Research outputs:

Effect of Noise Reduction Gain Errors on Simulated Cochlear Implant Speech Intelligibility

It has been suggested that the most important factor for obtaining high speech intelligibility in noise with cochlear implant (CI) recipients is to preserve the low-frequency amplitude modulations of speech across time and frequency by, for example, minimizing the amount of noise in the gaps between speech segments. In contrast, it has also been argued that the transient parts of the speech signal, such as speech onsets, provide the most important information for speech intelligibility. The present study investigated the relative impact of these two factors on the potential benefit of noise reduction for CI recipients by systematically introducing noise estimation errors within speech segments, speech gaps, and the transitions between them. The introduction of these noise estimation errors directly induces errors in the noise reduction gains within each of these regions. Speech intelligibility in both stationary and modulated noise was then measured using a CI simulation tested on normal-hearing listeners. The results suggest that minimizing noise in the speech gaps can improve intelligibility, at least in modulated noise. However, significantly larger improvements were obtained when both the noise in the gaps was minimized and the speech transients were preserved. These results imply that the ability to identify the boundaries between speech segments and speech gaps may be one of the most important factors for a noise reduction algorithm because knowing the boundaries makes it possible to minimize the noise in the gaps as well as enhance the low-frequency amplitude modulations of the speech.
The impact of noise power estimation on speech intelligibility in cochlear-implant speech coding strategies

The advanced combination encoder (ACE™) is an established speech-coding strategy in cochlear-implant processing that selects a number of frequency channels based on amplitudes. However, speech intelligibility outcomes with this strategy are limited in noisy conditions. To improve speech intelligibility, either noise-dominant channels can be attenuated prior to ACE™ with noise reduction or, alternatively, channels can be selected based on estimated signal-to-noise ratios. A noise power estimation stage is, therefore, required. This study investigated the impact of noise power estimation in noise-reduction and channel-selection strategies. Results imply that estimation with improved noise-tracking capabilities does not necessarily translate into increased speech intelligibility.

General information
State: Published
Organisations: Hearing Systems, Department of Electrical Engineering, Department of Health Technology, Cochlear, Electromagnetic Systems Group
Contributors: Bentsen, T., Mauger, S. J., Kressner, A. A., May, T., Dau, T.
Pages: 818-821
Publication date: 1 Feb 2019
Peer-reviewed: Yes

Publication information
Journal: Journal of the Acoustical Society of America
Volume: 145
Issue number: 2
ISSN (Print): 0001-4966
Ratings:
BFI (2019): BFI-level 2
Web of Science (2019): Indexed yes
BFI (2018): BFI-level 2
Web of Science (2018): Indexed yes
BFI (2017): BFI-level 2
Scopus rating (2017): CiteScore 1.77 SJR 0.695 SNIP 1.224
Web of Science (2017): Impact factor 1.605
Web of Science (2017): Indexed yes
BFI (2016): BFI-level 2
Scopus rating (2016): CiteScore 1.83 SJR 0.819 SNIP 1.271
Web of Science (2016): Impact factor 1.547
Web of Science (2016): Indexed yes
BFI (2015): BFI-level 2
Scopus rating (2015): CiteScore 1.77 SJR 0.854 SNIP 1.416
Web of Science (2015): Impact factor 1.572
Web of Science (2015): Indexed yes
BFI (2014): BFI-level 2
Scopus rating (2014): CiteScore 1.8 SJR 0.887 SNIP 1.402
Web of Science (2014): Impact factor 1.503
Effects of Fast-Acting Hearing-Aid Compression on Audibility, Forward Masking and Speech Perception

Dynamic range compression (DRC) is a widely-used compensation strategy in hearing aids. However, the choice of the compression parameters, such as time constants, is still the subject of an ongoing debate. This contribution evaluates the efficacy of fast-acting DRC as a hearing-loss compensation strategy in a range of experimental conditions. First, fast-acting DRC was investigated considering temporal masking of narrowband stimuli. The results of a model-driven evaluation showed that the measures of temporal resolution can be improved with fast-acting compression with a very short release time (10 ms). Second, the effects of compression on speech audibility and noise-induced forward masking were evaluated in a highly-controlled scenario. The application of very short compression time constants was shown to improve HI listeners' consonant recognition performance. Finally, despite the benefits of fast-acting compression apparent in controlled conditions, it may introduce distortion in realistic scenarios, such as a reduction in the signal-to-noise ratio (SNR). A novel signal-to-noise-ratio-aware compensation strategy is discussed, which switches between fast- and slow-acting compression depending on the presence of the target signal and therefore preserves the natural relationship between the target and the background. An objective evaluation of the algorithm is presented and its potential applications are discussed.

General information
State: Published
Organisations: Department of Electrical Engineering, Hearing Systems, Sonova US Inc., Technical University of Denmark
Contributors: Kowalewski, B., May, T., Fereczkowski, M., Zaar, J., Strelcyk, O., MacDonald, E. N., Dau, T.
Number of pages: 9
Pages: 1-9
Publication date: 2018

Host publication information
Title of host publication: Proceedings of 2018 Joint Conference - Acoustics
ISBN (Print): 9781538671139
Keywords: Hearing aids, Compression, Speech intelligibility, Psychoacoustics, Audio signal processing
DOI: 10.1109/ACOUSTICS.2018.8502390
Source: FindIt
Source-ID: 2441007129
Research output: Research - peer-review › Article in proceedings – Annual report year: 2018

Effects of Slow- and Fast-Acting Compression on Hearing-Impaired Listeners' Consonant–Vowel Identification in Interrupted Noise

There is conflicting evidence about the relative benefit of slow- and fast-acting compression for speech intelligibility. It has been hypothesized that fast-acting compression improves audibility at low signal-to-noise ratios (SNRs) but may distort the speech envelope at higher SNRs. The present study investigated the effects of compression with a nearly instantaneous attack time but either fast (10ms) or slow (500ms) release times on consonant identification in hearing-impaired listeners. Consonant–vowel speech tokens were presented at a range of presentation levels in two conditions: in the presence of interrupted noise and in quiet (with the compressor “shadow-controlled” by the corresponding mixture of speech and noise). These conditions were chosen to disentangle the effects of consonant audibility and noise-induced forward masking on speech intelligibility. A small but systematic intelligibility benefit of fast-acting compression was found in both the quiet and the noisy conditions for the lower speech levels. No detrimental effects of fast-acting compression were observed when the speech level exceeded the level of the noise. These findings suggest that fast-acting compression provides an audibility benefit in fluctuating interferers when compared with slow-acting compression while not substantially affecting the perception of consonants at higher SNRs.

General information
State: Published
Organisations: Hearing Systems, Department of Electrical Engineering, Sonova US Inc.
Contributors: Kowalewski, B., Zaar, J., Fereczkowski, M., MacDonald, E., Strelcyk, O., May, T., Dau, T.
Pages: 1-12
Publication date: 2018
Peer-reviewed: Yes

Publication information
Journal: Trends in Hearing
Volume: 22
ISSN (Print): 2331-2165
Ratings:
BFI (2019): BFI-level 1
Web of Science (2019): Indexed yes
BFI (2018): BFI-level 1
Robust speech dereverberation with a neural network-based post-filter that exploits multi-conditional training of binaural cues

This study presents an algorithm for binaural speech dereverberation based on the supervised learning of short-term binaural cues. The proposed system combined a delay-and-sum beamformer (DSB) with a neural network-based post-filter that attenuated reverberant components in individual time-frequency (T-F) units. A multi-conditional training (MCT) procedure was used to simulate the uncertainties of short-term binaural cues in response to room reverberation by mixing the direct part of head related impulse responses (HRIRs) with diffuse noise. Despite being trained with only anechoic HRIRs, the proposed dereverberation algorithm was tested in a variety of reverberant environments and achieved considerable improvements relative to a coherence-based approach in terms of three objective metrics reflecting speech quality and speech intelligibility. Moreover, a systematic evaluation showed that the proposed system generalized very well to a wide range of acoustic conditions, including various measured binaural room impulse responses (BRIRs) reflecting different reverberation times, azimuth positions spanning the entire frontal hemifield, various source-receiver distances as well as different artificial heads.

General information
State: Published
Organisations: Department of Electrical Engineering, Hearing Systems
Contributors: May, T.
Pages: 406-414
Publication date: 2018
Peer-reviewed: Yes

Publication information
Journal: IEEE/ACM Transactions on Audio, Speech, and Language Processing
Volume: 26
Issue number: 2
ISSN (Print): 2329-9290
Ratings:
BFI (2019): BFI-level 2
Web of Science (2019): Indexed yes
BFI (2018): BFI-level 2
Web of Science (2018): Indexed yes
BFI (2017): BFI-level 2
Scopus rating (2017): CiteScore 4.4 SJR 0.867 SNIP 4.2
Web of Science (2017): Impact factor 2.95
Web of Science (2017): Indexed yes
Scopus rating (2016): CiteScore 3.5 SJR 0.762 SNIP 3.429
Web of Science (2016): Impact factor 2.491
Signal-to-Noise-Ratio-Aware Dynamic Range Compression in Hearing Aids

Fast-acting dynamic range compression is a level-dependent amplification scheme which aims to restore audibility for hearing impaired listeners. However, when being applied to noisy speech at positive signal-to-noise ratios (SNRs), the gain function typically changes rapidly over time as it is driven by the short-term fluctuations of the speech signal. This leads to an amplification of the noise components in the speech gaps, which reduces the output SNR and distorts the acoustic properties of the background noise. An adaptive compression scheme is proposed here which utilizes information about the SNR in different frequency channels to adaptively change the characteristics of the compressor. Specifically, fast-acting compression is applied to speech-dominated time-frequency (T-F) units where the SNR is high, while slow-acting compression is used to effectively linearize the processing for noise-dominated T-F units where the SNR is low. A systematic evaluation of this SNR-aware compression scheme showed that the effective compression of speech components embedded in noise was similar to that of a conventional fast-acting system, whereas natural fluctuations in
the background noise were preserved in a similar way as when a slow-acting compressor was applied.

**General information**
State: Published
Organisations: Department of Electrical Engineering, Hearing Systems
Contributors: May, T., Kowalewski, B., Dau, T.
Pages: 1–12
Publication date: 2018
Peer-reviewed: Yes

**Publication information**
Journal: Trends in Hearing
Volume: 22
ISSN (Print): 2331-2165
Ratings:
- BFI (2019): BFI-level 1
- Web of Science (2019): Indexed yes
- BFI (2018): BFI-level 1
- Web of Science (2018): Indexed yes
- Scopus rating (2017): CiteScore 1.95
- Web of Science (2017): Impact factor 2
- Web of Science (2017): Indexed yes
- Scopus rating (2016): CiteScore 3.61
- Web of Science (2016): Impact factor 3.024
- Web of Science (2016): Indexed yes
- Scopus rating (2015): CiteScore 2.21
- Web of Science (2015): Impact factor 2.158
- Web of Science (2015): Indexed yes
- Scopus rating (2014): CiteScore 0
- Web of Science (2014): Impact factor 0
- Web of Science (2014): Indexed yes
Original language: English
Keywords: Wide dynamic range compression, Signal-to-noise ratio, Hearing-aid signal processing

**Electronic versions:**
2331216518790903.pdf
DOIs:
10.1177/2331216518790903
Source: PublicationPreSubmission
Source-ID: 152113173
Research output: Research - peer-review; Journal article – Annual report year: 2018

The benefit of combining a deep neural network architecture with ideal ratio mask estimation in computational speech segregation to improve speech intelligibility
Computational speech segregation attempts to automatically separate speech from noise. This is challenging in conditions with interfering talkers and low signal-to-noise ratios.
Recent approaches have adopted deep neural networks and successfully demonstrated speech intelligibility improvements. A selection of components may be responsible for the success with these state-of-the-art approaches: the system architecture, a time frame concatenation technique and the learning objective. The aim of this study was to explore the roles and the relative contributions of these components by measuring speech intelligibility in normal-hearing listeners. A substantial improvement of 25.4 percentage points in speech intelligibility scores was found going from a subband-based architecture, in which a Gaussian Mixture Model-based classifier predicts the distributions of speech and noise for each frequency channel, to a state-of-the-art deep neural network-based architecture. Another improvement of 13.9 percentage points was obtained by changing the learning objective from the ideal binary mask, in which individual time-frequency units are labeled as either speech- or noise-dominated, to the ideal ratio mask, where the units are assigned a continuous value between zero and one. Therefore, both components play significant roles and by combining them, speech intelligibility improvements were obtained in a six-talker condition at a low signal-to-noise ratio.

**General information**
State: Published
Organisations: Department of Electrical Engineering, Hearing Systems
The impact of exploiting spectro-temporal context in computational speech segregation

Computational speech segregation aims to automatically segregate speech from interfering noise, often by employing ideal binary mask estimation. Several studies have tried to exploit contextual information in speech to improve mask estimation accuracy by using two frequently-used strategies that (1) incorporate delta features and (2) employ support vector machine (SVM) based integration. In this study, two experiments were conducted. In Experiment I, the impact of exploiting spectro-temporal context using these strategies was investigated in stationary and six-talker noise. In Experiment II, the delta features were explored in detail and tested in a setup that considered novel noise segments of the six-talker noise. Computing delta features led to higher intelligibility than employing SVM based integration and intelligibility increased with the amount of spectral information exploited via the delta features. The system did not, however, generalize well to novel segments of this noise type. Measured intelligibility was subsequently compared to extended short-term objective intelligibility, hit–false alarm rate, and the amount of mask clustering. None of these objective measures alone could account for measured intelligibility. The findings may have implications for the design of speech segregation systems, and for the selection of a cost function that correlates with intelligibility.

General information
State: Published
Organisations: Department of Electrical Engineering, Hearing Systems
Contributors: Bentsen, T., Kressner, A. A., Dau, T., May, T.
Pages: 248-259
Publication date: 2018
Peer-reviewed: Yes

Publication information
Journal: Journal of the Acoustical Society of America
Volume: 143
Issue number: 1
ISSN (Print): 0001-4966
Ratings:
BFI (2019): BFI-level 2
Web of Science (2019): Indexed yes
BFI (2018): BFI-level 2
Web of Science (2018): Indexed yes
BFI (2017): BFI-level 2
Scopus rating (2017): CiteScore 1.77 SJR 0.695 SNIP 1.224
Web of Science (2017): Impact factor 1.605
Web of Science (2017): Indexed yes
BFI (2016): BFI-level 2
Scopus rating (2016): CiteScore 1.83 SJR 0.819 SNIP 1.271
Web of Science (2016): Impact factor 1.547
Web of Science (2016): Indexed yes
BFI (2015): BFI-level 2
Scopus rating (2015): CiteScore 1.77 SJR 0.854 SNIP 1.416
Web of Science (2015): Impact factor 1.572
Web of Science (2015): Indexed yes
BFI (2014): BFI-level 2
Scopus rating (2014): CiteScore 1.8 SJR 0.887 SNIP 1.402
Web of Science (2014): Impact factor 1.503
Web of Science (2014): Indexed yes
BFI (2013): BFI-level 2
Scopus rating (2013): CiteScore 2 SJR 0.707 SNIP 1.937
Web of Science (2013): Impact factor 1.555
ISI indexed (2013): ISI indexed yes
Web of Science (2013): Indexed yes
An accurate estimation of the broadband input signal-to-noise ratio (SNR) is a prerequisite for many hearing-aid algorithms. An extensive comparison of three SNR estimation algorithms was performed. Moreover, the influence of the duration of the analysis window on the SNR estimation performance was systematically investigated.

The most accurate approach utilized an estimation of the clean speech power spectral density (PSD) and the noisy speech power across a sliding window of 1280 ms and achieved an overall SNR estimation error below 3 dB across a wide variety of background noises and input SNRs.

State: Published
Effects of slow- and fast-acting compression on hearing impaired listeners' consonant-vowel identification in interrupted noise

There is conflicting evidence about the relative benefit of slow- and fast-acting compression for speech intelligibility. It has been hypothesized that fast-acting compression improves audibility at low signal-to-noise ratios (SNRs) but may distort the speech envelope at higher SNRs. The present study investigated the effects of compression with nearly instantaneous attack time but either fast (10 ms) or slow (500 ms) release times on consonant identification in hearing-impaired listeners. Consonant-vowel speech tokens were presented at several presentation levels in two conditions: in the presence of interrupted noise and in quiet (with the compressor "shadow controlled" by the corresponding mixture of speech and noise). These conditions were chosen to disentangle the effects of consonant audibility and noise-induced forward masking on speech intelligibility. A small but systematic intelligibility benefit of fast-acting compression was found in both the quiet and the noisy conditions for the lower speech levels. No negative effects of fast-acting compression were observed when the speech level exceeded the level of the noise. These findings suggest that fast-acting compression provides an audibility benefit in fluctuating interferers as compared to slow-acting compression, while not substantially affecting the perception of consonants at higher SNRs.

Exploiting Deep Neural Networks and Head Movements for Robust Binaural Localization of Multiple Sources in Reverberant Environments

This paper presents a novel machine-hearing system that exploits deep neural networks (DNNs) and head movements for robust binaural localization in reverberant environments. DNNs are used to learn the relationship between the source azimuth and binaural cues, consisting of the complete cross-correlation function (CCF) and interaural level differences (ILDs). In contrast to many previous binaural hearing systems, the proposed approach is not restricted to localization of sound sources in the frontal hemifield. Due to the similarity of binaural cues in the frontal and rear hemifields, front–back confusions often occur. To address this, a head movement strategy is incorporated in the localization model to help reduce the front–back errors. The proposed DNN system is compared to a Gaussian-mixture-model-based system that employs interaural time differences (ITDs) and ILDs as localization features. Our experiments show that the DNN is able to exploit information in the CCF that is not available in the ITD cue, which together with head movements substantially improves localization accuracies under challenging acoustic scenarios, in which multiple talkers and room reverberation are present.
Influence of binary mask estimation errors on robust speaker identification

Missing-data strategies have been developed to improve the noise-robustness of automatic speech recognition systems in adverse acoustic conditions. This is achieved by classifying time-frequency (T-F) units into reliable and unreliable components, as indicated by a so-called binary mask. Different approaches have been proposed to handle unreliable feature components, each with distinct advantages. The direct masking (DM) approach attenuates unreliable T-F units in the spectral domain, which allows the extraction of conventionally used mel-frequency cepstral coefficients (MFCCs). Instead of attenuating unreliable components in the feature extraction front-end, full marginalization (FM) discards unreliable feature components in the classification back-end. Finally, bounded marginalization (BM) can be used to combine the evidence from both reliable and unreliable feature components during classification. Since each of these approaches utilizes the knowledge about reliable and unreliable feature components in a different way, they will respond differently to estimation errors in the binary mask. The goal of this study was to identify the most effective strategy to exploit knowledge about reliable and unreliable feature components in the context of automatic speaker identification (SID). A systematic evaluation under ideal and non-ideal conditions demonstrated that the robustness to errors in the binary mask varied substantially across the different missing-data strategies. Moreover, full and bounded marginalization showed complementary performances in stationary and non-stationary background noises and were subsequently combined using a simple score fusion. This approach consistently outperformed individual SID systems in all considered experimental conditions.
Web of Science (2014): Impact factor 1.256
BFI (2013): BFI-level 1
Scopus rating (2013): CiteScore 2.71 SJR 0.566 SNIP 2.091
Web of Science (2013): Impact factor 1.548
ISI indexed (2013): ISI indexed yes
BFI (2012): BFI-level 1
Scopus rating (2012): CiteScore 2.38 SJR 0.594 SNIP 2.382
Web of Science (2012): Impact factor 1.283
ISI indexed (2012): ISI indexed yes
BFI (2011): BFI-level 2
Scopus rating (2011): CiteScore 2.13 SJR 0.726 SNIP 2.41
Web of Science (2011): Impact factor 1.267
ISI indexed (2011): ISI indexed yes
BFI (2010): BFI-level 2
Scopus rating (2010): SJR 0.537 SNIP 1.824
Web of Science (2010): Impact factor 1.012
Web of Science (2010): Indexed yes
BFI (2009): BFI-level 1
Scopus rating (2009): SJR 0.733 SNIP 2.266
Web of Science (2009): Indexed yes
BFI (2008): BFI-level 2
Scopus rating (2008): SJR 0.767 SNIP 1.481
Scopus rating (2007): SJR 0.513 SNIP 1.269
Scopus rating (2006): SJR 0.574 SNIP 1.959
Scopus rating (2005): SJR 0.455 SNIP 1.736
Scopus rating (2004): SJR 0.396 SNIP 1.616
Scopus rating (2003): SJR 0.263 SNIP 1.587
Scopus rating (2002): SJR 0.688 SNIP 1.865
Scopus rating (2001): SJR 0.717 SNIP 1.563
Scopus rating (2000): SJR 0.234 SNIP 1.806
Scopus rating (1999): SJR 0.224 SNIP 0.706
Original language: English
Keywords: Bounded marginalization, Direct masking, Estimated binary mask, Full marginalization, Ideal binary mask, Missing data, Speaker identification
Electronic versions:
Influence of binary mask estimation
DOIs:

Bibliographical note
Published under a Creative Commons license
Source: FindIt
Source-ID: 2349817058
Research output: Research - peer-review › Journal article – Annual report year: 2017

Investigating the effects of noise-estimation errors in simulated cochlear implant speech intelligibility
A recent study suggested that the most important factor for obtaining high speech intelligibility in noise with cochlear implant recipients is to preserve the low-frequency amplitude modulations of speech across time and frequency by, for example, minimizing the amount of noise in speech gaps. In contrast, other studies have argued that the transients provide the most information. Thus, the present study investigates the relative impact of these two factors in the framework of noise reduction by systematically correcting noise-estimation errors within speech segments, speech gaps, and the transitions between them. Speech intelligibility in noise was measured using a cochlear implant simulation tested on normal-hearing listeners. The results suggest that minimizing noise in the speech gaps can substantially improve intelligibility, especially in modulated noise. However, significantly larger improvements were obtained when both the noise in the gaps was minimized and the speech transients were preserved. These results imply that the correct identification of the boundaries between speech segments and speech gaps is the most important factor in maintaining high intelligibility in cochlear implants. Knowing the boundaries will make it possible for algorithms to both minimize the noise in the gaps and enhance the low frequency amplitude modulations.
Prediction of speech intelligibility based on a correlation metric in the envelope power spectrum domain

A powerful tool to investigate speech perception is the use of speech intelligibility prediction models. Recently, a model was presented, termed correlation-based speech envelope power spectrum model (sEPSMcorr) [1], based on the auditory processing of the multi-resolution speech-based Envelope Power Spectrum Model (mr-sEPSM) [2], combined with the correlation back-end of the Short-Time Objective Intelligibility measure (STOI) [3]. The sEPSMcorr can accurately predict NH data for a broad range of listening conditions, e.g., additive noise, phase jitter and ideal binary mask processing.

Preserving spatial perception in rooms using direct-sound driven dynamic range compression

Fast-acting hearing-aid compression systems typically distort the auditory cues involved in the spatial perception of sounds in rooms by enhancing low-level reverberant energy portions of the sound relative to the direct sound. The present study investigated the benefit of a direct-sound driven compression system that adaptively selects appropriate time constants to preserve the listener's spatial impression. Specifically, fast-acting compression was maintained for time-frequency units dominated by the direct sound while the processing of the compressor was linearized for time-frequency units dominated by reverberation. This compression scheme was evaluated with normal-hearing listeners who indicated their perceived location and distribution of sound images in the horizontal plane for virtualized speech. The experimental results confirmed that both independent compression at each ear and linked compression across ears resulted in broader, sometimes internalized, sound images as well as image splits. In contrast, the linked direct-sound driven compression system provided the listeners with a spatial perception similar to that obtained with linear processing that served as the reference condition. The independent direct-sound driven compressor created a sense of movement of the sound between the two ears, suggesting that preserving the interaural level differences via linked compression is advantageous with the proposed direct-sound driven compression scheme.
A correlation metric in the envelope power spectrum domain for speech intelligibility prediction

A powerful tool to investigate speech perception is the use of speech intelligibility prediction models. Recently, a model was presented, termed correlation-based speech-based envelope power spectrum model (sEPSPMcorr) [1], based on the auditory processing of the multi-resolution speech-based Envelope Power Spectrum Model (mr-sEPSPM) [2], combined with the correlation back-end of the Short-Time Objective Intelligibility measure (STOI) [3]. The sEPSPMcorr can accurately predict NH data for a broad range of listening conditions, e.g., additive noise, phase jitter and ideal binary mask processing.

General information
State: Published
Organisations: Hearing Systems, Department of Electrical Engineering
Contributors: Relaño-Iborra, H., May, T., Zaar, J., Scheidiger, C., Dau, T.
Publication date: 2016
Peer-reviewed: Yes
Event: Poster session presented at ARCHES/ICANHEAR 2016, Zurich, Switzerland.
Electronic versions: ARCHES_poster_final3.pdf
Research output: Research - peer-review › Poster – Annual report year: 2017

Assessing and modeling apparent source width perception

General information
State: Published
Organisations: Department of Electrical Engineering, Hearing Systems, Technical University of Denmark
Contributors: Käsbach, J., Hahmann, M., May, T., Dau, T.
Number of pages: 4
Publication date: 2016

Host publication information
Title of host publication: Proceedings of DAGA 2016
Publisher: Deutsche Gesellschaft für Akustik e.V.

Bibliographical note
Proceedings of DAGA (German acoustic association) 2016 in Aachen, Germany
Source: PublicationPreSubmission
Source-ID: 123771527
Assessing the contribution of binaural cues for apparent source width perception via a functional model

In echoic conditions, sound sources are not perceived as point sources but appear to be expanded. The expansion in the horizontal dimension is referred to as apparent source width (ASW). To elicit this perception, the auditory system has access to fluctuations of binaural cues, the interaural time differences (ITDs), interaural level differences (ILDs) and the interaural coherence (IC). To quantify their contribution to ASW, a functional model of ASW perception was exploited using the TWO!EARS auditory-front-end (AFE) toolbox. The model determines the left and right-most boundary of a sound source using a statistical representation of ITDs and ILDs based on percentiles integrated over time and frequency. The model's performance was evaluated against psychoacoustic data obtained with noise, speech and music signals in loudspeaker-based experiments. A robust model prediction of ASW was achieved using a cross-correlation based estimation with either IC or ITDs, in contrast to a combination of ITDs and ILDs where the performance slightly decreased.

Comparing the influence of spectro-temporal integration in computational speech segregation

The goal of computational speech segregation systems is to automatically segregate a target speaker from interfering maskers. Typically, these systems include a feature extraction stage in the front-end and a classification stage in the back-end. A spectrotemporal integration strategy can be applied in either the front-end, using the so-called delta features, or in the back-end, using a second classifier that exploits the posterior probability of speech from the first classifier across a spectro-temporal window. This study systematically analyzes the influence of such stages on segregation performance, the error distributions and intelligibility predictions. Results indicated that it could be problematic to exploit context in the back-end, even though such a spectro-temporal integration stage improves the segregation performance. Also, the results emphasized the potential need of a single metric that comprehensively predicts computational segregation performance and correlates well with intelligibility. The outcome of this study could help to identify the most effective spectro-temporal integration strategy for computational segregation systems.

Outcome measures based on classification performance fail to predict the intelligibility of binary-masked speech

To date, the most commonly used outcome measure for assessing ideal binary mask estimation algorithms is based on the difference between the hit rate and the false alarm rate (H-FA). Recently, the error distribution has been shown to substantially affect intelligibility. However, H-FA treats each mask unit independently and does not take into account how errors are distributed. Alternatively, algorithms can be evaluated with the short-time objective intelligibility (STOI) metric.
using the reconstructed speech. This study investigates the ability of H-FA and STOI to predict intelligibility for binary-masked speech using masks with different error distributions. The results demonstrate the inability of H-FA to predict the behavioral intelligibility and also illustrate the limitations of STOI. Since every estimation algorithm will make errors that are distributed in different ways, performance evaluations should not be made solely on the basis of these metrics.

**General information**
State: Published
Organisations: Department of Electrical Engineering, Hearing Systems, Georgia Institute of Technology
Contributors: Kressner, A. A., May, T., Rozell, C. J.
Pages: 3033–3036
Publication date: 2016
Peer-reviewed: Yes

**Publication information**
Journal: Journal of the Acoustical Society of America
Volume: 139
Issue number: 6
ISSN (Print): 0001-4966
Ratings:
BFI (2019): BFI-level 2
Web of Science (2019): Indexed yes
BFI (2018): BFI-level 2
Web of Science (2018): Indexed yes
BFI (2017): BFI-level 2
Scopus rating (2017): CiteScore 1.77 SJR 0.695 SNIP 1.224
Web of Science (2017): Impact factor 1.605
Web of Science (2017): Indexed yes
BFI (2016): BFI-level 2
Scopus rating (2016): CiteScore 1.83 SJR 0.819 SNIP 1.271
Web of Science (2016): Impact factor 1.547
Web of Science (2016): Indexed yes
BFI (2015): BFI-level 2
Scopus rating (2015): CiteScore 1.77 SJR 0.854 SNIP 1.416
Web of Science (2015): Impact factor 1.572
Web of Science (2015): Indexed yes
BFI (2014): BFI-level 2
Scopus rating (2014): CiteScore 1.8 SJR 0.887 SNIP 1.402
Web of Science (2014): Impact factor 1.503
Web of Science (2014): Indexed yes
BFI (2013): BFI-level 2
Scopus rating (2013): CiteScore 2 SJR 0.707 SNIP 1.937
Web of Science (2013): Impact factor 1.555
ISI indexed (2013): ISI indexed yes
Web of Science (2013): Indexed yes
BFI (2012): BFI-level 2
Scopus rating (2012): CiteScore 1.75 SJR 0.771 SNIP 1.619
Web of Science (2012): Impact factor 1.646
ISI indexed (2012): ISI indexed yes
Web of Science (2012): Indexed yes
BFI (2011): BFI-level 2
Scopus rating (2011): CiteScore 1.68 SJR 0.686 SNIP 1.624
Web of Science (2011): Impact factor 1.55
ISI indexed (2011): ISI indexed yes
Web of Science (2011): Indexed yes
BFI (2010): BFI-level 2
Scopus rating (2010): SJR 0.734 SNIP 1.511
Web of Science (2010): Impact factor 1.644
A speech intelligibility prediction model is proposed that combines the auditory processing front end of the multi-resolution speech-based envelope power spectrum model [mr-sEPSM; Jørgensen, Ewert, and Dau (2013). J. Acoust. Soc. Am. 134(1), 436–446] with a correlation back end inspired by the short-time objective intelligibility measure [STOI; Taal, Hendriks, Heusdens, and Jensen (2011). IEEE Trans. Audio Speech Lang. Process. 19(7), 2125–2136]. This “hybrid” model, named sEPSMcorr, is shown to account for the effects of stationary and fluctuating additive interferers as well as for the effects of non-linear distortions, such as spectral subtraction, phase jitter, and ideal time frequency segregation (ITFS). The model shows a broader predictive range than both the original mr-sEPSM (which fails in the phase-jitter and ITFS conditions) and STOI (which fails to predict the influence of fluctuating interferers), albeit with lower accuracy than the source models in some individual conditions. Similar to other models that employ a short-term correlation-based back end, including STOI, the proposed model fails to account for the effects of room reverberation on speech intelligibility. Overall, the model might be valuable for evaluating the effects of a large range of interferers and distortions on speech intelligibility, including consequences of hearing impairment and hearing-instrument signal processing.
A machine-hearing system exploiting head movements for binaural sound localisation in reverberant conditions

This paper is concerned with machine localisation of multiple active speech sources in reverberant environments using two (binaural) microphones. Such conditions typically present a problem for ‘classical’ binaural models. Inspired by the human ability to utilise head movements, the current study investigated the influence of different head movement strategies on binaural sound localisation. A machine-hearing system that exploits a multi-step head rotation strategy for sound localisation was found to produce the best performance in simulated reverberant acoustic space. This paper also reports the public release of a free binaural room impulse responses (BRIRs) database that allows the simulation of head rotation used in this study.

General information
State: Published
Organisations: Department of Electrical Engineering, Hearing Systems, University of Sheffield, Technische Universität Berlin
Contributors: May, T., Ma, N., Wierstorf, H., Brown, G.
Pages: 2699-2703
Publication date: 2015

Host publication information
Title of host publication: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2015)
Publisher: IEEE
ISBN (Print): 9781467369978
Keywords: Head movements, Binaural localisation, Machine hearing, Reverberation, Binaural room impulse response
Electronic versions:
MaMayWierstorfBrown_ICASSP2015.pdf
DOIs: 10.1109/ICASSP.2015.7178461
Source: PublicationPreSubmission
Source-ID: 105091897
Research output: Research - peer-review › Article in proceedings – Annual report year: 2015
Exploiting deep neural networks and head movements for binaural localisation of multiple speakers in reverberant conditions

This paper presents a novel machine-hearing system that exploits deep neural networks (DNNs) and head movements for binaural localisation of multiple speakers in reverberant conditions. DNNs are used to map binaural features, consisting of the complete crosscorrelation function (CCF) and interaural level differences (ILDs), to the source azimuth. Our approach was evaluated using a localisation task in which sources were located in a full 360-degree azimuth range. As a result, front-back confusions often occurred due to the similarity of binaural features in the front and rear hemifields. To address this, a head movement strategy was incorporated in the DNN-based model to help reduce the front-back errors. Our experiments show that, compared to a system based on a Gaussian mixture model (GMM) classifier, the proposed DNN system substantially reduces localisation errors under challenging acoustic scenarios in which multiple speakers and room reverberation are present.

Robust localisation of multiple speakers exploiting head movements and multi-conditional training of binaural cues

This paper addresses the problem of localising multiple competing speakers in the presence of room reverberation, where sound sources can be positioned at any azimuth on the horizontal plane. To reduce the amount of front-back confusions which can occur due to the similarity of interaural time differences (ITDs) and interaural level differences (ILDs) in the front and rear hemifield, a machine hearing system is presented which combines supervised learning of binaural cues using multi-conditional training (MCT) with a head movement strategy. A systematic evaluation showed that this approach substantially reduced the amount of front-back confusions in challenging acoustic scenarios. Moreover, the system was able to generalise to a variety of different acoustic conditions not seen during training.
The role of temporal resolution in modulation-based speech segregation

This study is concerned with the challenge of automatically segregating a target speech signal from interfering background noise. A computational speech segregation system is presented which exploits logarithmically-scaled amplitude modulation spectrogram (AMS) features to distinguish between speech and noise activity on the basis of individual time-frequency (T-F) units. One important parameter of the segregation system is the window duration of the analysis-synthesis stage, which determines the lower limit of modulation frequencies that can be represented but also the temporal acuity with which the segregation system can manipulate individual T-F units. To clarify the consequences of this trade-off on modulation-based speech segregation performance, the influence of the window duration was systematically investigated.

Computational speech segregation based on an auditory-inspired modulation analysis

A monaural speech segregation system is presented that estimates the ideal binary mask from noisy speech based on the supervised learning of amplitude modulation spectrogram (AMS) features. Instead of using linearly scaled modulation filters with constant absolute bandwidth, an auditory-inspired modulation filterbank with logarithmically scaled filters is employed. To reduce the dependency of the AMS features on the overall background noise level, a feature normalization stage is applied. In addition, a spectro-temporal integration stage is incorporated in order to exploit the context information about speech activity present in neighboring time-frequency units. In order to evaluate the generalization performance of the system to unseen acoustic conditions, the speech segregation system is trained with a limited set of low signal-to-noise ratio (SNR) conditions, but tested over a wide range of SNRs up to 20dB. A systematic evaluation of the system demonstrates that auditory-inspired modulation processing can substantially improve the mask estimation accuracy in the presence of stationary and fluctuating interferers.
Web of Science (2017): Impact factor 1.605
Web of Science (2017): Indexed yes
BFI (2016): BFI-level 2
Scopus rating (2016): CiteScore 1.83 SJR 0.819 SNIP 1.271
Web of Science (2016): Impact factor 1.547
Web of Science (2016): Indexed yes
BFI (2015): BFI-level 2
Scopus rating (2015): CiteScore 1.77 SJR 0.854 SNIP 1.416
Web of Science (2015): Impact factor 1.572
Web of Science (2015): Indexed yes
BFI (2014): BFI-level 2
Scopus rating (2014): CiteScore 1.8 SJR 0.887 SNIP 1.402
Web of Science (2014): Impact factor 1.503
Web of Science (2014): Indexed yes
BFI (2013): BFI-level 2
Scopus rating (2013): CiteScore 2 SJR 0.707 SNIP 1.937
Web of Science (2013): Impact factor 1.555
ISI indexed (2013): ISI indexed yes
Web of Science (2013): Indexed yes
BFI (2012): BFI-level 2
Scopus rating (2012): CiteScore 1.75 SJR 0.771 SNIP 1.619
Web of Science (2012): Impact factor 1.646
ISI indexed (2012): ISI indexed yes
Web of Science (2012): Indexed yes
BFI (2011): BFI-level 2
Scopus rating (2011): CiteScore 1.68 SJR 0.686 SNIP 1.624
Web of Science (2011): Impact factor 1.55
ISI indexed (2011): ISI indexed yes
Web of Science (2011): Indexed yes
BFI (2010): BFI-level 2
Scopus rating (2010): SJR 0.734 SNIP 1.511
Web of Science (2010): Impact factor 1.644
Web of Science (2010): Indexed yes
BFI (2009): BFI-level 2
Scopus rating (2009): SJR 0.778 SNIP 1.692
Web of Science (2009): Indexed yes
BFI (2008): BFI-level 2
Scopus rating (2008): SJR 0.83 SNIP 1.657
Web of Science (2008): Indexed yes
Scopus rating (2007): SJR 0.838 SNIP 1.635
Web of Science (2007): Indexed yes
Scopus rating (2006): SJR 0.739 SNIP 1.678
Web of Science (2006): Indexed yes
Scopus rating (2005): SJR 0.946 SNIP 1.728
Web of Science (2005): Indexed yes
Scopus rating (2004): SJR 0.77 SNIP 1.761
Web of Science (2004): Indexed yes
Scopus rating (2003): SJR 0.875 SNIP 1.695
Web of Science (2003): Indexed yes
Scopus rating (2002): SJR 0.785 SNIP 1.572
Web of Science (2002): Indexed yes
Scopus rating (2001): SJR 0.727 SNIP 1.483
Web of Science (2001): Indexed yes
Generalization of Supervised Learning for Binary Mask Estimation

This paper addresses the problem of speech segregation by estimating the ideal binary mask (IBM) from noisy speech. Two methods will be compared, one supervised learning approach that incorporates a priori knowledge about the feature distribution observed during training. The second method solely relies on a frame-based speech presence probability (SPP) estimation, and therefore, does not depend on the acoustic condition seen during training. We investigate the influence of mismatches between the acoustic conditions used for training and testing on the IBM estimation performance and discuss the advantages of both approaches.

General information
State: Published
Organisations: Department of Electrical Engineering, Hearing Systems, University of Oldenburg
Contributors: May, T., Gerkmann, T.
Number of pages: 5
Publication date: 2014

Requirements for the evaluation of computational speech segregation systems
Recent studies on computational speech segregation reported improved speech intelligibility in noise when estimating and applying an ideal binary mask with supervised learning algorithms. However, an important requirement for such systems in technical applications is their robustness to acoustic conditions not considered during training. This study demonstrates that the spectro-temporal noise variations that occur during training and testing determine the achievable segregation performance. In particular, such variations strongly affect the identification of acoustical features in the system associated with perceptual attributes in speech segregation. The results could help establish a framework for a systematic evaluation of future segregation systems.

General information
State: Published
Organisations: Department of Electrical Engineering, Hearing Systems
Contributors: May, T., Dau, T.
Number of pages: 7
Pages: 398-404
Publication date: 2014
Peer-reviewed: Yes
Web of Science (2018): Indexed yes
BFI (2017): BFI-level 2
Scopus rating (2017): CiteScore 1.77 SJR 0.695 SNIP 1.224
Web of Science (2017): Impact factor 1.605
Web of Science (2017): Indexed yes
BFI (2016): BFI-level 2
Scopus rating (2016): CiteScore 1.83 SJR 0.819 SNIP 1.271
Web of Science (2016): Impact factor 1.547
Web of Science (2016): Indexed yes
BFI (2015): BFI-level 2
Scopus rating (2015): CiteScore 1.77 SJR 0.854 SNIP 1.416
Web of Science (2015): Impact factor 1.572
Web of Science (2015): Indexed yes
BFI (2014): BFI-level 2
Scopus rating (2014): CiteScore 1.8 SJR 0.887 SNIP 1.402
Web of Science (2014): Impact factor 1.503
Web of Science (2014): Indexed yes
BFI (2013): BFI-level 2
Scopus rating (2013): CiteScore 2 SJR 0.707 SNIP 1.937
Web of Science (2013): Impact factor 1.555
ISI indexed (2013): ISI indexed yes
Web of Science (2013): Indexed yes
BFI (2012): BFI-level 2
Scopus rating (2012): CiteScore 1.75 SJR 0.771 SNIP 1.619
Web of Science (2012): Impact factor 1.646
ISI indexed (2012): ISI indexed yes
Web of Science (2012): Indexed yes
BFI (2011): BFI-level 2
Scopus rating (2011): CiteScore 1.68 SJR 0.686 SNIP 1.624
Web of Science (2011): Impact factor 1.55
ISI indexed (2011): ISI indexed yes
Web of Science (2011): Indexed yes
BFI (2010): BFI-level 2
Scopus rating (2010): SJR 0.734 SNIP 1.511
Web of Science (2010): Impact factor 1.644
Web of Science (2010): Indexed yes
BFI (2009): BFI-level 2
Scopus rating (2009): SJR 0.778 SNIP 1.692
Web of Science (2009): Indexed yes
BFI (2008): BFI-level 2
Scopus rating (2008): SJR 0.83 SNIP 1.657
Web of Science (2008): Indexed yes
Scopus rating (2007): SJR 0.838 SNIP 1.635
Web of Science (2007): Indexed yes
Scopus rating (2006): SJR 0.739 SNIP 1.678
Web of Science (2006): Indexed yes
Scopus rating (2005): SJR 0.946 SNIP 1.728
Web of Science (2005): Indexed yes
Scopus rating (2004): SJR 0.77 SNIP 1.761
Web of Science (2004): Indexed yes
Scopus rating (2003): SJR 0.875 SNIP 1.695
Web of Science (2003): Indexed yes
Scopus rating (2002): SJR 0.785 SNIP 1.572
The effect of interaural-time-difference fluctuations on apparent source width
For the perception of spaciousness, the temporal fluctuations of the interaural time differences (ITDs) and interaural level differences (ILDs) provide important binaural cues. One major characteristic of spatial perception is apparent source width (ASW), which describes the perceived width of a sound image. The temporal fluctuations of the binaural cues cause the signals at a listeners’ ears to be decorrelated. Therefore, ASW has traditionally been measured by using the interaural cross- correlation (IACC). In particular, ITD fluctuations (below 2kHz) have been suggested to be the dominant cue for the perception of ASW. However, the contribution of the ITD statistics on the percept of ASW has not yet been clarified. In the present study, the impact of ITD fluctuations in different frequency bands on the perceived ASW was investigated. In a psychoacoustic evaluation, a source signal was convolved with individual binaural room impulse responses (BRIRs) and presented to the listener via headphones. The obtained signals were passed through a gammatone filterbank with an analysis and synthesis stage which enabled the modification of the ITD fluctuation statistics in individual frequency bands. The ITD fluctuations of broadband noise stimuli were compressed while the effect of this compression on the ILD statistics was kept minimal. The IACC was kept the same for stimuli with compression below 2kHz and for the uncompressed noise which should lead to the same ASW percept in the two conditions. However, the psychoacoustic data showed a reduced ASW for the modified signals, particularly in conditions with an applied compression around 1 kHz. In contrast, above 2kHz, the compression had no effect on ASW, whereas the IACC increased. The results suggest that the broadband IACC can be a misleading objective measure of ASW and that ITD fluctuations around 1kHz are crucial for ASW perception.

The importance of binaural cues for the perception of apparent source width at different sound pressure levels

The importance of binaural cues for the perception of apparent source width at different sound pressure levels

The effect of interaural-time-difference fluctuations on apparent source width
For the perception of spaciousness, the temporal fluctuations of the interaural time differences (ITDs) and interaural level differences (ILDs) provide important binaural cues. One major characteristic of spatial perception is apparent source width (ASW), which describes the perceived width of a sound image. The temporal fluctuations of the binaural cues cause the signals at a listeners’ ears to be decorrelated. Therefore, ASW has traditionally been measured by using the interaural cross- correlation (IACC). In particular, ITD fluctuations (below 2kHz) have been suggested to be the dominant cue for the perception of ASW. However, the contribution of the ITD statistics on the percept of ASW has not yet been clarified. In the present study, the impact of ITD fluctuations in different frequency bands on the perceived ASW was investigated. In a psychoacoustic evaluation, a source signal was convolved with individual binaural room impulse responses (BRIRs) and presented to the listener via headphones. The obtained signals were passed through a gammatone filterbank with an analysis and synthesis stage which enabled the modification of the ITD fluctuation statistics in individual frequency bands. The ITD fluctuations of broadband noise stimuli were compressed while the effect of this compression on the ILD statistics was kept minimal. The IACC was kept the same for stimuli with compression below 2kHz and for the uncompressed noise which should lead to the same ASW percept in the two conditions. However, the psychoacoustic data showed a reduced ASW for the modified signals, particularly in conditions with an applied compression around 1 kHz. In contrast, above 2kHz, the compression had no effect on ASW, whereas the IACC increased. The results suggest that the broadband IACC can be a misleading objective measure of ASW and that ITD fluctuations around 1kHz are crucial for ASW perception.

The importance of binaural cues for the perception of apparent source width at different sound pressure levels

The importance of binaural cues for the perception of apparent source width at different sound pressure levels

The effect of interaural-time-difference fluctuations on apparent source width
For the perception of spaciousness, the temporal fluctuations of the interaural time differences (ITDs) and interaural level differences (ILDs) provide important binaural cues. One major characteristic of spatial perception is apparent source width (ASW), which describes the perceived width of a sound image. The temporal fluctuations of the binaural cues cause the signals at a listeners’ ears to be decorrelated. Therefore, ASW has traditionally been measured by using the interaural cross- correlation (IACC). In particular, ITD fluctuations (below 2kHz) have been suggested to be the dominant cue for the perception of ASW. However, the contribution of the ITD statistics on the percept of ASW has not yet been clarified. In the present study, the impact of ITD fluctuations in different frequency bands on the perceived ASW was investigated. In a psychoacoustic evaluation, a source signal was convolved with individual binaural room impulse responses (BRIRs) and presented to the listener via headphones. The obtained signals were passed through a gammatone filterbank with an analysis and synthesis stage which enabled the modification of the ITD fluctuation statistics in individual frequency bands. The ITD fluctuations of broadband noise stimuli were compressed while the effect of this compression on the ILD statistics was kept minimal. The IACC was kept the same for stimuli with compression below 2kHz and for the uncompressed noise which should lead to the same ASW percept in the two conditions. However, the psychoacoustic data showed a reduced ASW for the modified signals, particularly in conditions with an applied compression around 1 kHz. In contrast, above 2kHz, the compression had no effect on ASW, whereas the IACC increased. The results suggest that the broadband IACC can be a misleading objective measure of ASW and that ITD fluctuations around 1kHz are crucial for ASW perception.

The importance of binaural cues for the perception of apparent source width at different sound pressure levels

The importance of binaural cues for the perception of apparent source width at different sound pressure levels

The effect of interaural-time-difference fluctuations on apparent source width
For the perception of spaciousness, the temporal fluctuations of the interaural time differences (ITDs) and interaural level differences (ILDs) provide important binaural cues. One major characteristic of spatial perception is apparent source width (ASW), which describes the perceived width of a sound image. The temporal fluctuations of the binaural cues cause the signals at a listeners’ ears to be decorrelated. Therefore, ASW has traditionally been measured by using the interaural cross- correlation (IACC). In particular, ITD fluctuations (below 2kHz) have been suggested to be the dominant cue for the perception of ASW. However, the contribution of the ITD statistics on the percept of ASW has not yet been clarified. In the present study, the impact of ITD fluctuations in different frequency bands on the perceived ASW was investigated. In a psychoacoustic evaluation, a source signal was convolved with individual binaural room impulse responses (BRIRs) and presented to the listener via headphones. The obtained signals were passed through a gammatone filterbank with an analysis and synthesis stage which enabled the modification of the ITD fluctuation statistics in individual frequency bands. The ITD fluctuations of broadband noise stimuli were compressed while the effect of this compression on the ILD statistics was kept minimal. The IACC was kept the same for stimuli with compression below 2kHz and for the uncompressed noise which should lead to the same ASW percept in the two conditions. However, the psychoacoustic data showed a reduced ASW for the modified signals, particularly in conditions with an applied compression around 1 kHz. In contrast, above 2kHz, the compression had no effect on ASW, whereas the IACC increased. The results suggest that the broadband IACC can be a misleading objective measure of ASW and that ITD fluctuations around 1kHz are crucial for ASW perception.
Environment-aware ideal binary mask estimation using monaural cues
We present a monaural approach to speech segregation that estimates the ideal binary mask (IBM) by combining amplitude modulation spectrogram (AMS) features, pitch-based features and speech presence probability (SPP) features derived from noise statistics. To maintain a high mask estimation accuracy in the presence of various background noises, the system employs environment-specific segregation models and automatically selects the appropriate model for a given input signal. Furthermore, instead of classifying each timefrequency (T-F) unit independently, the a posteriori probabilities of speech and noise presence are evaluated by considering adjacent TF units. The proposed system achieves high classification accuracy.

Extracting Sound-Source-Distance Information from Binaural Signals
A novel method for the estimation of the distance of a sound source from binaural speech signals is proposed. The method relies on several statistical features extracted from such signals and their binaural cues. Firstly, the standard
deviation of the difference of the magnitude spectra of the left and right binaural signals is used as a feature for this method. In addition, an extended set of additional statistical features that can improve distance detection is extracted from an auditory front-end which models the peripheral processing of the human auditory system. The method incorporates the above features into two classification frameworks based on Gaussian mixture models and Support Vector Machines and the relative merits of those frameworks are evaluated. The proposed method achieves distance detection when tested in various acoustical environments and performs well in unknown environments. Its performance is also compared to an existing binaural distance detection method.

General information
State: Published
Organisations: Department of Electrical Engineering, Hearing Systems, University of Patras, University of Oldenburg
Contributors: Georganti, E., May, T., van de Par, S., Mourjopoulos, J.
Pages: 1727-1741
Publication date: 2013
Peer-reviewed: Yes

Publication information
Journal: IEEE Transactions on Audio, Speech and Language Processing
Volume: 21
Issue number: 8
ISSN (Print): 1558-7916
Ratings:
BFI (2019): BFI-level 2
Web of Science (2019): Indexed yes
BFI (2018): BFI-level 2
Web of Science (2018): Indexed yes
BFI (2017): BFI-level 2
Scopus rating (2017): CiteScore 4.4 SJR 0.867 SNIP 4.2
Web of Science (2017): Impact factor 2.95
Web of Science (2017): Indexed yes
Scopus rating (2016): CiteScore 3.5 SJR 0.762 SNIP 3.429
Web of Science (2016): Impact factor 2.491
Scopus rating (2015): CiteScore 2.4 SJR 1.247 SNIP 2.4
Web of Science (2015): Impact factor 1.225
Web of Science (2015): Indexed yes
Scopus rating (2014): CiteScore 4.58 SJR 1.468 SNIP 2.936
Web of Science (2014): Impact factor
Scopus rating (2013): CiteScore 4.03 SJR 1.021 SNIP 3.242
Web of Science (2013): Impact factor 2.625
ISI indexed (2013): ISI indexed yes
Web of Science (2013): Indexed yes
Scopus rating (2012): CiteScore 3.05 SJR 0.918 SNIP 2.672
Web of Science (2012): Impact factor 1.675
ISI indexed (2012): ISI indexed yes
Web of Science (2012): Indexed yes
Scopus rating (2011): CiteScore 2.56 SJR 0.892 SNIP 2.743
Web of Science (2011): Impact factor 1.498
ISI indexed (2011): ISI indexed yes
Web of Science (2011): Indexed yes
Scopus rating (2010): SJR 1.127 SNIP 2.502
Web of Science (2010): Impact factor 1.668
Scopus rating (2009): SJR 1.116 SNIP 2.798
Scopus rating (2008): SJR 0.931 SNIP 2.389
Scopus rating (2007): SJR 0.613 SNIP 2.745
Web of Science (2007): Indexed yes
Scopus rating (2006): SJR 1.256 SNIP 3.301
Web of Science (2006): Indexed yes
A Binaural Scene Analyzer for Joint Localization and Recognition of Speakers in the Presence of Interfering Noise Sources and Reverberation

In this study, we present a binaural scene analyzer that is able to simultaneously localize, detect and identify a known number of target speakers in the presence of spatially positioned noise sources and reverberation. In contrast to many other binaural cocktail party processors, the proposed system does not require a priori knowledge about the azimuth position of the target speakers. The proposed system consists of three main building blocks: binaural localization, speech source detection, and automatic speaker identification. First, a binaural front-end is used to robustly localize relevant sound source activity. Second, a speech detection module based on missing data classification is employed to determine whether detected sound source activity corresponds to a speaker or to an interfering noise source using a binary mask that is based on spatial evidence supplied by the binaural front-end. Third, a second missing data classifier is used to recognize the speaker identities of all detected speech sources. The proposed system is systematically evaluated in simulated adverse acoustic scenarios. Compared to state-of-the-art MFCC recognizers, the proposed model achieves significant speaker recognition accuracy improvements.

General information
State: Published
Organisations: University of Oldenburg, Philips Research
Contributors: May, T., van de Par, S., Kohlrausch, A.
Pages: 2016-2030
Publication date: 2012
Peer-reviewed: Yes

Publication information
Journal: IEEE Transactions on Audio, Speech and Language Processing
Volume: 20
Issue number: 7
ISSN (Print): 1558-7916
Ratings:
BFI (2019): BFI-level 2
Web of Science (2019): Indexed yes
BFI (2018): BFI-level 2
Web of Science (2018): Indexed yes
BFI (2017): BFI-level 2
Scopus rating (2017): CiteScore 4.4 SJR 0.867 SNIP 4.2
Web of Science (2017): Impact factor 2.95
Web of Science (2017): Indexed yes
Scopus rating (2016): CiteScore 3.5 SJR 0.762 SNIP 3.429
Web of Science (2016): Impact factor 2.491
Scopus rating (2015): CiteScore 2.4 SJR 1.247 SNIP 2.4
Web of Science (2015): Impact factor 1.225
Web of Science (2015): Indexed yes
Scopus rating (2014): CiteScore 4.58 SJR 1.468 SNIP 2.936
Web of Science (2014): Impact factor
Scopus rating (2013): CiteScore 4.03 SJR 1.021 SNIP 3.242
Web of Science (2013): Impact factor 2.625
Binaural Scene Analysis: Localization, Detection and Recognition of Speakers in Complex Acoustic Scenes

The human auditory system has the striking ability to robustly localize and recognize a specific target source in complex acoustic environments while ignoring interfering sources. Surprisingly, this remarkable capability, which is referred to as auditory scene analysis, is achieved by only analyzing the waveforms reaching the two ears. Computers, however, are presently not able to compete with the performance achieved by the human auditory system, even in the restricted paradigm of confronting a computer algorithm based on binaural signals with a highly constrained version of auditory scene analysis, such as localizing a sound source in a reverberant environment or recognizing a speaker in the presence of interfering noise. In particular, the problem of focusing on an individual speech source in the presence of competing speakers, termed the cocktail party problem, has been proven to be extremely challenging for computer algorithms. The primary objective of this thesis is the development of a binaural scene analyzer that is able to jointly localize, detect and recognize multiple speech sources in the presence of reverberation and interfering noise. The processing of the proposed system is divided into three main stages: localization stage, detection of speech sources, and recognition of speaker identities. The only information that is assumed to be known a priori is the number of target speech sources that are present in the acoustic mixture. Furthermore, the aim of this work is to reduce the performance gap between humans and machines by improving the performance of the individual building blocks of the binaural scene analyzer. First, a binaural front-end inspired by auditory processing is designed to robustly determine the azimuth of multiple, simultaneously active sound sources in the presence of reverberation. The localization model builds on the supervised learning of azimuth-dependent binaural cues, namely interaural time and level differences. Multi-conditional training is performed to incorporate the uncertainty of these binaural cues resulting from reverberation and the presence of interfering noise sources. Second, a speech detection module that exploits the distinct spectral characteristics of speech and noise signals is developed to automatically select azimuthal positions that are likely to correspond to speech sources. Due to the established link between the localization stage and the recognition stage, which is realized by the speech detection module, the proposed binaural scene analyzer is able to selectively focus on a predefined number of speech sources that are positioned at unknown spatial locations, while ignoring interfering noise sources emerging from other spatial directions. Third, the speaker identities of all detected speech sources are recognized in the final stage of the model. To reduce the impact of environmental noise on the speaker recognition performance,
Missing data classifier is combined with the adaptation of speaker models using a universal background model. This combination is particularly beneficial in nonstationary background noise.

**General information**
State: Published
Organisations: Technical University of Denmark, Eindhoven University of Technology
Contributors: May, T., Van de Par, S., Kohlrausch, A.
Number of pages: 195
Publication date: 2012

**Publication information**
ISBN (Print): 978-90-386-3230-8
Original language: English
Electronic versions: 735521.pdf

**Bibliographical note**
PhD thesis
Source: dtu
Source-ID: u::7383
Research output: Research › Ph.D. thesis – Annual report year: 2012

**Blind estimation of the number of speech source in reverberant multisource scenarios based on binaural signals**
In this paper we present a new approach for estimating the number of active speech sources in the presence of interfering noise sources and reverberation. First, a binaural front-end is used to detect the spatial positions of all active sound sources, resulting in a binary mask for each candidate position. Then, each candidate position is characterized by a set of features. In addition to exploiting the overall spectral shape, a new set of mask-based features is proposed which aims at characterizing the pattern of the estimated binary mask. The decision stage for detecting a speech source is based on a support vector machine (SVM) classifier. A systematic analysis shows that the proposed algorithm is able to blindly determine the number and the corresponding spatial positions of speech sources in multisource scenarios and generalizes well to unknown acoustic conditions.

**General information**
State: Published
Organisations: Department of Electrical Engineering, Hearing Systems, University of Oldenburg
Contributors: May, T., van de Par, S.
Number of pages: 4
Publication date: 2012

**Host publication information**
Title of host publication: Proceedings of IWAENC 2012
Keywords: Binaural processing, Binary mask, computational computational
Electronic versions:
May, Par - 2012 - International Workshop on Acoustic Signal Enhancement (IWAENC).pdf
Source: dtu
Source-ID: u::7297
Research output: Research - peer-review › Article in proceedings – Annual report year: 2012

**Noise-Robust Speaker Recognition Combining Missing Data Techniques and Universal Background Modeling**
Although the field of automatic speaker recognition (ASR) has been the subject of extensive research over the past decades, the lack of robustness against background noise has remained a major challenge. This paper describes a noise-robust speaker recognition system that combines missing data (MD) recognition with the adaptation of speaker models using a universal background model (UBM). For MD recognition, the identification of reliable and unreliable feature components is required. For this purpose, the signal-to-noise ratio (SNR) based mask estimation performance of various state-of-the-art noise estimation techniques and noise reduction schemes is compared. Speaker recognition experiments show that the usage of a UBM in combination with missing data recognition yields substantial improvements in recognition performance, especially in the presence of highly non-stationary background noise at low SNRs.

**General information**
State: Published
Organisations: University of Oldenburg, Philips Research
Contributors: May, T., van de Par, S., Kohlrausch, A.
Pages: 108-121
Publication date: 2012
Peer-reviewed: Yes
On the statistics of Binaural Room Transfer Functions
The well-known property of the spectral standard deviation of Room Transfer Functions (RTFs), that is, its convergence to 5.57 dB, is extended to reverberant Binaural Room Transfer Functions (BRTFs). The BRTFs are related to the anechoic Head Related Transfer Functions (HRTFs) and the corresponding RTFs. Consequently, the statistical properties of the RTFs and HRTFs can be systematically related to the statistical properties of the BRTFs. In this work, the standard deviation of BRTFs measured in different types of rooms, for various source/receiver distances and azimuth angles is computed. The derived values are compared to the ones obtained from the single channel RTFs measured at the same positions. Their relationship to the 5.57 dB value is discussed.

A Probabilistic Model for Robust Localization Based on a Binaural Auditory Front-End
Although extensive research has been done in the field of machine-based localization, the degrading effect of reverberation and the presence of multiple sources on localization performance has remained a major problem. Motivated by the ability of the human auditory system to robustly analyze complex acoustic scenes, the associated peripheral stage is used in this paper as a front-end to estimate the azimuth of sound sources based on binaural signals. One classical approach to localize an acoustic source in the horizontal plane is to estimate the interaural time difference (ITD) between both ears by searching for the maximum in the cross-correlation function. Apart from ITDs, the interaural level difference (ILD) can contribute to localization, especially at higher frequencies where the wavelength becomes smaller than the diameter of the head, leading to ambiguous ITD information. The interdependency of ITD and ILD on azimuth is a complex pattern that depends also on the room acoustics, and is therefore learned by azimuth-dependent Gaussian mixture models (GMMs). Multiconditional training is performed to take into account the variability of the binaural features which results from multiple sources and the effect of reverberation. The proposed localization model outperforms state-of-the-art localization techniques in simulated adverse acoustic conditions.
Binaural detection of speech sources in complex acoustic scenes

In this paper we present a novel system that is able to simultaneously localize and detect a predefined number of speech sources in complex acoustic scenes based on binaural signals. The system operates in two steps: First, the acoustic scene is analyzed by a binaural front-end that detects relevant sound source activity. Second, a speech detection module selects source positions from a set of candidate positions that are most likely speech. The proposed method is evaluated in simulated multi-source scenarios consisting of two speech sources, three interfering noise sources and reverberation. © 2011 IEEE.

General information
State: Published
Organisations: Eindhoven University of Technology, University of Oldenburg
Contributors: May, T., Par, S. V. D., Kohlrausch, A.
Publication date: 2011

Host publication information
Simultaneous localization and identification of speakers in noisy and reverberant environments
Whereas the human auditory system has remarkable capabilities to focus on a particular target source in complex multi-source scenarios, it has remained a challenging task to develop algorithms that are able to retrieve information about sound sources in a complex acoustic scene (e.g. to localize and identify active speech sources). A robust binaural scene recognizer will be presented that is able to simultaneously localize and classify a predefined number of target speech sources in the presence of reverberation and interfering noise. The model consists of three stages: localization stage, detection of speech sources, and recognition of speaker identities. First, a binaural front-end is used to localize relevant sound source activity. Based on this localization information, a binary mask is determined which identifies the activity of individual sound sources on a time-frequency (T-F) basis. The localization is based on the supervised learning of azimuth-dependent binaural features, namely interaural time and level differences (ITDs and ILDs). Secondly, a speech detection module determines whether the corresponding source type is speech or noise for all sound sources that have been found. For this purpose the estimated binary mask and the corresponding spectral features are passed to a missing data classifier for each sound source candidate. Finally, the speaker identity of all detected speech sources is recognized. The proposed system is analyzed in simulated, adverse conditions including interfering noise, reverberation and the presence of multiple target sources. Compared to a state-of-the art MFCC recognizer, the proposed model achieves significant speaker recognition accuracy improvements.

General information
State: Published
Organisations: Unknown
Contributors: May, T., Van De Par, S., Kohlrausch, A.
Pages: 2121-2126
Publication date: 2011

Host publication information
Title of host publication: Proceedings of Forum Acusticum
(Proceedings of Forum Acusticum).
Source: dtu
Source-ID: n:oai:DTIC-ART:compendex/370501920::27482
Research output: Research - peer-review › Article in proceedings – Annual report year: 2011

Speaker Distance Detection Using a Single Microphone
A method to detect the distance of a speaker from a single microphone in a room environment is proposed. Several features, related to statistical parameters of speech source excitation signals, are introduced and are shown to depend on the distance between source and receiver. Those features are used to train a pattern recognizer for distance detection. The method is tested using a database of speech recordings in four rooms with different acoustical properties. Performance is shown to be independent of the signal gain and level, but depends on the reverberation time and the characteristics of the room. Overall, the system performs well especially for close distances and for rooms with low reverberation time and it appears to be robust to small distance mismatches. Finally, a listening test is conducted in order to compare the results of the proposed method to the performance of human listeners.

General information
State: Published
Organisations: University of Patras, University of Oldenburg, Philips Research
Contributors: Georganti, E., May, T., van de Par, S., Harma, A., Mourjopoulos, J.
Pages: 1949-1961
Publication date: 2011
Peer-reviewed: Yes

Publication Information
Journal: I E E E Transactions on Audio, Speech and Language Processing
Volume: 19
Issue number: 7
ISSN (Print): 1558-7916
Single-channel sound source distance estimation based on statistical and source-specific features

In this paper we study the problem of estimating the distance of a sound source from a single microphone recording in a room environment. The room effect cannot be separated from the problem without making assumptions about the properties of the source signal. Therefore, it is necessary to develop methods of distance estimation separately for different types of source signals. In this paper, we focus on speech signals. The proposed solution is to compute a number of statistical and source specific features from the speech signal and to use pattern recognition techniques to develop a
robust distance estimator for speech signals. Experiments with a database of real speech recordings showed that the proposed model is capable of estimating source distance with acceptable performance for applications such as ambient telephony.

**General information**
State: Published
Organisations: Eindhoven University of Technology, Philips Research, University of Patras
Contributors: Georganti, E., May, T., van de Par, S., Hämä, A., Mourjopoulos, J.
Number of pages: 12
Publication date: 2009

**Host publication information**
Title of host publication: Proceedings of the 126th AES convention
ISBN (Print): 9781615671663
Research output: Research - peer-review › Article in proceedings – Annual report year: 2009

**The effect of spectro-temporal integration in a probabilistic model for robust acoustic localization**
A robust acoustic localization model will be presented, which is based on the supervised learning of azimuth-dependent binaural feature maps consisting of interaural time differences (ITD) and interaural level differences (ILD). Motivated by the robust localization performance of the human auditory system, the associated peripheral stage is used in this study as a front-end for binaural cue extraction. Multi-conditional training is performed to take into account the variability of the binaural features which results from the combination of multiple sources, the effect of reverberation and changes in the source/receiver configuration. One way of accumulating evidence of possible sound source locations is to combine information across auditory channels. Alternatively, integrating evidence across groups of time-frequency (T-F) units, so-called fragments, which are believed to belong to a single source, was reported to significantly improve ITD-based localization performance [Christensen et al., Proc. of Interspeech, 2769-2772 (2007)]. Instead of accumulating the localization cue directly, the proposed model combines likelihoods, taking into account the uncertainty which is associated with the azimuth estimate of a particular T-F unit. Various procedures of controlling the spectro-temporal integration will be discussed and the influence on sound source localization will be presented.

**General information**
State: Published
Organisations: Technical University of Denmark, Philips Research
Contributors: May, T., van de Par, S., Kohlrausch, A.
Pages: 126-134
Publication date: 2009

**Host publication information**
Title of host publication: Proceedings of ISAAR 2009 : Binaural Processing and Spatial Hearing.
ISBN (Print): 87-990013-2-2
Electronic versions:
May, van de Par and Kohlrausch - ISAAR - 2010.pdf
Source: dtu
Source-ID: u::7413
Research output: Research - peer-review › Article in proceedings – Annual report year: 2009

**Constant complexity reverberation for any reverberation time**
A new artificial reverberation system is proposed which is based on perceptually relevant components in reverberated audio and as such allows for a very efficient implementation. The system first separates the signal into transient and steady-state components. The transient signal is reverberated by using an efficient time-varying recursive filter while the steady-state signal is processed separately with an all-pass filter. In contrast to common reverberation systems, the complexity of the recursive filter is determined solely by the duration of the transients and is therefore independent of the reverberation time.

**General information**
State: Published
Organisations: Philips Lighting
Contributors: May, T., Schobben, D.
Pages: 1163-1172
Publication date: 2007

**Host publication information**
Title of host publication: 122nd Convention Audio Engineering Society Convention
Publisher: Audio Engineering Society
Article number: 7040
Distant teaching of chamber music via local area networks
In this paper we present a study on teaching chamber music via internet. The application for this setup is for a highly reputed teacher to teach professional musicians at a very high level. Usually, all participants would have to fly from all over the world in order to work together. Therefore, it would be of great value, if these teaching lessons could be done via internet. Several audio and video devices and different audio setups have been tested. The results indicate that MPEG 2 broadcast devices with two microphones are suitable for this task.

General information
State: Published
Organisations: Department of Electrical Engineering, Hearing Systems, University of Oldenburg, Technical University of Braunschweig, L3S Research Center
Contributors: Bitzer, J., May, T., Kurtisi, Z., Loesch, T.
Pages: 259-266
Publication date: 2006

Host publication information
Title of host publication: 120th Convention Audio Engineering Society Convention
Publisher: Audio Engineering Society
Article number: 6709
Keywords: Modeling and Simulation, Acoustics and Ultrasonics, Audio and video, Chamber music, Two microphones, Motion Picture Experts Group standards, Internet
Source: FindIt
Source-ID: 250231009
Research output: Research - peer-review › Article in proceedings – Annual report year: 2006

Projects:

Characterizing neutral mechanisms of attention-driven speech processing
Fuglsang, S., PhD Student, Hearing Systems
Dau, T., Main Supervisor, Department of Electrical Engineering
Hjortkjær, J., Supervisor, Department of Electrical Engineering
May, T., Supervisor, Department of Electrical Engineering
Andersen, T., Examiner, Department of Applied Mathematics and Computer Science
De Vos, M., Examiner
Shamma, S., Examiner
Anden EU-finansiering
15/05/2015 → 13/11/2018
Award relations: Characterizing neutral mechanisms of attention-driven speech processing
Project: PhD

Characterizing neural mechanisms of attention-driven speech processing
Fuglsang, S., PhD Student, Department of Electrical Engineering
Dau, T., Main Supervisor, Department of Electrical Engineering
Hjortkjær, J., Supervisor, Department of Electrical Engineering
May, T., Supervisor, Department of Electrical Engineering
Anden EU-finansiering
15/05/2015 → 14/05/2018
Award relations: Characterizing neural mechanisms of attention-driven speech processing
Project: PhD

Assessing hearing-aid signal processing based on variations of the Turing test
Kowalewski, B., PhD Student, Department of Health Technology
Dau, T., Main Supervisor, Department of Electrical Engineering
Fereczkowski, M., Supervisor, Department of Electrical Engineering
MacDonald, E., Supervisor, Department of Electrical Engineering
Computational speech segregation inspired by principles of auditory processing
Bentsen, T., PhD Student, Department of Electrical Engineering
Dau, T., Main Supervisor, Department of Electrical Engineering
Kressner, A. A., Supervisor, Department of Health Technology
May, T., Supervisor, Department of Electrical Engineering
Marozeau, J., Examiner, Department of Health Technology
Tchorz, J., Examiner
Bleeck, S., Examiner
Samfinansieret - Andet
15/06/2014 → 07/03/2018
Award relations: Computational speech segregation inspired by principles of auditory processing
Project: PhD

Correlations between physical and perceptual parameters of acoustic scenarios. Implications for auditory modelling and sound field design
Käsbach, J., PhD Student, Department of Electrical Engineering
Dau, T., Main Supervisor, Department of Electrical Engineering
Epp, B., Supervisor, Department of Electrical Engineering
May, T., Supervisor, Department of Health Technology
MacDonald, E., Examiner, Department of Electrical Engineering
Tchorz, J., Examiner
Whitmer, W. M., Examiner
Technical University of Denmark
01/12/2011 → 02/11/2016
Award relations: Correlations between physical and perceptual parameters of acoustic scenarios. Implications for auditory modelling and sound field design
Project: PhD

Modeling perceptual externalization in the normal, impaired and aided-impaired auditory system
Hassager, H. G., PhD Student, Department of Electrical Engineering
Dau, T., Main Supervisor, Department of Electrical Engineering
May, T., Supervisor, Department of Electrical Engineering
Udesen, J., Supervisor, Department of Electrical Engineering
MacDonald, E., Examiner, Department of Electrical Engineering
Majdak, P. S., Examiner
Hohmann, V., Examiner
Technical University of Denmark
01/03/2013 → 10/05/2017
Award relations: Modeling perceptual externalization in the normal, impaired and aided-impaired auditory system
Project: PhD