åhlander, Roland Rydell, Jonas Brunskog and Anders Lofqvist

*Corresponding author’s address: Technical University of Denmark, Building 352, Oersteds plads, Kongens Lyngby, DK-2800, Kongens Lyngby, Denmark, dpg@elektro.dtu.dk

Many teachers suffer from voice problems and classroom acoustics has been considered as one of the potential hazards for this. The present study examines how classroom acoustics interacts with the voices of 14 teachers without voice problems and 13 teachers with voice problems. The assessment of the voice problems was made with a questionnaire and a laryngological examination. During teaching, the sound pressure level at the teacher’s position was monitored. The teacher’s voice level and the activity noise level were separated using mixed Gaussians. In addition, objective acoustic parameters of Reverberation Time and Voice Support were measured in the 30 empty classrooms of the study. An empirical model shows that the measured voice levels depended on the activity noise levels and the voice support. Teachers with and without voice problems were differently affected by the voice support of the classroom. The results thus suggest that teachers with voice problems are more aware of classroom acoustic conditions than their healthy colleagues and make use of the more supportive rooms to lower their voice levels. This behavior may result from an adaptation process of the teachers with voice problems to preserve their voices. [Work supported by AFA.]
INTRODUCTION

Voice is the primary working tool of teachers, and a good voice is essential for communicating with students. Nowadays, many teachers suffer from voice problems. A recent study reported that around 13% of the active school teachers in southern Sweden self-reported voice problems [1]. Voice health problems are a major concern, not only due to the required clinical assistance and the personal consequences in job dissatisfaction and lack of self-esteem, but also due to the financial impact that the teachers’ absence produces in the global budget of the country [2]. Investigating possible causes for voice disorders from the testimonies of affected teachers, Vilkman points out “bad classroom acoustics” as one of the hazards for voice health [3].

The present study analyzed the average voice levels used at work by teachers with and without voice problems as a function of relevant environmental acoustic parameters. Two acoustic parameters were considered important: the activity noise level, due to the presence of students and other noise sources during teaching, and the voice support offered by the classroom. Three steps were necessary in the study: first, the choice of teachers and the assessment of voice problems. Second, the monitoring of the teacher’s voice levels and the activity noise levels during teaching, and last, the measurement of objective acoustic parameters in the empty classrooms.

METHOD

Choice of teachers

A total of 27 teachers in 5 different schools in the south of Sweden, at educational levels ranging from primary school to high school, were considered for this study. The participants were selected as a follow-up to an epidemiological study [1].

The teachers were classified into two groups: one group (test; $N_T = 13$, 2 male/11 female) containing the teachers with voice problems and another group (control, $N_C = 14$; 2 male/12 female) with those teachers having no remarkable voice problems. The assessment of voice problems was made by means of the VHI-T (Voice Handicap Index with Throat subscale) questionnaire [4] and a laryngological examination.

Measurements during teaching

The teachers were equipped with an IEC 61672-compliant, type 2, sound level meter SVANTEK SV-102. This device measured and stored the A-weighted sound pressure level (SPL), using an exponential averaging with “fast” time constant, sampled at 1 s intervals. The microphone capsule was attached to the teachers’ clothing neck, as a lapel microphone, at a distance of about 15 cm from the mouth.

The sound level meter operated for one working day. For each teacher, two SPL sequences were studied. One of them corresponded to a lesson at the beginning of the day and another one to a lesson at the last hour. The duration of the lessons was between 30 and 45 minutes. An example sequence is shown in Fig. 1 and the corresponding histogram is shown as gray bars in Fig. 2.

In these SPL sequences, it was assumed that the SPL from the teacher’s voice was several dB higher than the SPL from activity (originated from students, ventilation noise and other external sources), because of the closer placement of the microphone to the teacher’s mouth (around 15 cm). The time fraction while the teacher was talking was noted as $\alpha$. The activity levels were obtained while the teacher was silent, during a time fraction $1 - \alpha$.

The teacher’s voice ($S$) and activity noise ($N$) levels were assumed to be random processes coming from normal distributions, with probability density functions $f_S(L)$ and $f_N(L)$, respectively, where $L$ indicates the A-weighted SPL. The means of these distributions are notated $L_{50,S}$ and $L_{50,N}$ (the symbol $L_{50}$ indicates the level that is exceeded during 50% of the time, also referred to as median level), and their standard deviations $\sigma_S$ and $\sigma_N$. As an example, these distributions are indicated in Fig. 2 with dash-dot and dashed lines, respectively. Thus,

\[
S \sim N(L_{50,S}; \sigma_S) \rightarrow f_S(L), \tag{1}
\]

\[
N \sim N(L_{50,N}; \sigma_N) \rightarrow f_N(L). \tag{2}
\]
FIGURE 1. A-weighted SPL at the lapel microphone worn by the teacher during one lesson

FIGURE 2. In gray, histogram computed from the A-weighted SPL values in Fig. 1. On top, scaled normal probability density functions corresponding to the activity noise (dashed line), the teacher’s voice (dash/dot line), and the addition of both processes (solid line).

The joint process corresponding to the observed A-weighted SPL values was regarded as having a probability density function \( f_{S+N}(L) \), obtained by overlapping the two normal distributions \( f_S(L) \) and \( f_N(L) \), scaled by their probability of occurrence in time (\( \alpha \) and \( 1 - \alpha \), respectively):

\[
f_{S+N}(L) = \alpha f_S(L) + (1 - \alpha) f_N(L).
\]  

According to this principle, a linear combination of two normal distributions was fitted to the A-weighted SPL histogram, by minimizing the squared error with the simplex algorithm implemented in the function \texttt{fminsearch} of \textsc{Matlab}. In this way, there were 5 estimated parameters (\( L_{50,S}, \sigma_S, \sigma_N, \) and \( \alpha \)) for each sequence, although only the A-weighted median levels for the teacher’s voice (\( L_{50,S} \)) and the activity noise (\( L_{50,N} \)) were used in the analysis. As an example, the probability density function fitted to the measured A-weighted SPL is shown with a solid line in Fig. 2. A similar approach to determine speech and noise levels in classrooms has been previously used [5].

**Classroom acoustic measurements**

Acoustic measurements were performed in the 30 classrooms where the teachers held their lessons, while they were empty.

**Reverberation time.** The reverberation time (RT) was calculated according to the standard ISO 3382-2 [6]. The sound source was a B&K Omnisource type 4295, placed at the teacher’s position and with the radiating opening at a
height of 1.6 m. Two 1/2” pressure-field microphones B&K type 4192 were used as receivers and were placed close to students’ seats at a height of 1.2 m. The 01dB Symphonie system, incorporating the MLS software module, was used to produce the measurement signal and send it to the loudspeaker via a power amplifier, acquire the signal from the microphones, calculate the impulse responses, and derive the $RT_{20}$. The measured RT values in the classrooms, corresponding to the average of the 500 Hz and 1 kHz octave frequency bands, are shown in Fig. 3. However, the RT was not used in the empirical model due to the lack of normality in the measured values. The three ‘outliers’ in reverberation time correspond to three sports halls that were used for gymnastics lessons.

Voice support. Instead, the focus in this research was on characterizing the acoustic conditions of classrooms as perceived by the teachers while talking. A parameter called Voice Support ($STV$) is introduced in this paper as a measure of how much the sound reflections at the room boundaries amplify the voice of the teacher at his/her ears (NOTE: The exact definition of $STV$ is given below).

The voice support is calculated from an impulse response corresponding to the airborne sound transmission between the mouth and the ears (or simply, mouth-to-ears impulse response). For this purpose, a Head and Torso Simulator (HaTS) B&K type 4128 was used. The HaTS included a loudspeaker at its mouth, and microphones at its ears. The HaTS was placed at a representative teaching position, with the mouth at a height of 1.5 m. The 01dB Symphonie system was used to produce the excitation signal and determine the mouth-to-ears impulse response from the measured signal at the microphones. The setup used to measure the mouth-to-ears impulse response is shown in Fig. 4.

From the measured mouth-to-ears impulse response $h(t)$ (example shown in Fig. 5), the direct sound $h_d(t)$ is obtained by applying a window $w(t)$ to the measured impulse response $h(t)$,

$$h_d(t) = h(t) \times w(t),$$

(4)
FIGURE 5. Example of a measured mouth-to-ears impulse response, with the windowing applied in order to calculate the direct and the reflected airborne sound components of one’s own voice.

where \( w(t) \) is

\[
w(t) = \begin{cases} 
1 & t < 4.5 \text{ ms} \\
0.5 + 0.5 \cos \left( \frac{2\pi(t - t_0)}{T} \right) & 4.5 \text{ ms} < t < 5.5 \text{ ms} \\
0 & t > 5.5 \text{ ms}
\end{cases}
\]  

(5)

with \( t_0 = 4.5 \text{ ms} \) and \( T = 2 \text{ ms} \). The reflected sound \( h_r(t) \) is the complementary signal

\[ h_r(t) = h(t) \times (1 - w(t)) = h(t) - h_d(t) \]  

(6)

From the above signals, the energy levels corresponding to the direct sound \( (L_{E,d}) \) and the reflected sound \( (L_{E,r}) \) are calculated as

\[ L_{E,d} = 10 \log \frac{\int_0^\infty h_d^2(t) \, dt}{E_0}, \]  

(7)

\[ L_{E,r} = 10 \log \frac{\int_0^\infty h_r^2(t) \, dt}{E_0}. \]  

(8)

From these two equations, the voice support \( ST_V \), in analogy to Gade’s objective support [7], is defined as the difference between the reflected sound and the direct sound from the mouth-to-ears impulse response,

\[ ST_V = L_{E,r} - L_{E,d}. \]  

(9)

The \( ST_V \) values measured in the 30 classrooms of the study, averaged for two HaTS positions and the two ears, without applying any filtering, are shown in Fig. 6. The average value is indicated with a solid line, whereas one standard deviation above and below the mean is indicated with dashed lines.

**Statistical method**

We used a multiple regression to analyze the combined influence of the covariates voice support \( (ST_V) \) and median activity noise \( (L_{50,N}) \) on the teachers’ median voice levels \( (L_{50,S}) \). The two covariates \( ST_V \) and \( L_{50,N} \) were fairly uncorrelated \( (\rho = -0.07) \). Additionally, we accounted for possible differences in voice use between the teachers of the test and control groups (with and without voice problems) by including a binary variable named Test/Control which indicated which group the teacher belonged to.

Since we considered the effect of \( ST_V \) and \( L_{50,N} \) to be potentially different for the teachers of the test and control groups, we included also the interaction between the Test/Control variable and the two covariates. Nevertheless, the
interaction between $L_{50,N}$ and Test/Control was found to be non-significant ($F_{1,48} = 0.15, P = 0.70$) and was left out from the final model.

We fitted the model in R [8] using the function $\text{lm}$. Prior to running the model, we applied the square root, affine transformation to the activity noise levels $\sqrt{75 - L_{50,N}}$, in order to obtain an approximately normal distribution of the observed values of the covariate. None of the measured noise levels was higher than 75 dB. This transformed variable, and $STV$, which already presented an absence of outliers and skew, were further $z$-transformed. We checked various diagnostics of model validity and stability (Cook’s distance, dfits, distribution of residuals, residuals plotted against predicted values) and none of these indicated obvious influential cases or outliers, nor obvious deviations from the assumptions of normality and homogeneity of residuals [9]. The significance of each variable in the model was assessed by means of F-tests resulting from an analysis of variance.

### RESULTS

Overall, the median voice levels were clearly influenced by the combination of predictor variables in the proposed statistical model ($R^2 = 0.69, F_{4,49} = 27.8, p < 0.001$):

\[
L_{50,S}(\text{test}) = 81.3 - 3.87 \times \sqrt{75 - L_{50,N}} - 0.72 \times STV \ [\text{dB}], \quad (10a)
\]

\[
L_{50,S}(\text{control}) = 102.9 - 3.87 \times \sqrt{75 - L_{50,N}} + 0.84 \times STV \ [\text{dB}]. \quad (10b)
\]

The effect of the transformed noise levels on the voice levels ($F_{1,49} = 92.2, p < 0.001$) was highly significant. The overall effect of the covariate voice support $STV$ ($F_{1,49} = 0.65, p = 0.43$) and the factor Test/Control ($F_{1,49} = 2.12, p = 0.15$) were not significant at the 5% level. However, the interaction between the $STV$ and the Test/Control variable was found to be highly significant ($F_{1,49} = 16.5, p < 0.001$).

The measured $L_{50,S}$ values as a function of $STV$ are shown in Fig. 7. For the average observed noise levels ($L_{50,N} = \overline{L}_{50,N}$), the model (10) is:

\[
L_{50,S}(\text{test}) = 69.8 - 0.72 \times STV \ [\text{dB}], \quad (11a)
\]

\[
L_{50,S}(\text{control}) = 91.4 + 0.84 \times STV \ [\text{dB}]. \quad (11b)
\]

For teachers without voice problems (control group), the median voice levels increased with the measured voice support at a rate of 0.8 dB/dB. On the other hand, teachers with voice problems (test group) lowered their voice levels the higher the voice support, at a rate of -0.7 dB/dB.
FIGURE 7. Median voice SPL used by teachers versus voice support measured in the empty classrooms. The solid lines show the regression model in (11). The two teacher groups make use of the voice support in significantly different ways.

FIGURE 8. Median voice SPL used by teachers versus median activity noise SPL. The solid lines show the regression model (12). As a consequence of the Lombard effect, the voice levels increase with the noise levels, equally for teachers with and without voice problems. However, teachers in the control group use higher voice levels than those in the test group.

The measured $L_{50,S}$ values as a function of $L_{50,N}$ are shown in Fig. 8. For the average observed voice support ($STV = \bar{STV}$), the model (10) is:

$$L_{50,S}(\text{test}) = 90.6 - 3.87 \times \sqrt{75 - L_{50,N}} \, [\text{dB}],$$

(12a)

$$L_{50,S}(\text{control}) = 92.0 - 3.87 \times \sqrt{75 - L_{50,N}} \, [\text{dB}].$$

(12b)

For all teachers, there was an increase of median voice level with the activity noise present during teaching. This increase was non-linear in the observed range of levels, being more relevant for the highest noise levels. Additionally, the teachers from the test group talked 1.4 dB on average softer than the teachers in the control group. However, this difference was not statistically significant with the number of teachers considered in this study.
DISCUSSION

Teachers from the test group (with voice problems) decreased their voice levels with increasing voice support (-0.7 dB/dB) in the classrooms, as opposed to the control group (without voice problems, 0.8 dB/dB). The behavior of the test group would be desirable for the prevention of voice problems. The measurements suggest that teachers from the test group made good use of the voice support as an adaptive mechanism to preserve their vocal health. This finding supports the results from a study by Kob et al. [10], who found that teachers with voice problems were more affected by poor classroom acoustics than their healthy colleagues. The behavior of the teachers in the test group follows the results of Brunskog et al. [11], who found that teachers lowered their voice levels as a function of the amplification offered by the room to their own voice. However, the behavior of teachers in the control group does not follow a logical pattern. A hypothetical answer would be that the voice support increases in rooms with sound reflecting boundaries, and the activity noise levels would increase in this case. Due to the Lombard effect, the talkers (students and teacher) would perceive increased noise levels and automatically raise their voices. However, the lack of correlation between voice support and activity noise invalidates this hypothesis.

Teachers from the test and control groups were equally affected by noise. Both groups increased their vocal intensity with increasing activity noise, in accordance with the Lombard effect. If the curves are approximated by straight lines for $L_{50,N}$ above 55 dB, the slope is 0.6 dB/dB, in good agreement with the literature (for example, Lazarus reports slopes between 0.5 dB/dB and 0.7 dB/dB [12]). The teachers from the test group talked on average 1.4 dB softer than the control group, although this difference was not significant. Nevertheless, this might be an additional indication that teachers with voice problems tried to limit their vocal effort in terms of vocal intensity.

CONCLUSIONS

The main conclusions from the field study are the following:

- Teachers with voice problems make a more efficient use of the voice support in classrooms than their healthy colleagues, probably as an adaptive mechanism to preserve their voice health.
- Teachers with and without voice problems react identically to changes in activity noise, according to the Lombard effect.

ACKNOWLEDGMENTS

The authors wish to express their gratitude to all the teachers and staff of schools participating in the measurements. This research has been funded by the Swedish company AFA Försäkring as a part of the project “Speaker’s comfort and voice health in classrooms”.

REFERENCES

Loudspeaker-based system for real-time own-voice auralization

David Pelegrín-García and Jonas Brunskog

Acoustic Technology, Department of Electrical Engineering, Technical University of Denmark, Kongens Lyngby DK-2800, Denmark

(Dated: March 30, 2011)

In order to investigate the influence of room acoustic conditions on voice production, a system for the real-time auralization of one’s own voice has been designed. This system combines computerized room acoustic models, psychoacoustic processing, short-delay convolution techniques, mixed-order Ambisonics encoding/decoding, and loudspeaker reproduction. Equalization filters are used on an individual basis to adjust the performance of the system to each particular talker, including the ratio between direct and reflected sound. The auditory cues of delay, amplitude, frequency response, and directionality corresponding to each sound reflection are preserved. Thus, this system is suitable for psychoacoustics and cross-modality research, integration in multimodal virtual reality systems or room acoustics enhancement.

PACS numbers: 43.55.Lb

I. INTRODUCTION

Recently, the field of voice ergonomics, which is defined as the study and action on all the factors that enhance performance in speech communication, decrease risks for voice disorders and enable recovery from a voice disorder, has received some attention. One of the topics of research in this field is the study of the interaction between room acoustics, noise, and voice production. This is of special relevance in the case of teachers, for whom the prevalence of voice problems is significantly higher than in the rest of the population.

In a recent investigation, Brunskog et al. studied the effect of the classroom acoustic conditions on the voice levels of a number of teachers. The same teacher had to move to a number of different rooms, which were sometimes located far away from each other. This methodology may have introduced some bias in the results by making comparisons among rooms or judgments about them slightly difficult. In addition, the acoustic conditions of the rooms were given beforehand, with very limited possibilities of adjusting them.

The present paper describes a tool to accurately recreate the acoustics of different rooms in a controllable way inside a loudspeaker array, located in a highly damped room. This will enable a more careful and flexible design of experimental conditions in research. The recreation of different room acoustic conditions is based on the reconstruction of the sound field of the simulated room using the method proposed by Favrot and Buchholz, although introducing slight modifications for the requirements of real-time performance. The reconstruction of the sound field is focused mainly on the voice of a talker at his own ears, so he/she has the experience of being in an acoustic environment different from the actual room. According to Kleiner et al., this system aims to "auralize" the talker’s voice in real time.

Previous auralization systems with the same aim have been reported in the literature. Kleiner and Bertson used a system with nine loudspeakers that could provide up to 50 early reflections obtained from delay lines. Shearer and Torres used a two-channel, headphone-based auralization system able to convolve in real time the voice of a talker with an impulse response calculated with a room acoustic simulation software. In a more recent work, Cabrera et al. used a pair of earspeakers to render a binaurally recorded sound-field, with the possibility of accounting for head rotations by means of head-tracking in the horizontal plane.

Similar systems have been built to investigate the importance of room acoustic conditions for singers. Marshall and Meyer used a system with 7 loudspeakers and 4 microphones that simulated 4 early reflections and late reverberation, with the particularity of allowing the presence of several performers at the same time. Noson et al. studied the preference of singers after introducing an additional reflection in realistic environments, with the aid of a microphone, a delay line and a loudspeaker. In more recent works, Yuen et al. and Stetson and Braasch used a two-channel convolution system able to recreate binaural sound fields through binaural impulse response measurements in real halls.

Other investigations, not focused on the talker’s voice, but on the effect of room acoustics on musical performance and subjective preference of musicians in stage, have used similar setups. Gade used a system with five loudspeakers and a microphone to generate sound fields consisting of a single reflection and a reverberation tail. Ueno and Tachibana designed a 6-loudspeaker system to simulate sound fields obtained through the measurement of the corresponding impulse responses in real rooms.

During the past few years, many technological advances have made it possible to implement techniques which were previously known but not technically possible. As an example, state-of-the-art PCs have sufficient processing capability to perform a number of simultaneous convolutions efficiently, without expensive and dedi-
cated DSP, as required one decade ago\textsuperscript{16}. There are several free software open source solutions available to perform efficient multiple channel convolutions with very low delay\textsuperscript{17,18}. The release of new multi-channel digital audio standards such as MADI\textsuperscript{7}, in combination with multi-channel sound cards, has simplified the connections from the system, expanded the possibilities of centralized convolution systems, and made the technology affordable for a larger number of people. In addition, state-of-the-art room acoustics simulation software provides fairly accurate predictions of the sound-fields in rooms\textsuperscript{19,20}. The system presented in this paper takes advantage of all these innovations to perform the real-time convolution of the own voice with a 29-channel simulated impulse response that, reproduced through 29 loudspeakers, generates the reflected 3D sound field of one’s own voice. These components is added to the sound of one’s own voice propagated directly through the air or through the body. The reconstruction of the reflected sound field is made according to a realistic approach. It combines the output of a room acoustics simulation program\textsuperscript{21} with the spatial and psychoacoustic decoding scheme proposed by Favrot and Buchholz\textsuperscript{4}, thus preserving delay, amplitude, spectrum, and directional cues of the simulated reflections. Very long impulse responses can be used, so the system does not put a restriction of the maximum length for practical use in room acoustics.

II. SYSTEM DESCRIPTION

A. Overview

A block diagram illustrating the overall real-time auralization system is shown in Fig. 1. As can be seen, there are two main parts, namely the pre-processing stage and the real-time processing, acquisition, and reproduction stage. The first part includes all the necessary steps to obtain the impulse responses of the environment that will be used in the auralization. This includes the design of a computerized room acoustic model, the calculation of an impulse response with room acoustics software and its encoding and decoding with mixed-order Ambisonics techniques for the given layout of the loudspeaker reproduction system. The second part contains all the elements of the system that apply the desired room impulse response to a talker’s speech signal in real time. These are: an acquisition part with a microphone and a sound card, a real-time section with a software convolver and an equalizer filter, and a reproduction system based on 29 loudspeakers.

B. Pre-processing

A very important part of the auralization system is the offline calculation, decoding, and storage of an accurate set of impulse responses ready to be used in the second block, which applies the room effect to a talker’s voice in real-time. This part of the system is an adaptation of the LoRA toolbox designed by Favrot and Buchholz\textsuperscript{4}. The LoRA toolbox is a software application that uses the output (impulse response with directional information) of an acoustics simulation program to encode it in Ambisonics and decode it to a particular reproduction layout, producing an IR for each loudspeaker. However, some modifications in the procedure and calculation are needed in order to match the requirements for self-voice auralization.

First, a computer-based room acoustic model is needed, which is then loaded into an acoustic simulation program. In the proposed system, Odeon is used\textsuperscript{21}, although other alternative solutions may also be used, as long as the interface with the LoRA toolbox is implemented satisfactorily. In the acoustic simulation, the source is located at the talker’s position, avoiding positions too close to the boundaries that could not be satisfactorily reproduced by the system due to the inherent latency (analyzed in section III.B). The receiver point is located 1 m in front of the source. Note that this position does not correspond to the position of the ears relative to the mouth (sound source). However, the reflection pattern is reasonably similar to the reflection pattern experienced at the position of the ears. In addition, the proposed calibration method takes advantage of this approximation, as will be discussed in section III.C. The source is oriented toward the audience and has a directivity pattern similar to the average human speech\textsuperscript{22,23}.

For rooms in a volume range of approximately $100 \text{ m}^3 < V < 1000 \text{ m}^3$, the used simulation parameters are 5000 rays, a maximum reflection order of 2000, a transition order of 3 reflections between early reflections and late reverberation, and a histogram resolution of 10 ms for the late reverberation. The length of the response is adjusted to correspond at least to the largest reverberation time among all frequency bands for the simulated room. The early part of the response is calculated through the image source method and the late part by ray-tracing. Although 5000 rays are usually a low number in this kind of simulations, it is not of critical importance here, since the fine structure of the late reverberation is not of interest, but only the envelope of the energy-time curve.

FIG. 1. Block diagram of the real-time convolution system.
The acoustic simulation program exports the discrete early reflections separately, each one with its delay, direction of incidence, and attenuation per frequency band. The late reverberation is exported as vectorial intensity (i.e., in first order Ambisonics format WXYZ) in each of the standard octave frequency bands from 63 Hz to 8 kHz at the defined time intervals. The combination of these two components is referred to as the **Directional IR** in Fig. 1. The LoRA toolbox is adapted to omit the direct sound from these files, because it will be produced by the talker himself during the real-time auralization. The early reflections are then encoded in fourth order Ambisonics and decoded into the corresponding loudspeaker layout for reproduction (see Fig. 6 in²). The envelope of the reverberation tail is decoded with a lower directional accuracy (first order Ambisonics) than the early reflections, which leads to a higher degree of diffuseness in the resulting multichannel IR. The decoded envelopes are filled with noise sequences uncorrelated among the different channels, in order to avoid coherent interference effects and coloration of the sound. The late reverberation is added to the early reflections and the resulting impulse responses for each loudspeaker are stored as separate WAV files with a 32 bit precision and a sampling frequency of 44.1 kHz.

### C. Real-time acquisition, processing, and reproduction

The real-time operations in the system can be separated into signal acquisition, processing (convolution), and reproduction.

1. **Signal acquisition**

The talker's speech signal is picked up with a headworn microphone DPA 4066-F, placed on the talker's cheek, then digitalized at 44.1 kHz/24 bit with a Behringer ADA8000 and sent into a PC with a RME HDSP ADA8000 and sent into a PC with a RME HDSP MADI audio interface connected to a RME ADI648 (MADI/ADAT converter). Although other placements of the microphone could be more suitable for research, as e.g. in Cabrera et al.⁹, the built-in fitting accessory was quite ergonomic and well adapted to the placement on the cheek. The microphone capsule is close enough to the mouth to avoid any severe influence of feedback (see analysis later in the paper). As in Pörschmann²⁴, the spectral distortion introduced by this placement of the microphone is corrected with an equalizer filter \( h_{EQ}(t) \), which adjusts the spectrum of the speech signal to match the spectrum of the on-axis speech signal at 1 m in front of the mouth. The calculation of the equalizer filter is done on an individual basis, as the placement of the microphone in relation to the mouth of the speaker differs among users. The measurement of the equalizer filter is used also to calibrate the system, as detailed in the next section. The justification for applying the equalizer filter is that the calculation of the impulse response in the simulation program assumes an on-axis source signal to provide a spectrally correct output. For practical reasons, the equalizer filter was pre-convolved with the stored multichannel room impulse responses, reducing the overall delay in the system during run-time operation. Nevertheless, the conceptual representation of Fig. 1 is still valid.

2. **Convolution**

The convolver is the most technically demanding element of the system. It should provide high quality audio, both regarding bit depth and sampling frequency, introduce the lowest possible delay between input and output, and convolve a number of long impulse responses. Lengths of hundreds of thousands of taps are typical for room impulse responses. In the present system, 29 simultaneous convolutions are required (one for each loudspeaker).

To perform the convolutions, a free software convolver—\( jconvolver \)—is used¹⁸. \( jconvolver \) is a multichannel software implementation of the variable block-size convolution scheme proposed by Gardner²⁵. It runs in a four-core PC under Fedora 8 Linux, patched with the real-time kernel module from Planet CCRMA and uses JACK audio server with ALSA sound driver architecture. The convolver is configured with a simple script that defines the input (the speech signal from the microphone), the 29 impulse responses, and adjustments of gain and delay to account for the position of the loudspeakers in the actual arrangement, which are at different distances from the center of the layout. With JACK, each of the outputs of the convolver are assigned to physical outputs of the audio interface.

In order to investigate the demands of the DSP software in relation to the process capability of the hardware (Quad core Intel PC with 8 GB of RAM), a small benchmark study was carried out. In Table I, the CPU load is measured as a function of the minimum block size (identical for JACK and \( jconvolver \)) and length of the impulse response, while calculating 29 impulse responses. In Table II, the CPU load is indicated for each combination of number of channels and minimum block size, for an impulse response of 65536 samples. The CPU load increases with the number of channels and the length of the impulse response, whereas it decreases with the block size. The drawback of the decrease in CPU load is an increase in latency, which is not desirable for real-time convolution. The measured low values of CPU load show that it is possible to run in parallel alternative processes to record or monitor an input or output signal, or also to run multiple instances of \( jconvolver \) in the same computer, so as to simulate more complex auditory environments, for example, adding a second sound source at a different position in the simulated room.

3. **Reproduction**

The output signals are converted into the analog domain with a MADI/ADAT converter RME ADI-648 and four Behringer ADA8000 devices, amplified, and sent to
TABLE I. Benchmarking. CPU load versus different combinations of minimum convolution block size and impulse response length, for 29 parallel convolutions and a 44.1 kHz sampling frequency. The latency introduced by \textit{jconvolver} is indicated in parentheses.

<table>
<thead>
<tr>
<th>IR length</th>
<th>Block size (latency)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>64 (2.9 ms)</td>
</tr>
<tr>
<td>22050</td>
<td>8.7 %</td>
</tr>
<tr>
<td>44100</td>
<td>9.2 %</td>
</tr>
<tr>
<td>88200</td>
<td>10.2 %</td>
</tr>
<tr>
<td>176400</td>
<td>13.4 %</td>
</tr>
</tbody>
</table>

TABLE II. Benchmarking. CPU load versus different combinations of minimum convolution block size and number of channels, for an impulse response of 65536 samples and a 44.1 kHz sampling frequency. The latency introduced by \textit{jconvolver} is indicated in parentheses.

<table>
<thead>
<tr>
<th>Number of channels</th>
<th>Block size (latency)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>64 (2.9 ms)</td>
</tr>
<tr>
<td>4</td>
<td>2.1 %</td>
</tr>
<tr>
<td>8</td>
<td>3.1 %</td>
</tr>
<tr>
<td>16</td>
<td>6.0 %</td>
</tr>
<tr>
<td>32</td>
<td>10.8 %</td>
</tr>
</tbody>
</table>

29 DYNAUDIO BM6 loudspeakers. The loudspeakers are arranged on the surface of a quasi-sphere with distances in the range 1.5 m–2.0 m from the center of the arrangement (see Fig. 2 for specific details of this layout). As the frequency response of the loudspeakers is fairly flat in the frequency range of interest for voice (100 Hz–10 kHz), no equalizers are introduced, as these could be detrimental for the audio quality with small displacements from the equalized position.

III. PRACTICAL CONSIDERATIONS

There are some practical issues that should be addressed so that this auralization system works as intended.

A. Acoustic conditions of the reproduction room

In the first place, the real-time auralization system requires an acoustically dry environment, ideally anechoic, so that the loudspeakers reproduce what they are meant to and not a combination of the simulated room and the test room itself (due to the sound reflections). The physical reproduction room, with dimensions 4.7m × 4.6m × 3.4m, is covered in its whole majority with sound absorbing materials, and its reverberation time, measured according to the standard ISO-3382, is shown in Table III. The value of 0.16 s at 125 Hz could seem a bit high for this application, but due to the fact that the reflected component of one’s own voice in this frequency band is much lower than the sound transmitted directly through the air or through the body, the influence is negligible. As with the loudspeakers, no inverse filtering of room acoustics is applied.

B. Delay / Latency

The term \textit{real-time} applied to this system can lead to some confusion or misunderstanding, as there is actually a certain latency or delay in the system. In a room impulse response, there is usually a time gap between the arrival of the direct sound and the first reflection. If the latency of the system is shorter than this gap, then it is possible to remove a number of samples corresponding to the latency, compensating for this delay without missing any reflection. In our system, the measured latency was 11.5 ms. This delay included the block size used in JACK (64 samples) at the input and the output, the block processing in \textit{jconvolver} (64 samples), the time of sound propagation from the loudspeakers to the ears, and smaller delays in other processes (A/D, D/A, etc).

Considering the sound propagation between the mouth and the ears, a time gap of 11.5 ms between the arrival of the direct sound and the first reflection corresponds to a reflection coming from a boundary at a distance of 2 m. Thus, reflections coming from walls closer than this distance cannot be simulated properly, with precise timing, level, and direction. As a consequence, the smallest volume of a box-shaped room with the source at its cen-
ter that can be accurately simulated is \((2 \times 2)^3 = 64\text{m}^3\). However, smaller rooms are highly dominated by modal effects in a broad frequency range and acoustic simulations with ray-tracing and image-source methods do not perform very accurately for these situations. As a rule of thumb, one limitation of the system is that it cannot simulate rooms smaller than the laboratory room.

Shorter latencies would be desired in this system, although a very obvious limit in our system is imposed by the distance to the loudspeakers. Reducing the distance to the loudspeakers might not be a good solution, because the number of loudspeaker would need to be reduced, reducing the accuracy on directional reproduction, or the loudspeaker would produce much more noticeable physical sound reflections, which should be avoided.

### C. Calibration

A correct calibration is crucial for providing a convincing experience while using the system. The calibration of this system has two goals: on the one hand, adjusting the frequency response in order to compensate for the location of the acquisition microphone in relation to the talker’s mouth, and on the other hand, adjusting the direct-to-reflected energy level difference to match that of a realistic situation.

Given a human speaker, the level difference between the direct sound at a distance of 1 m in front of him and the reflected sound (direct-to-reflected ratio) is noted as \(\Delta L\). This difference must be the same, regardless of the fact that it is obtained by simulation or in a real scenario. The proposed calibration method, summarized in Fig. 3 aims at replicating the level difference obtained by simulation in the real-time auralization system.

The first step, only performed once, is the calculation of an impulse response \(h(t)\) produced by a single reflecting plane in front of a human speaker by means of acoustic simulation,

\[
h(t) \approx \delta(t - t_d) + 10^{\Delta L/20}\delta(t - t_r),
\]

where \(h(t)\) is calculated with the same calculation parameters as for an arbitrary room. That is, the source has the directivity features of a human speaker, and the receiver is located 1 m in front of it. The reflecting plane is in this case located 1.5 m in front of the receiver point. The plane is oriented normally to the line that connects source and receiver. In this way, the direct sound (with delay \(t_d\) and level \(L_1\), regarded as the reference) and the reflection (with delay \(t_r\) and level \(L_2\)) originate from the same direction and have the same spectral distribution, ignoring the effect of air absorption and the finite size of the plane. The level difference between the two components is \(\Delta L = L_2 - L_1\) dB. The IR corresponding to the single reflection, excluding the direct sound, is noted as \(h_{ref}(t)\), processed with the LoRA toolbox, and stored,

\[
h_{ref}(t) \approx 10^{\Delta L/20}\delta(t - t_r),
\]

and the corresponding Fourier transformed version:

\[
20\log_{10}|H_{ref}(f)| = \Delta L \text{ dB}.
\]

The second step requires the presence of a human talker in the loudspeaker room. The talker is equipped with the headworn microphone, which requires a careful fixing to the talker’s head in order to preserve the relative position to the mouth throughout the operation. A measurement microphone B&K type 4192 is placed 1 m in front of the mouth. Next, the talker is asked to speak continuously during 30 s, staring at a reference sign so that the mouth is aligned with the measurement microphone. Both signals from the measurement microphone \(x_f(t)\) and the headworn microphone \(x_n(t)\) are recorded simultaneously. The goal of the calibration procedure is to obtain an ideal equalizer filter \(\tilde{h}_{EQ}(t)\) that applied to \(x_n(t)\) and reproduced through the system (with a gain symbolized \(G_{pb}\), where "pb" stands for "playback") produces \(x_f(t)\) at the center position,

\[
G_{pb}x_n(t) * \tilde{h}_{EQ}(t) = x_f(t),
\]

or in the frequency domain,

\[
G_{pb}X_n(f)\tilde{H}_{EQ}(f) = X_f(f),
\]

from which the ideal filter results,

\[
\tilde{H}_{EQ}(f) = \frac{X_f(f)}{G_{pb}X_n(f)}.
\]

The gain of the system \(G_{pb}\) is still unknown and requires another measurement. The room should be empty and the measurement microphone has to be moved to the center of the laboratory room, so that its position corresponds to the point between the two ears when a talker would be present. The previously recorded signal from the headworn microphone \(x_n(t)\) is routed to the input of the convolver, which is loaded with the single reflection, \(h_{ref}(t)\). The output of the convolver is sent to the amplifiers and reproduced through the loudspeakers.

At the same time, the measurement microphone records the resulting signal, \(\tilde{x}_{ref}(t)\),

\[
\tilde{x}_{ref}(t) = G_{pb}x_n(t) * h_{ref}(t),
\]

and the corresponding Fourier transform:

\[
\tilde{X}_{ref}(f) = G_{pb}X_n(f)H_{ref}(f).
\]

From the previous signals, it is possible to calculate the filter \(\tilde{H}_{EQ}\):

\[
\tilde{H}_{EQ}(f) = \frac{X_f(f)H_{ref}(f)}{X_{ref}(f)}.
\]

Making use of eq. (3) and using logarithms:

\[
20\log_{10}|\tilde{H}_{EQ}(f)| = 20\log_{10}|X_f(f)| - 20\log_{10}|\tilde{X}_{ref}(f)| + \Delta L.
\]

The fourth step is a practical implementation of eq. (10). It uses the signals corresponding to the on-axis direct sound, \(x_f(t)\), and the reflection, \(\tilde{x}_{ref}(t)\), as inputs. The signals \(x_f(t)\) and \(\tilde{x}_{ref}(t)\) are processed with a spectral analyzer that calculates the energy level of the signals in
FIG. 3. Steps involved in the calibration process. Step 1: calculation, by means of simulation, of a reference impulse response consisting of the direct sound and a single reflection. Step 2: Measurement of the on-axis speech signal at 1 m. Step 3: Playback, processing, and recording of the reflection. Step 4: Comparison of the direct sound and the reflection to obtain the personalized equalizer filter.

one-third octave frequency bands between 31.5 Hz and 16 kHz. The level difference between the two components is calculated and the target level difference $\Delta L$ is added. The result is the magnitude frequency response (in one-third octave bands) of the equalizer filter. The magnitude frequency response at frequencies other than the standardized one-third octave center frequencies are obtained by interpolation. The response is band-pass filtered to eliminate frequencies lower than 50 Hz, which are not likely to have been produced by the human voice, and frequencies higher than 10 kHz, to prevent unstable feedback in the system. The resulting filter $h_{EQ}(t)$ (slightly different from the ideal $\tilde{h}_{EQ}(t)$) is a 2048-tap FIR filter obtained by minimum phase reconstruction of the magnitude frequency response described in the previous steps.

As an example, Fig. 4 shows the magnitude frequency response of the equalizer filter $h_{EQ}(t)$ calculated for the same talker with slightly different microphone positions. As can be seen, these filters are fairly consistent, with a standard deviation of about 1.8 dB (averaged across frequency).

D. Feedback

The presence of the acquisition microphone and the loudspeakers in the same room generates a closed loop which introduces some feedback (unstable or not) in the system. Inspired by the method of Rokutanda et al.\textsuperscript{28}, the feedback is derived from an IR measurement $h_{FB}(t)$ at the headworn microphone using a Head And Torso Simulator B&K type 4128 (HATS) while the auralization system is running an arbitrary room simulation (see the complete system in Fig.5). The mouth-loudspeaker of the HATS is driven with an amplified pseudo-random noise signal (MLS). By calculating the cross-correlation of this signal and the signal at the headworn microphone, $h_{FB}(t)$ is obtained. It also contains the effect of the mouth radiator. The early part of $h_{FB}(t)$, in this case, contains the direct sound plus some reflections from the loudspeaker room and the torso, and the rest of $h_{FB}(t)$ is the feedback.
activity of the human talker matches the simulated directivity pattern, which has been chosen with this orientation. When the human talker turns around, the directivity pattern which is simulated by the system is still facing to the front. There is a mismatch between the actual and the simulated directivity pattern which emphasizes reflections from some directions and attenuates reflections from other directions. As a result, the simulated sound field is wrong. However, slight movements of the head do not give rise to a serious error. A measure of the error ϵ produced by head rotations in azimuth ϕ0 and elevation θ0, could be quantified by the following formula:

$$\epsilon(\phi_0, \theta_0) = 1 - \frac{\iint_{4\pi} D(\phi, \theta) D(\phi - \phi_0, \theta - \theta_0) d\Omega}{\iint_{4\pi} D^2(\phi, \theta) d\Omega}, \quad (11)$$

where $D$ is the linear directivity pattern of the simulated human head (assumed to be equal to an average talker long term speech directivity), $\phi$ and $\theta$ are the spherical coordinates (see Fig. 8), $\phi = 0, \theta = 0$ is the design orientation of the talker, and $\Omega$ indicates the solid angle. The head rotations in the radial direction are ignored in this analysis. Figure 9 shows the error graphically on a logarithmic scale: 10log(1 − ϵ). As expected, the accuracy decreases with frequency, as the voice becomes more directive. It is worthwhile to point out that the error produced by azimuthal rotations is lower than the error that would be produced by the same rotations in elevation. Azimuthal head rotations are more likely to occur than elevational ones. The contour lines at −3 dB show that azimuthal head rotations in the range $-30^\circ \leq \phi_0 \leq 30^\circ$ do not introduce severe inaccuracy of the simulated sound field. However, this error could be minimized, and the accuracy drastically improved by introducing a head tracking system that used the information about the head orientation to dynamically update the multichannel impulse response.

### E. Misalignment error

The system for the real-time auralization of one’s own voice assumes a speaker at the exact center of the loudspeaker array facing to the front. In this case, the directivity of the human talker matches the simulated directivity pattern, which has been built-up on a number of assumptions which are worthwhile summarizing:

#### F. Summary of assumptions

The real-time auralization system for the own voice has been built-up on a number of assumptions which are worthwhile summarizing:

---

**FIG. 5.** Setup used to measure the impulse response $h_{fbx}(t)$ between the mouth and the headworn microphone, from which the feedback in the system is evaluated.

**FIG. 6.** Feedback impulse response $h_{fbx}(t)$

**FIG. 7.** Feedback-to-direct as a function of the frequency for different gains of the simulated room impulse responses
IV. APPLICATIONS

The first experiences trying the system have been very positive, in the sense that it generates a convincing impression of being in environments different from the reproduction room, matching the expectations that talkers have about the simulated environments.

The system for real-time auralization of one’s own voice finds one of its main applications in psychophysics or cross-modality research. It is possible to investigate how people perceive environments by using exclusively aural cues produced with their own voices, study the subjective effects of the acoustic environment on voice production, or study the preference of theater actors in different acoustical settings. The system described in this article is being used at the time of publication in a research project where the relation between classroom acoustical conditions and the vocal behavior of a teacher is investigated.

The system could easily be adapted for use with music instruments. In this case, it would be necessary to make the computer acoustic simulation with the directivity pattern of the desired musical instrument, and perform the calibration exactly described in this paper, but replacing the headworn microphone with a microphone to pick the sound from the instrument. However, this microphone needs to be mounted on the instrument to reject feedback and avoid the variation of the acoustic path between source and acquisition microphone during operation.

Furthermore, this system could be used as a part of larger virtual reality systems, in order to achieve a more immersive experience. Applying some of the techniques here described and simplifying the reproduction method, this kind of system might also find place in digital entertainment.

Acknowledgments

Thanks to Sylvain Favrot and Jörg Buchholz for the development of the hardware arrangement and the LoRA toolbox, which allowed us to obtain the present system. Thanks to Anders Christian Gade for his valuable ideas regarding calibration and quality of the system. This research is financed by the Swedish organization AFA Försäkring.

12. C. Yuen, P. Calamia, and N. Xiang, “Investigation of voice stage support: A subjective preference test using an aural-
FIG. 9. Accuracy of the simulation $10 \log(1 - \epsilon)$ as a function of the azimuth ($\phi_0$) and elevation ($\theta_0$) head rotations for the different frequency bands. The 0 dB value corresponds to the perfect alignment of the talker to the simulated orientation. The -3 dB contour lines are indicated in solid.

\begin{align*}
21 & C. L. Christensen, Odeon room acoustics program, version 9.1, user manual, industrial, auditorium and combined editions (January 2008).
Speakers’ comfort and voice level variation in classrooms: laboratory research

David Pelegrín-García and Jonas Brunskog

Acoustic Technology, Department of Electrical Engineering, Technical University of Denmark, Kongens Lyngby DK-2800, Denmark

(Dated: September 12, 2011)

Teachers adjust their voice levels under different classroom acoustics conditions, even in the absence of background noise. Laboratory experiments have been conducted in order to understand further this relationship and to determine optimum room acoustic conditions for speaking. Under simulated acoustic environments, talkers do modify their voice levels linearly with the measure voice support, and the slope of this relationship is referred to as room effect. The magnitude of the room effect depends highly on the instruction used and on the individuals. Group-wise, the average room effect ranges from -0.93 dB/dB, with free speech, to -0.1 dB/dB with other less demanding communication tasks as reading and talking at short distances. The room effect for some individuals can be as strong as -1.7 dB/db. A questionnaire investigation showed that the preferred acoustic conditions for talking in classrooms, in the absence of background noise, are indicated by reverberation times around 0.5 s, measured from an impulse response between the mouth and the ears of a talker. Teachers with self-reported voice problems prefer higher reverberation times and more supportive rooms to speak in than their healthy colleagues.

PACS numbers: 43.55.Hy, 43.55.Mn

I. INTRODUCTION

The adjustment of vocal intensity (or voice level, indicated by the on-axis, free-field at 1 m SPL, or the sound power level) is regulated by a number of sensory inputs, including auditory feedback. In a review paper, Lane and Tranel point out a number of elements that contribute to alter auditory feedback, and as consequence, the voice levels. These elements are background noise, altered sidetone (amplified playback of one’s own voice), hearing loss, and room acoustics. The modification of voice levels as a result of an altered auditory feedback is generally called the Lombard effect. Lane and Tranel cited two works that reported observations of talkers speaking louder in acoustically “dead” rooms than in acoustically “live” rooms, arguing that this effect is a consequence of a psychological public loop used by talkers to adjust their voice level to the requirements of a given communication situation.

In recent investigations, the authors reported talkers’ voice level variations in rooms of -13.5 dB per each dB of change in the objective measure room gain $G_{RC}$. The room gain $G$ is a measure of the degree of amplification offered by the room to the voice of a talker at his ears, compared to anechoic conditions. Alternatively, the reported voice level variations were of -0.78 dB per each dB of change in the voice support $ST_V$ (named in analogy to Gade’s objective support, commonly used in the design of stages in concert halls). The voice support is a measure of the strength of the reflections of a talker’s voice at his ears relative to the strength of the direct sound.

The slope of the linear relationship between room gain or voice support and voice level is referred to as room effect in this paper. In the previous example, the room effect was -13.5 dB/dB$_G$ (dB of voice level over dB of room gain, denoted by the subscript $G$) or -0.78 dB/dB$_S$ (dB of voice level over dB of voice support, denoted by the subscript $S$). The voice support and the room gain are linked to each other through

$$ST_V = 10 \log(10^{G_{RC}/10} - 1).$$

However, these quantities are strongly dependent on the physical room volume. In rooms of similar proportions, the volume and the average communication distance between a talker and a group of listeners are correlated (i.e., in small rooms, listeners tend to be closer to the talker than in larger rooms). Talkers increase their voice level linearly with the logarithm of the distance to the listener. In a later study, Pelegrin-Garcia et al. quantified the room effect on talkers addressing listeners at different distances. In the original study, only the room effect at 6 m was reported. Figure 1 shows the linear regressions for the voice levels as a function of the room gain for different talker-to-listener distances (as a replot of the data of Fig. 4 in Pelegrin-Garcia et al.10). The measured room effect was -1.6 dB/dB$_G$ for a communication distance of 1.5 m, -3.6 dB/dB$_G$ for 3 m, -3.6 dB/dB$_G$ for 6 m and -3.7 dB/dB$_G$ for a distance of 12 m. Not only the absolute voice level, but also the absolute value of the room effect increased with communication distance. Nevertheless, only four acoustic environments—including an anechoic chamber and a reverberation room, thus non-representative of real-world conditions—were tested and the communication task (describing a map) was different from that used by Brunskog et al. who instructed talkers to give an oral presentation. The much higher value of room effect found by Brunskog et al. (-13.5 dB/dB$_G$)...

---

a) Electronic address: dpg@elektro.dtu.dk
included the effect of rising the voice level to address listeners at longer distances in the largest rooms (which had the lowest room gain values), corresponding therefore to an ecological location of the listeners.

Pelegrin-Garcia et al.\textsuperscript{11} determined the variations in voice level that would cause the loudness level of a talker’s own voice (i.e., the autophonic level) to be constant under different room acoustic conditions, characterized by the room gain or the voice support. The magnitude of the room effect for a talker addressing a listener at 1.5 m (-1.6 dB/dB\textsubscript{G}) is comparable to the variations in voice level that keep the autophonic level constant (-1.5 dB/dB\textsubscript{G} for a range of room gain between 0 and 0.8 dB), indicating that, in situations of low vocal demands, talkers just adjust their voice level to hear themselves equally loud.

Determining the magnitude of the room effect in more representative acoustic conditions would give a better understanding of the interaction between a talker and the physical environment, and would be a useful guideline to assess the consequences of different acoustics designs in rooms for speech. This is of special interest regarding teachers, who are one of the work forces with higher voice demands, who suffer from voice problems much more often than the rest of the population,\textsuperscript{12,13,14} and who consider bad classroom acoustics as a potential hazard.\textsuperscript{12} The present study aims at determining the room effect under simulated acoustic conditions, without the influence of visual cues. The experiments reported here have similarities to the ones conducted by Ueno et al.\textsuperscript{15} to analyze objective changes in musicians’ performance under different simulated acoustic conditions.

The present study has a secondary aim: to find out the preferred acoustic conditions for speaking in classrooms by means of questionnaires. In a similar research, Shearer and Torres\textsuperscript{16} observed that talkers have a preference for acoustic settings that naturally amplify their voices but dislike too long reverberation times because of the loss in clarity.

\section*{II. EXPERIMENTAL METHOD}

Four laboratory experiments in connection with the research project were carried out. The primary goal in all of them was to determine the relationship between objective acoustic parameters and the voice level adjustment. Secondary goals were to determine optimal acoustic conditions for speech. Two of the experiments (PRE1 and PRE2) were considered pilot studies, another one (A) aimed at analyzing differences in performance due to teaching experience. Hence, there were two groups: teachers (A1) and students (A2). The last experiment (B) aimed at detecting differences between a group of teachers with healthy voices (B1) and a group of teachers with self-reported voice problems (B2). A summary of the experiments, with the corresponding subject groups, number of subjects and conditions, methods, and setups used, is shown in Table I.

\subsection*{A. Setup}

As shown in Table I, there were two different setups in the experiments: a loudspeaker-based system and an earphone-based system. Both systems generated simulated acoustic sound fields with the voice of a talker in real time.

The first system was especially designed for the experiments PRE1, PRE2, and A. It consisted of 29 loudspeakers placed on the surface of an imaginary sphere (with a radius of 2 m) around a subject in a highly damped room. The speech signal from a talker (subject) in the center was picked with a headworn microphone, convolved in real time with a room impulse response (RIR)—which defined the acoustic condition—and recorded for analysis.

A block diagram of the system is shown in Fig. 2. In a pre-processing stage, computer room models corresponding to the desired acoustic conditions were built with a CAD program. These models were imported into an acoustical simulation software [Odeon (Odeon A/S; Koge, Denmark)], which computed the airborne acoustic path from the mouth to the ears of a talker. The output of the acoustic simulation software was a RIR containing information about the arrival time, level, and direction of sound reflections. The direct sound was excluded from the RIR because it was always present during the experiments (i.e., a talker can hear himself without the presence of a loudspeaker auralization system). The RIR from the acoustic simulation software was processed with the so-called LoRA (Loudspeaker Room Auralization) toolbox\textsuperscript{17} in order to obtain an appropriate format of the RIR consisting of 29 impulse responses (one for each loudspeaker). With the LoRA toolbox, the early reflections in the RIR were initially encoded with fourth order Ambisonics, whereas the envelope of the vector intensity histograms from the late reflections in the RIR was used to generate Gaussian noise encoded with first order Ambisonics. Afterwards, the LoRA toolbox decoded the Ambisonics format RIRs into 29 impulse responses (in WAV-format) suitable for the actual loud-
TABLE I. Summary of the experiments, identifying the group of subjects and its number, the number of conditions, whether there was a questionnaire study, the kind of instruction used, and the technical setup.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Subjects</th>
<th># Subjects</th>
<th># Conditions</th>
<th>Questionnaire?</th>
<th>Instruction</th>
<th>Setup</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pilot PRE1</td>
<td>Mixed teachers-students</td>
<td>5</td>
<td>5</td>
<td>NO</td>
<td>Simulated lecture</td>
<td>Loudspeaker</td>
</tr>
<tr>
<td>Pilot PRE2</td>
<td>Teachers</td>
<td>5</td>
<td>10</td>
<td>NO</td>
<td>Simulated lecture</td>
<td>Loudspeaker</td>
</tr>
<tr>
<td>A A1 Teachers</td>
<td></td>
<td>13</td>
<td>10</td>
<td>YES</td>
<td>Reading a text</td>
<td>Loudspeaker</td>
</tr>
<tr>
<td>A A2 Students</td>
<td></td>
<td>13</td>
<td>10</td>
<td>YES</td>
<td>Reading a text</td>
<td>Loudspeaker</td>
</tr>
<tr>
<td>B B1 Teachers (healthy voice)</td>
<td></td>
<td>11</td>
<td>10</td>
<td>YES</td>
<td>Describing a map</td>
<td>Earphones</td>
</tr>
<tr>
<td>B B2 Teachers (non-healthy voice)</td>
<td></td>
<td>10</td>
<td>10</td>
<td>YES</td>
<td>Describing a map</td>
<td>Earphones</td>
</tr>
</tbody>
</table>

FIG. 2. Loudspeaker based setup for experiments PRE1, PRE2, and A, which added the acoustic effect of a room to the voice of a talker by means of convolution in real time.

The real-time part of the system consisted of processing, acquisition, and reproduction. The acquisition of the voice from the talker was done with a headworn microphone DPA (DPA Microphones A/S; Allerød, Denmark) model 4066 located at 6 cm from the edge of the lips in the line between the mouth and the right ear. The signal was digitized and processed with an equalizer (EQ) filter, which adjusted the spectrum of the voice to be identical to the free-field, on-axis speech signal. The acoustic effect of the room was obtained by convolution of the voice of a talker with the RIR using the open-source convolution software jconvolver. Finally, the resulting signals were amplified and reproduced through the 29 loudspeakers. The acquisition and reproduction took place in a highly damped room to minimize the effect of natural sound reflections.

The second system was the same as used in previous research and had the same effect as the loudspeaker-based system on adding to the voice of the talker an auditory sensation of being in a room. Instead of using loudspeakers to add the simulated sound reflections to the voice of a talker, it used earphones, specially designed not to block the airborne direct sound. The generation of the acoustic conditions was similar to the process described before: a computer room model was inserted into an acoustic simulation software, which calculated the binaural RIR from the mouth to the ears, from which the direct sound was excluded by cropping.

B. Conditions

The different room acoustic conditions were defined by the RIRs loaded into the convolution software. The objective acoustic parameters of voice support ($ST_v$), reverberation time at the ears ($T_{30,ears}$, defined later on in this section), and reverberation time ($T_{20}$) corresponding to the different conditions are summarized in Table II for the different experiments. These parameters were derived from objective IR measurements in the laboratory facility, while an acoustic condition was active, using a dummy head B&K (Brüel & Kjær Sound & Vibration Measurement A/S; Nærum, Denmark) Head And Torso Simulator (HATS) type 4128 at the position of the talker, with a loudspeaker at its mouth and microphones at the eardrums. The $T_{20}$ was calculated after removing the first 5 ms of the IR, in order to avoid the strong influence of the direct sound. The $T_{30,ears}$ was calculated as the decay from -5 to -35 dB in the backwards integrated energy curve from the IR measured between the mouth and the ears of the dummy head, as shown in Fig. 3. The conditions under which $T_{30,ears}$ is determined, with a strong influence of the direct sound, result in a dependence of both the decay time of the sound in the room and the direct-to-reverberant sound level difference. All the parameters were averaged for the left and the right ear. The conditions were presented in random order for each subject. (NOTE: The reverberation time at the ears $T_{30,ears}$ is not intended to be a new measure of reverberation time, but a specification of the conditions used to determine the reverberation time. A prediction model showing the dependence of $T_{30,ears}$ with the “standard” reverberation time and the room volume is described in the Appendix.)

The conditions in experiments PRE1, A, and B were obtained by acoustic simulation of classrooms with different geometries and absorption layouts. The five conditions in the pilot experiment PRE1 were a subset of the ten conditions in experiment A. The goal of the pilot experiment PRE2 was to study the response to artificial RIRs. For this reason, all conditions were generated by applying different overall gains to a single RIR, and there-
TABLE II. Objective parameters of voice support ($ST_v$), reverberation time at the ears ($T_{30,ears}$), and reverberation time ($T_{20}$) characterizing the acoustic conditions of the different experiments

<table>
<thead>
<tr>
<th>Condition</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ST$_v$, dB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PRE1</td>
<td>-12.2</td>
<td>-15.1</td>
<td>-14.6</td>
<td>-14.6</td>
<td>-15.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$T_{30,ears}$, s</td>
<td>0.73</td>
<td>0.20</td>
<td>0.41</td>
<td>0.88</td>
<td>0.05</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$T_{20}$, s</td>
<td>0.96</td>
<td>0.36</td>
<td>0.67</td>
<td>1.51</td>
<td>0.08</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PRE2</td>
<td>-15.8</td>
<td>-15.8</td>
<td>-15.6</td>
<td>-15.3</td>
<td>-14.8</td>
<td>-14.1</td>
<td>-13.3</td>
<td>-12.5</td>
<td>-11.7</td>
<td>-10.9</td>
</tr>
<tr>
<td></td>
<td>$T_{30,ears}$, s</td>
<td>0.05</td>
<td>0.11</td>
<td>0.32</td>
<td>0.47</td>
<td>0.60</td>
<td>0.69</td>
<td>0.77</td>
<td>0.83</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>$T_{20}$, s</td>
<td>0.07</td>
<td>0.52</td>
<td>0.84</td>
<td>0.99</td>
<td>1.15</td>
<td>1.09</td>
<td>1.10</td>
<td>1.09</td>
<td>1.09</td>
</tr>
<tr>
<td>A</td>
<td>-12.2</td>
<td>-14.0</td>
<td>-14.6</td>
<td>-15.1</td>
<td>-14.8</td>
<td>-15.6</td>
<td>-12.8</td>
<td>-14.6</td>
<td>-14.7</td>
<td>-15.9</td>
</tr>
<tr>
<td></td>
<td>$T_{30,ears}$, s</td>
<td>0.73</td>
<td>0.65</td>
<td>0.88</td>
<td>0.20</td>
<td>0.34</td>
<td>0.23</td>
<td>0.45</td>
<td>0.41</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td>$T_{20}$, s</td>
<td>0.96</td>
<td>0.98</td>
<td>1.51</td>
<td>0.36</td>
<td>0.51</td>
<td>0.52</td>
<td>0.68</td>
<td>0.67</td>
<td>1.16</td>
</tr>
<tr>
<td>B</td>
<td>-6.1</td>
<td>-10.5</td>
<td>-11.3</td>
<td>-16.8</td>
<td>-14.7</td>
<td>-19.3</td>
<td>-7.2</td>
<td>-11.1</td>
<td>-11.6</td>
<td>-23.2</td>
</tr>
<tr>
<td></td>
<td>$T_{30,ears}$, s</td>
<td>0.68</td>
<td>0.62</td>
<td>0.85</td>
<td>0.17</td>
<td>0.32</td>
<td>0.22</td>
<td>0.43</td>
<td>0.41</td>
<td>0.79</td>
</tr>
<tr>
<td></td>
<td>$T_{20}$, s</td>
<td>0.77</td>
<td>0.78</td>
<td>1.02</td>
<td>0.29</td>
<td>0.41</td>
<td>0.42</td>
<td>0.53</td>
<td>0.55</td>
<td>0.92</td>
</tr>
</tbody>
</table>

C. Subjects

There were different number of subjects in each experiment (see Table I). In the pilot experiment PRE1, there were five subjects with ages 23 to 35 yr. They were three fellow students and two teachers from the research group in acoustics who had good hearing (no hearing loss greater than 25 dB HL below 4 kHz) and vocal health (no self-reported voice problems) at the time of the experiments.

In the pilot PRE2, the subjects were five male teachers with ages 29 to 65 yr, from secondary school to university levels. The subjects had good hearing and voice health, with the same criteria as in PRE1.

In the experiment A, there were 13 teachers (group A1: four females, nine males) of secondary school, high school, and university, with ages 30 to 67 yr. There were also 13 students (group A2: 12 males, one female) with ages 24 to 28 yr. None of the subjects had self-reported voice problems or hearing loss greater than 25 dB HL.

In the experiment B, there were a total of 21 teachers divided into two groups according to their vocal health: a group of 11 teachers with healthy voices (group B1: two males, nine females) and ages 26 to 63 yr and a group of 10 teachers with self-reported voice problems (group B2: one male, nine females) and ages 29 to 62 yr. These teachers were selected from a questionnaire study and participated in previous clinical and field research.

D. Instruction

In the pilot experiments PRE1 and PRE2, the subjects were instructed to give a lecture of 3 minutes in their mother tongue to an imaginary group of 30 students under each condition. The subjects were instructed about this beforehand, and they could repeat the lecture on each condition.

In experiment A, the subjects were instructed to read a text (Goldilocks’ passage) in English—although it was
not the mother tongue of most of them—during 2.5 minutes at each condition, addressing a listener located at a distance of 2 m, simulated with a dummy head to provide a visual reference distance cue. At each condition, the subjects had to start reading the text from the beginning.

In experiment B, the subjects were given a map which contained a number of labelled items and a path connecting two points (the maps have been used in previous research\cite{10,22}). They were instructed to describe the route between the starting point and the finish point, indicating the items along the path (e.g. “go to the west until you find the harbor”), while trying to enable eye-contact with the experimenter, seated at 3 m distance in front of them. A different map (out of ten) was used at each condition. The order of the maps was randomized differently for each subject. All teachers performed the task in Swedish, their mother tongue. However, the experimenter did not understand this language.

In all experiments, the start and the end of a condition was indicated by means of acoustic signals.

E. Questionnaires

In experiments A and B, the subjects had to rate a set of questions or statements regarding the experience of talking under a certain acoustic condition, by making a vertical tick in a continuous horizontal line of 100 mm length, after every experimental condition. These statements were the following:

1. I would feel exhausted if I were talking in this classroom for a whole lesson
2. The classroom is good to speak in
3. The classroom enhances and supports my speech
4. I must raise my voice in order to be heard in the classroom
5. The sound system makes my voice sound unnatural
6. I noticed echo phenomena in the classroom
7. Rate the degree of reverberance that you perceived in the classroom
8. Rate how you perceive your voice now

In questions 1 to 6, the extremes of the lines were totally disagree (left) and strongly agree (right). In question 7, the extremes were very low (left) and very high (right). Question 8 had extremes no voice problems (left) and extremely severe problems (right). This last question had the aim of detecting anomalous performance in certain conditions.

F. Post-processing of the speech signals

The voice recordings were processed to determine the phonated or voiced segments in speech with the average magnitude difference function method\cite{23} implemented in MATLAB. The length of the segments was 50 ms. In these segments, the fundamental frequency F0 was determined. The segments with too high or too low F0 (due to erroneous detection in the algorithm) were considered unvoiced in the analysis. Next, the equivalent Sound Pressure Level (SPL) of the phonated segments at the position of the headworn microphone was calculated. In separated measurements in an anechoic chamber, the SPL of atalker was measured simultaneously with the headworn microphone and a free-field microphone at 1 m in front of the talker. The difference between the SPL at the two positions was determined, and this quantity was used to report all the SPL values in the investigation as on-axis, free field SPL_{ff,1m} at 1 m in front of the talker, simply referred to as voice level. Other parameters, like mean F0, standard deviation of F0, or relative phonation time, were calculated but led to non-significant results, and are therefore not reported here.

G. Statistical Analysis

The statistical analysis was carried out in the open-source statistics software R.\cite{24}

1. Voice level

In all cases, the focus of the experiments was to characterize the dependence of the voice level with the room acoustic conditions. The room acoustic conditions were described by only one variable at a time (objective acoustic parameters STV, T20, or T30_{ears}). An initial correlation analysis was performed, comparing the voice level with each of the objective acoustic parameters. The strongest correlations were found between the voice level and STV in most of the experiments, therefore STV was used as the main predictor for the voice level.

In order to evaluate the fixed effects, an ANCOVA (ANalysis of COVAriance)\cite{25} model was used, with STV as the only explanatory variable, and subject as a factor. The interaction between the subject and the explanatory variable was allowed. The significance of variables and interactions is shown by means of an ANOVA table (Analysis OF VAriance) in the results section.

In a next step, mixed-effects model with random slopes were fitted to the data. However, assumptions of normality of the random effects were not fulfilled. Therefore, no generalization of the effects observed in this study to a greater population is aimed for.

The voice level measurements of each subject were used together to fit a line with a certain slope. The slopes from all subjects in any of the experiments were a sample of the ideal slope distribution expected from that particular experiment. The comparisons between experiments were done with an ANOVA on the slope samples derived from the individuals and post-hoc Tukey HSD tests.
2. Subjective data

In order to reduce the differences across subjects due to criteria and scale, z-scores\(^{26}\) were obtained from the answers to the questions. I.e., for a specific subject and question, the z-score was calculated as the difference from the average value and divided by the sample standard deviation.

The answers to different questions were very interrelated among them. A principal component analysis (PCA)\(^{27}\) was used to reduce the redundancy in the set of answers to the questions. Prior to this, a linear relation between responses was observed (if there was a relation at all). Initial PCAs revealed question 8 not to load strongly any principal component (PC) with eigenvalues higher than 1. Since question 8 was only weakly correlated with other questions, it was excluded from the PCAs.

The PCAs revealed two PCs with eigenvalues higher than 1 in the analysis of all experiments. After varimax rotation, performed with the function `factanal()` in R, the loadings of the different questions on the two PCs with eigenvalues higher than one are shown in Table III. One of the PCs (PC1 in experimental groups A1, B1, and B2; PC2 in group A2) was related mainly to the questions 1 to 4, whereas the other PC was related to questions 5 to 7. The first one can be interpreted as the overall quality of the room, whereas the second one can be linked to the reverberance.

It is important to remark that the score of the first PC decreased with the perception of exhaustiveness of speaking in a classroom during a lesson, increased with the perception of the classroom as being good to speak in, increased with the perceived support and enhancement, and decreased with the sensation of having to increase the voice level.

The scores of the two PCs were used in regression analyses with objective acoustic parameters as explanatory variables. After initial inspection of the relation between the PCs and the objective parameters, it was observed that \(T_{30,ears}\) defined the main trends in the PCs (instead of \(ST_V\)) and it was the only variable used.

III. RESULTS

A. Voice level

Statistical models for the voice level in the different experiments, with \(ST_V\) as a linear predictor, subject as a factor, and an interaction between subject and \(ST_V\), were assessed by means of ANOVAs (Table IV). In all cases, the effect of the factor subject was highly significant, pointing out the importance of the individual differences in voice level. This is better observed in the sum of squares (see Table IV), which is about an order of magnitude higher for subject compared to other sources of variability. In all cases but for group A2 (\(F_{1,104} = 1.2; p = 0.274\)), the effect of \(ST_V\) on voice level (or room effect) was highly significant. The interaction between subject and \(ST_V\) was highly significant in group A1 (\(F_{12,104} = 3.86; p < 0.001\)), significant in pilot experiment PRE2 (\(F_{4,40} = 3.20; p = 0.023\), nearly non-significant in groups A2 (\(F_{12,103} = 1.8; p = 0.054\)) and B2 (\(F_{5,79} = 1.8; p = 0.09\), and non-significant at all for pilot experiment PRE1 (\(F_{4,15} = 1.0; p = 0.43\) and group B1 (\(F_{10,87} = 0.86; p = 0.57\)).

The estimates of the average intercepts and slopes (room effect) in the linear models, together with their standard error, the residual standard error, and the coefficient of determination \(R^2\), are shown in Table V. The coefficients of determination \(R^2\) were very high (higher than 0.87 in all cases), which is explained by the high differences in voice level across subjects and the large amount of variability explained by taking into account the factor subject in the analysis. All the average room effects were negative, indicating a tendency of the talkers to lower their voice levels as the voice support in a room increased. The average room effect for the pilot experiments PRE1 and PRE2 was -0.89 and -0.96 dB/dBS, respectively. For experiment A, the average room effect was -0.35 dB/dBS in group A1 and -0.11 dB/dBS (non-significant effect) in group A2. For experiment B, the average slopes were -0.12 dB/dBS for group B1 and -0.07 dB/dBS for group B2.

The average relationship between voice level and \(ST_V\) is shown in Fig. 4 as straight solid lines. The confidence intervals, calculated from a simple linear regression model (without taking into account the effect of subject) are shown in the same figure as dashed curves. On the left plot, the confidence intervals are further away from the regression line due to the low number of subjects participating in the experiments.

The slopes in the relationship between voice level and \(ST_V\) (the room effect) were the most important quantities in the analysis. For each subject, a regression line was fit to the data, and the room effects from all individuals are summarized in Fig. 5 in the form of a histogram, identifying to which experiments the room effect belonged to. The shape of the histogram was non-symmetric, with a larger presence of values in the negative tail than in the positive tail. In general, it can be seen that the room effects from the pilot experiments were more negative than other experiments (also seen from the more negative average room effects in Table V). The room effects from group A2 were indifferently positive and negative, and there was a high concentration of room effects from experiment B around 0—though slightly negative. Because the room effects were not following a normal distribution, it would not have been appropriate to use random slopes mixed-effects model to characterize the data.

An ANOVA revealed significant differences (\(F_{5,51} = 7.43; p < 0.001\)) in room effect across experimental groups. A Tukey HSD post-hoc analysis was applied to determine which experimental pairs of groups produced the differences (Table VI). There were significant differences between the pilot tests PRE1, PRE2 and all the other experimental groups (A1, A2, B1, and B2), although the difference between PRE1 and A1 was only significant at the 10% level (\(p = 0.093\)). Differences among experimental groups A1, A2, B1, and B2 were
TABLE III. Loadings of the principal components (PC) of the questions 1 to 7 with eigenvalues larger than 1, after a varimax rotation.

<table>
<thead>
<tr>
<th>Question</th>
<th>PC1</th>
<th>PC2</th>
<th>PC3</th>
<th>PC4</th>
<th>PC5</th>
<th>PC6</th>
<th>PC7</th>
<th>Eigen-value</th>
<th>% variance explained</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>-0.80</td>
<td>0.05</td>
<td>0.86</td>
<td>-0.63</td>
<td>-0.12</td>
<td>-0.01</td>
<td>0.24</td>
<td>1.71</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A2</td>
<td>0.29</td>
<td>-0.19</td>
<td>0.08</td>
<td>-0.23</td>
<td>0.66</td>
<td>0.81</td>
<td>0.81</td>
<td>1.46</td>
<td>30</td>
</tr>
<tr>
<td>B1</td>
<td>-0.87</td>
<td>0.20</td>
<td>-0.13</td>
<td>-0.03</td>
<td>0.58</td>
<td>0.79</td>
<td>0.79</td>
<td>1.36</td>
<td>26</td>
</tr>
<tr>
<td>B2</td>
<td>-0.70</td>
<td>0.20</td>
<td>0.05</td>
<td>-0.23</td>
<td>0.59</td>
<td>0.84</td>
<td>0.84</td>
<td>1.49</td>
<td>32</td>
</tr>
</tbody>
</table>

TABLE IV. Analysis of variance table for the voice level of the different experiments and subject groups, according to an ANCOVA model with STv as the explanatory variable, interacting with the subjects. Significance levels: *** ($p < 0.001$), ** ($0.001 < p < 0.01$), * ($0.01 < p < 0.05$), . ($0.05 < p < 0.1$), — not significant ($p > 0.1$)

<table>
<thead>
<tr>
<th></th>
<th>Degrees of freedom</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F value</th>
<th>p-value</th>
<th>Significance level</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRE1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>STv</td>
<td>1</td>
<td>31</td>
<td>30.5</td>
<td>40.1</td>
<td>1.30E-05</td>
<td>***</td>
</tr>
<tr>
<td>Subject</td>
<td>4</td>
<td>370</td>
<td>92.5</td>
<td>121.6</td>
<td>3.10E-11</td>
<td>***</td>
</tr>
<tr>
<td>STv*Subject</td>
<td>4</td>
<td>3</td>
<td>0.80</td>
<td>1.0</td>
<td>0.43</td>
<td>—</td>
</tr>
<tr>
<td>Residuals</td>
<td>15</td>
<td>11</td>
<td>0.80</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PRE2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>STv</td>
<td>1</td>
<td>84</td>
<td>84</td>
<td>49.0</td>
<td>1.90E-08</td>
<td>***</td>
</tr>
<tr>
<td>Subject</td>
<td>4</td>
<td>2065</td>
<td>516</td>
<td>301.90</td>
<td>&lt; 2E-16</td>
<td>***</td>
</tr>
<tr>
<td>STv*Subject</td>
<td>4</td>
<td>22</td>
<td>5</td>
<td>3.20</td>
<td>0.0230</td>
<td>*</td>
</tr>
<tr>
<td>Residuals</td>
<td>40</td>
<td>68</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>STv</td>
<td>1</td>
<td>19</td>
<td>19.4</td>
<td>25.99</td>
<td>1.50E-06</td>
<td>***</td>
</tr>
<tr>
<td>Subject</td>
<td>12</td>
<td>1580</td>
<td>131.7</td>
<td>176.23</td>
<td>&lt; 2E-16</td>
<td>***</td>
</tr>
<tr>
<td>STv*Subject</td>
<td>12</td>
<td>35</td>
<td>2.9</td>
<td>3.86</td>
<td>6.9E-05</td>
<td>***</td>
</tr>
<tr>
<td>Residuals</td>
<td>104</td>
<td>78</td>
<td>0.7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>STv</td>
<td>1</td>
<td>2</td>
<td>1.8</td>
<td>1.2</td>
<td>0.274</td>
<td>—</td>
</tr>
<tr>
<td>Subject</td>
<td>12</td>
<td>1852</td>
<td>154.4</td>
<td>101.2</td>
<td>&lt; 2E-16</td>
<td>***</td>
</tr>
<tr>
<td>STv*Subject</td>
<td>12</td>
<td>33</td>
<td>2.8</td>
<td>1.8</td>
<td>0.0540</td>
<td>.</td>
</tr>
<tr>
<td>Residuals</td>
<td>103</td>
<td>157</td>
<td>1.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>STv</td>
<td>1</td>
<td>39</td>
<td>39.0</td>
<td>34.21</td>
<td>8.50E-08</td>
<td>***</td>
</tr>
<tr>
<td>Subject</td>
<td>10</td>
<td>1127</td>
<td>112.7</td>
<td>98.96</td>
<td>&lt; 2E-16</td>
<td>***</td>
</tr>
<tr>
<td>STv*Subject</td>
<td>10</td>
<td>10</td>
<td>1</td>
<td>0.86</td>
<td>0.57</td>
<td>—</td>
</tr>
<tr>
<td>Residuals</td>
<td>87</td>
<td>99</td>
<td>1.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>STv</td>
<td>1</td>
<td>11</td>
<td>10.8</td>
<td>7.4</td>
<td>0.0079</td>
<td>**</td>
</tr>
<tr>
<td>Subject</td>
<td>9</td>
<td>753</td>
<td>83.7</td>
<td>57.5</td>
<td>&lt; 2E-16</td>
<td>***</td>
</tr>
<tr>
<td>STv*Subject</td>
<td>9</td>
<td>23</td>
<td>2.6</td>
<td>1.8</td>
<td>0.0901</td>
<td>.</td>
</tr>
<tr>
<td>Residuals</td>
<td>79</td>
<td>115</td>
<td>1.5</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

not significant.

B. Subjective data

The plots of the PC defining overall quality for speaking is shown in the top row of Fig. 6. The average value of the PC followed a non-linear relationship with $T_{30,ears}$ for experimental groups A1, A2, and B1, which was mod-
TABLE V. Parameter estimates of linear regression models relating voice level to $ST_v$ in each of the experiments, considering subject a blocking factor and interaction between subject and $ST_v$. The residual standard error and the coefficient of determination $R^2$ are also shown.

<table>
<thead>
<tr>
<th></th>
<th>Intercept [dB]</th>
<th>Slope [dB/dB$_g$]</th>
<th>Residual [dB]</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean Std. Error</td>
<td>Mean Std. Error</td>
<td>Std. Error</td>
<td></td>
</tr>
<tr>
<td>PRE1</td>
<td>46.2 2.63</td>
<td>-0.89 0.14</td>
<td>0.87 0.97</td>
<td></td>
</tr>
<tr>
<td>PRE2</td>
<td>44.0 4.14</td>
<td>-0.96 0.25</td>
<td>1.31 0.97</td>
<td></td>
</tr>
<tr>
<td>A1</td>
<td>54.8 1.96</td>
<td>-0.35 0.14</td>
<td>0.86 0.96</td>
<td></td>
</tr>
<tr>
<td>A2</td>
<td>60.0 1.87</td>
<td>-0.11 0.13</td>
<td>1.24 0.92</td>
<td></td>
</tr>
<tr>
<td>B1</td>
<td>58.1 0.99</td>
<td>-0.12 0.02</td>
<td>1.07 0.92</td>
<td></td>
</tr>
<tr>
<td>B2</td>
<td>60.2 1.00</td>
<td>-0.07 0.03</td>
<td>1.21 0.87</td>
<td></td>
</tr>
</tbody>
</table>

The results from the laboratory experiments have shown that there is a tendency of speakers to lower their voice level as a function of $ST_v$, i.e., there is a room effect on speakers’ voices, although it varies in magnitude in different experiments. The extent to which it varies is both individual and communication scenario-based. The proof that it is an individual attribute, is that only a few room effect values were measured at the lower tail of the histogram in Fig. 5.

The individual characteristic of the room effect might be due to the sensitivity of the speaker toward changing acoustic conditions, acquired through knowledge in acoustics or experience. No further investigation has been done to determine the causes for this effect, but it could be potentially developed on the rest of the individuals by training for adjusting their voices as a function of the room acoustics conditions.

A. Effect of experience

If teaching experience was a decisive factor contributing to the room effect, it would be expected that the teacher group A1 would have more negative room effect than the students group A2. The average room effect
for teachers in group A1 was -0.35 dB/dB, whereas for students in group A2 it was -0.11 dB/dB (see Table V). The difference in room effect between the two groups was non-significant (t-test: $t_{24} = -1.26; p = 0.59$, see Table VI). The lower mean value for the teacher group could be due to extreme individual differences: one of the individual room effects in the teacher group was 1.45 dB/dB, whereas one of the individual room effects in the student group was +0.75 dB/dB. A possible reason for the non-significant difference between groups is that reading a text does not point out differences between the two groups, because teachers do not base their teaching activity on reading texts. Another explanation could be due to all students were specialized in acoustics and many of them had heard about the aims of the research beforehand. Thus, experiment A cannot tell whether there are actual differences in room effect between teachers and students in a realistic teaching scenario.

B. Effect of voice health

The average room effect for teachers with healthy voices (group B1) was -0.12 dB/dB, whereas for teachers with self-reported non-healthy voices (group B2) it was -0.07 dB/dB. The difference between the two groups was non-significant ($t_{14.7} = -1.52; p = 0.15$). In a previous investigation, the same teachers were studied in a real teaching scenario. Although the results of that study were problematic from the methodological perspective, because most of the individuals talked in just one environment, the group-wise differences were statistically significant. It is possible that teachers perceived the laboratory scenario as non-demanding and did not stress their voices as if it was a real teaching situation. The low communication demands in the laboratory scenario were given by the absence of noise in the simulations and by the presence of one single listener at three meters distance who did not understand the language of the talkers.

On the other hand, subjective preference was slightly different for the two groups of teachers. As can be seen in Fig. 6, the PC related to the overall quality of the room for speaking followed a non-linear trend—as a function of the reverberation time at ears—in the voice-healthy group, whereas it followed a linear trend for the voice-unhealthy group. The non-linear trend was also observed for the other healthy groups (A1 and A2), indicating an optimum/most preferred reverberation time at the ears $T_{30, ears}$ of about 0.5 s. The teachers with non-healthy voices preferred classrooms with higher reverberation time at the ears, trading intelligibility for amplification of their voices. This is in good agreement with the results of the study in a real teaching scenario, where teachers with voice problems made a more efficient use of their voices, lowering them with increasing voice support.

C. Effect of instruction

The average room effect under the three different instructions was -0.93 dB/dB for free speech in the pilot tests, -0.23 dB/dB for reading a text aloud in experiment A, and -0.10 dB/dB describing a map in experiment B. The spread around these values is summarized in the boxplot of Fig. 7. The room effect values were significantly different between the free speech and reading a text aloud ($t_{18.8} = -4.2; p < 0.001$), between free
speech and describing a map \((t_{9.3} = -6.1; \ p < 0.001)\), but not between reading a text aloud and describing a map \((t_{26.8} = -1.39; \ p = 0.18)\). The variance of the measured room effects in the case of describing a map was significantly lower than in the case of reading a text \((F\text{-test}, F_{25.20} = 33.3; \ p < 0.001)\).

The fact that the room effect was significantly higher (in absolute terms) for free speech than for reading a text or describing a map seems related to the demands of the communication scenario. In the case of free speech, talkers had to address an imaginary group of 30 students. It is possible that the talkers imagined the group of students being located according to the perceived size of the room, and thus varied their voice levels to reach audiences at different distances. However, the imaginary location of the audience was a non-controlled variable. This cue could be controlled by including a visual reference in the experimental setup, e.g. by means of three-dimensional images or a virtual reality system. On the other hand, the room effect values of -0.89 dB/dB\(_S\) in the pilot experiment PRE1 and -0.96 dB/dB\(_S\) in PRE2 are not significantly different from the value of -0.78 dB/dB\(_S\) measured in real classrooms\(^6\) with an ecological distribution of listeners.

In the case of reading a text, addressing a one-person-audience at a distance of 2 m is not a challenge. In previous works\(^5,10,11\) the authors found that the voice level variation addressing a listener at 1.5 m under different room acoustics conditions was equivalent to keeping the autophonic level constant under the same conditions. For a range of \(ST_V\) between -16 and -12 dB (for experiment A, see Table II), Pelegrin-Garcia et al.\(^11\) predict a room effect of -0.08 dB/dB\(_S\) that keeps the autophonic level constant, comparable to the -0.11 dB/dB\(_S\) measured for the student group A2. The high variance of the reading task might have been caused by different attitudes of the subjects toward the task. For example, some teachers read the text as a story for small kids, changing their voice quality with the characters in the dialogues, whereas other teachers read the text through, totally dispassionate.

The instruction of describing a map was given in order to achieve a more realistic communication scenario. For the range of \(ST_V\) in experiment B (between -23.2 and -6.1 dB), the average slope that would keep the autophonic level constant is -0.1 dB/dB\(_S\), which is almost identical to the average slope in experiment B. Apparently, the talkers in experiment B just kept their autophonic level constant. There was no room effect beyond the Lombard effect, in opposition with the findings of Pelegrin-Garcia et al.\(^10\) in real rooms, who found higher room effects at a communication distance of 3 m. This observation questions the realism of the communication scenario in laboratory. Moreover, there was no premium on understanding what the talkers said, because the experimenter did not understand the language of the talkers, and because of the awareness of being in a laboratory.

**D. Subjective preference**

Two main PCs appeared after the analysis of the questionnaires. One of them was related to the sensation of reverberance, and was highly correlated in a linear fashion with the reverberation time measured between the mouth and the ears of a talker.

The other PC was related to the overall quality of the room for speaking in. For all subjects with healthy voices (experimental groups A1, A2, and B1), the relationship between this component and the reverberation time at the ears was non-linear, with a maximum point located at an abscissa of about 0.5 s (range 0.45 to 0.55 s). Lower reverberation times at ears are perceived as uncomfortable because they correspond to acoustically dry environments which do not support the voice. Higher reverberation times are not preferred because one of the consequences of reverberation is the decrease in speech intelligibility (note that the reverberation times referred to in this paper are measured from the mouth to the ears of a dummy head—which are strongly influenced by the direct sound—and represent a combined measure of reverberation time and direct-to-reverberant ratio. Standard measures of reverberation are related to the values presented here through the prediction model described in the Appendix). These results point in the same direction as found by Shearer and Torres,\(^16\) who also found a preference for reverberation times neither too short nor too long.

Some teachers rated poorer the acoustics conditions with the highest reverberation times, because they imagined that an audience of children in that condition would become very noisy. Although that aspect was not desired for the evaluation of the questionnaire, it is nevertheless a factor to take into account during the classroom acoustics design. In fact, preferred reverberation times at ears in real settings might be lower than measured in laboratory (0.5 s) due to the increase of activity noise levels with reverberation time in a multitalker situation,\(^28\) consequence of the Lombard effect.\(^2\)

Teachers with non-healthy voices apparently prefer higher reverberation times because their autophonic levels are increased. It is not known whether this opinion
would be held in the presence of activity noise from students, but it is likely that the increase in noise would reduce their preference scores.

V. CONCLUSIONS

Laboratory experiments have been conducted in order to determine whether talkers modify their voice levels as a function of the room acoustics conditions and to determine optimum room acoustics conditions for speaking.

Under simulated acoustics conditions, talkers do modify their voice levels linearly with the quantity voice support, and the slope of this relationship is referred to as **room effect**. The room effect depends highly on the instruction used and on individuals. With free speech, it had an average value of -0.93 dB/dB, whereas with other tasks it had average values between -0.35 dB/dB and -0.07 dB/dB. Typical values were around -0.1 dB/dB, which is approximately the value of the room effect that keeps the loudness of a talker’s voice constant.

The preferred acoustic condition for talking in a classroom, in the absence of background noise, is indicated by a reverberation time of around 0.5 s, measured from an impulse response between the mouth and the ears of a dummy head, evaluating the decay levels from -5 to -35 dB.

Teachers with voice problems perceive their environment differently than teachers without voice problems, preferring higher reverberation times and more supportive rooms to speak in.

**Acknowledgments**

This research has been funded by the Swedish organization AFA Försäkring as a part of the project “Speaker’s comfort and voice health in classrooms”. The valuable feedback from Torben Poulsen at the Technical University of Denmark (DTU) has contributed to improve the quality of the manuscript. Thanks to Jesper Hagedorn, former DTU student, for conducting the tests in experiment A2 and to Dr. Viveka Lyberg-Åhlander, from Lund University, for her help in selecting the teachers for experiment B. Thanks to Prof. Marianne Gullberg from the Humanities Lab at Lund University for putting at our disposal the experimental facilities in experiment B. The authors would like to express their gratitude to all the participants in the laboratory experiments.

**APPENDIX: A MODEL FOR REVERBERATION TIME AT THE EARS**

The suggested quantity of reverberation time at the ears $T_{30,ears}$ is derived from an impulse response characterizing the airborne acoustic path between the mouth and the ears $h(t)$. A model of $h(t)$ corresponding to an average of impulse responses at different positions in a room has three components: first, the direct sound propagating from the mouth to the ears. Second, a reflection from the floor, considering that the mouth and the ears of the speaker are at a height $d = 1.5$ m above the floor. Third, an exponential decay or reverberation tail. The discrete early reflections other than the floor reflection are assumed to vary in intensity and in time and direction of arrival with different positions and are considered part of the reverberation tail after averaging. It is assumed that the direct sound and the floor reflection are Dirac delta functions arriving at $t = 0$ and $t = t_0$, respectively. The reverberation process is assumed to start at $t = 0$. Furthermore, it is assumed that the energy of these components sums directly. The energy density time curve $E(t) = h^2(t)$ is illustrated in Fig. 8.

![Energy density time curve assumed for the prediction of $T_{30,ears}$, showing the main components in the airborne acoustic path between the mouth to the ears: the direct sound, the floor reflection, and the reverberation tail](image)

As in the prediction model for voice support proposed by Pelegrin-Garcia *et al.*, the different quantities are based on the physical effects of a head and torso simulator Bruel & Kjaer type 4128, also used in section II. The direct sound has an energy level

$$L_d = L_W + K,$$

where $L_W$ is the voice power level and $K$ is the measured difference between the voice power level and the sound pressure level at the ears in free field. As the choice of $L_W$ is rather arbitrary, $L_W = -K$ is chosen. Therefore the energy density of the direct sound $E_d(t)$ is

$$E_d(t) = \delta(t)$$

so that the energy is 1 and $L_d = 0$. The energy level from the floor reflection at the point of the ears $L_{FR}$ is

$$L_{FR} = L_W + 10 \log \left( \frac{Q^*}{4\pi(2d)^2} \right) + \Delta L_{HRFF},$$

where the middle term is the propagation factor, assuming no sound absorption at the floor, $Q^*$ is the directivity factor of speech in the downward direction (derived from Chu and Warnock), and $\Delta L_{HRFF}$ is the HRTF correction factor (for using ears-like receiver instead of an omnidirectional microphone). As an approximation, $\Delta L_{HRFF}$ is assumed to correspond to a diffuse-field situation. The energy density of the floor reflection $E_{FR}(t)$ is

$$E_{FR}(t) = A\delta(t - t_0),$$
with parameters

\[ 10 \log A = 10 \log \left( \frac{Q^*}{4\pi(2d)^2} \right) + \Delta L_{\text{HRTF}} - K \]  

(A.5)

and

\[ t_0 = \frac{2d}{c}, \]  

(A.6)

where \( c \) is the speed of sound in the air. The energy level of the reverberant tail measured at the ears \( L_{\text{rev}} \) is

\[ L_{\text{rev}} = L_W + 10 \log \left( \frac{4}{R} \right) + \Delta L_{\text{HRTF}}, \]  

(A.7)

with \( R = S\bar{\alpha}/(1-\bar{\alpha}) \) being the so-called “room constant”, \( S \) the total surface area of the room, and \( \bar{\alpha} \) the mean absorption coefficient, which is derived from the volume \( V \) and the reverberation time \( T \) through Sabine’s formula \( \bar{\alpha} = 4 \ln(10^6) V/(cST) \). The energy density curve for the reverberant tail \( E_{\text{rev}}(t) \) is

\[ E_{\text{rev}}(t) = Be^{-t/\tau} \quad t > 0, \]  

(A.8)

with parameters

\[ \tau = \frac{T \log e}{3}, \]

(A.9)

\[ 10 \log B = 10 \log \left( \frac{4}{R} \right) + \Delta L_{\text{HRTF}} - K - 10 \log \tau, \]  

(A.10)

which ensure that

\[ 10 \log \left[ \int_0^\infty E_{\text{rev}}(t) \, dt \right] = L_{\text{rev}}. \]  

(A.11)

The total energy density \( E(t) \) is therefore \( E(t) = E_d(t) + E_{\text{FR}}(t) + E_{\text{rev}}(t) \) (see Fig. 8)

\[ E(t) = \delta(t) + A\delta(t-t_0) + Be^{-t/\tau} \quad t > 0. \]  

(A.12)

By applying Schroeder’s backward integration, the backward integrated energy curve \( R(t) = \int_t^\infty E(t) \, dt \) is

\[ R(t) = \begin{cases} 
1 + A + \tau B, & t = 0 \\
A + \tau Be^{-t/\tau}, & 0 < t < t_0 \\
\tau Be^{-t/\tau}, & t > t_0.
\end{cases} \]  

(A.13)

Introducing the level of the backward integrated energy curve \( L_R(t) = 10 \log R(t)/R(0) \), the quantity \( T_{30,\text{ears}} \) is finally found as

\[ T_{30,\text{ears}} = 2(t_{L_R=-35} - t_{L_R=-5}). \]  

(A.14)

The time values at which the level of the backward integrated energy curve are -5 dB \( (t_{L_R=-5}) \) and -35 dB \( (t_{L_R=-35}) \) do not have closed mathematical expressions and are obtained with a search algorithm in MATLAB.

Considering the octave band of 1 kHz, \( Q^* = 0.60, K = 4.0 \, \text{dB}, \Delta L_{\text{HRTF}} = 4 \, \text{dB}. \) The \( T_{30,\text{ears}} \) as a function of the room volume and the diffuse-field reverberation time \( T \) in the 1 kHz octave band is shown in Fig. 9. Although other values of the parameters can be used for different frequency bands, the model is not intended to generalize to a combination of frequency bands.

**FIG. 9.** Reverberation time at ears as a function of the room volume and the diffuse-field reverberation time in the 1 kHz octave band, as predicted with the model of Eq. (A.14)

12. E. Vilkman, “Voice problems at work: A challenge for oc-
30 W. Chu and A. Warnock, Detailed directivity of sound fields around human talkers, Institute for Research in Construc-