Binaural detection of speech sources in complex acoustic scenes

May, Tobias; Par, Steven Van De; Kohlrausch, Armin

Published in:
2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics

Link to article, DOI:
10.1109/ASPAA.2011.6082276

Publication date:
2011

Document Version
Publisher's PDF, also known as Version of record

Citation (APA):
May, T., Par, S. V. D., & Kohlrausch, A. (2011). Binaural detection of speech sources in complex acoustic scenes. In *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* IEEE. https://doi.org/10.1109/ASPAA.2011.6082276

BINAURAL DETECTION OF SPEECH SOURCES IN COMPLEX ACOUSTIC SCENES

Tobias May[*] *and Steven van de Par*

University of Oldenburg
26111 Oldenburg, Germany
{tobias.may, steven.van.de.par}@uni-oldenburg.de

Armin Kohlrausch

Eindhoven University of Technology and
Philips Research, Eindhoven, The Netherlands
armin.kohlrausch@philips.com

ABSTRACT

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.

Index Terms— binaural localization, speech detection, computational auditory scene analysis, missing data recognition

1. INTRODUCTION

One of the most striking facts about the human auditory system is the ability to focus on a target sound source in difficult acoustical environments and to recognize whether a source is speech or noise by only analyzing the waveforms reaching both ears [1]. So far this robust recognition can not be achieved with computational algorithms when facing multi-source scenarios with reverberation and interfering noise. Nevertheless, knowledge about the location of the target sources would be very useful for a wide range of applications such as hearing aids and teleconference systems, e.g. to steer a beamformer or to control processing parameters.

Recently, several algorithms have been proposed to localize multiple target sources with two microphones [2, 3, 4]. However, in these studies all active sources were assumed to be speech sources and none of the aforementioned methods are able to determine whether the localized source activity belongs to a speech source.

The current paper presents a method that is able to automatically detect a predefined number of N speech sources in the presence of reverberation and multiple noise sources. For this purpose a robust front-end for binaural localization [4] is combined with a speech detection module that determines whether detected sound source activity corresponds to a speaker or to interfering noise. The speech detection module builds on missing data (MD) classification [5] which is a promising approach to deal with multiple overlapping sources by only considering time-frequency (T-F) units that are believed to be dominated by the target source. A two-class Bayesian classifier is trained to discriminate between speech and noise sources. Furthermore, the ability of three different features to reflect the distinct characteristics of speech and noise is evaluated.

2. SYSTEM DESCRIPTION

The proposed system is shown in Fig. 1 and consists of two blocks, namely the localization stage ① and the speech detection module ②. In the following, the individual stages will be described in detail.

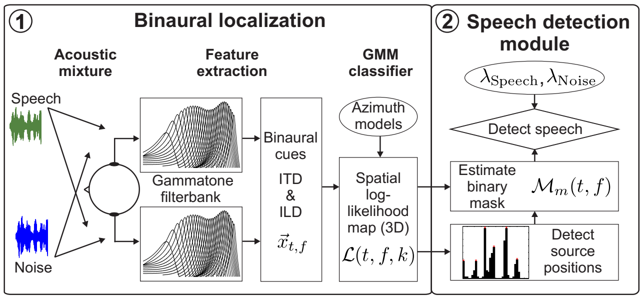

Figure 1: The block diagram of the binaural speech source detector.

2.1. Binaural analysis

The localization stage is based on an auditory front-end that simulates the processing of the human auditory system. The acoustic input to the proposed model is a binaural signal consisting of speech and noise sources that are positioned at unknown spatial locations (see Section 3.1 for more details). The binaural signal (sampled at a rate of 16 kHz) is first split into auditory channels using a bank of $Q = 32$ gammatone filters that cover the range of 80 to 5000 Hz. The processing of the inner hair cells is simulated by half-wave rectification and square-root compression, and the resulting response of the f-th auditory channel is denoted as h_f. Afterwards, interaural time (ITD, based on cross-correlation analysis) and level differences (ILD) are independently estimated for each auditory channel using overlapping frames of 20 ms with a 10 ms shift [4]. These two binaural cues are combined in a two-dimensional binaural feature space

$$\vec{x}_{t,f} = \{\hat{\text{itd}}_{t,f},\, \hat{\text{ild}}_{t,f}\}, \tag{1}$$

where t is the frame number and f indexes the gammatone channel. Based on these two binaural cues, the likelihood for each source location is determined by a Gaussian mixture model (GMM) classifier that has learned the azimuth-dependent distribution of ITDs and ILDs (cf. [4]): Multi-conditional training is performed to incorporate the uncertainty of ITDs and ILDs which results from multiple sound sources, changes in the source-receiver configuration, and the effect of reverberation. Given a set of K sound source directions $\{\varphi_1, \dots, \varphi_K\}$ that are modeled by a set of frequency-dependent GMMs $\{\lambda_{f,\varphi_1}, \dots, \lambda_{f,\varphi_K}\}$, a three-dimensional spatial log-likelihood map can be computed that represents the probability that the k-th sound source direction is active at frame t and frequency channel f:

$$\mathcal{L}(t,f,k) = \log p\left(\vec{x}_{t,f} | \lambda_{f,\varphi_k}\right), \tag{2}$$

where $p\left(\vec{x}_{t,f} | \lambda_{f,\varphi_k}\right)$ is a Gaussian mixture density with U weighted component densities. A number of $U = 15$ Gaussian components was used for all gammatone channels and azimuth directions. In this study, $K = 37$ sound source directions spaced by

[*] This project has been supported by Philips Research.

978-1-4577-0693-6/11/$26.00 ©2011 IEEE

5° within the range of $[-90^\circ, 90^\circ]$ are considered. The evidence about a sound source location is integrated across all Q gammatone channels, and the resulting azimuth estimate is used to determine the most probable sound source position for each frame:

$$\hat{P}(t) = \arg\max_{k} \sum\nolimits_{f=1}^{Q} \mathcal{L}(t, f, k). \qquad (3)$$

To obtain an estimate of all active sources, all frame-based azimuth estimates $\hat{P}(t)$ are pooled together to form an azimuth histogram $H[k]$. $H[k]$ represents the number of azimuth estimates that are assigned to the k-th sound source direction. Peaks within this azimuth histogram indicate relevant sound source activity and the corresponding histogram bin indices are used to form a set of $\mathcal{A}$ speech source candidate positions $L = \{\ell_1, \ldots, \ell_{\mathcal{A}}\}$. Each bin index ℓ_m corresponds to a local peak in the azimuth histogram.

2.2. Detection of speech sources

In the first stage, a histogram of frame-based azimuth estimates has been used to detect relevant sound source activity. Based on such a histogram, however, it is not possible to decide whether the detected activity corresponds to a speech source or to an interfering noise source. However, assuming that all sources are spatially separated, the spatial information can be used to determine and isolate the contribution of individual sound sources on a T-F basis. To achieve this, the spatial log-likelihood map $\mathcal{L}(t, f, k)$ is used to construct a binary mask $\mathcal{M}_m(t, f)$ by grouping T-F units according to common azimuth locations. For each T-F unit the most likely position among all $\mathcal{A}$ candidate positions is determined, and the individual T-F unit is added to the corresponding mask:

$$\mathcal{M}_m(t, f) = \begin{cases} 1 & \text{if } m = \arg\max\limits_{k \in L} \mathcal{L}(t, f, k) \\ 0 & \text{otherwise.} \end{cases} \qquad (4)$$

Based on such a mask, missing data classification [5] is performed to decide whether the corresponding source type is speech or noise. For classification two GMMs with 32 components and diagonal covariance matrices have been trained to approximate the probability distribution of the feature space $\mathcal{F}(t, f)$ that is extracted separately for speech and noise files. The first GMM, denoted as speech model λ_{Speech}, is trained with features based on a large pool of monaural speech files selected from the SSC database [6]. The second GMM, termed noise model λ_{Noise}, reflects the feature distribution of all types of noise files drawn from the NOISEX database [7]. About 29 minutes of training material was used for each GMM. A detailed description of the feature extraction process is given in Section 2.3.

Both the estimated binary mask $\mathcal{M}_m(t, f)$ and the extracted feature space $\mathcal{F}(t, f)$ are passed to the missing data recognizer. The log-likelihood ratio p_m reflects evidence for the m-th speech source candidate

$$p_m = \log\left(\frac{p(\mathcal{F}|\lambda_{\text{Speech}})}{p(\mathcal{F}|\lambda_{\text{Noise}})}\right). \qquad (5)$$

A speech source is detected if the likelihood for the speech model is larger than the likelihood for the noise model

$$p_m \begin{cases} \geq 0 & \text{accept } \lambda_{\text{Speech}} \\ < 0 & \text{reject } \lambda_{\text{Speech}}. \end{cases} \qquad (6)$$

Based on this criterion, all active sound sources are classified to be either speech or noise. After classification a set of $\hat{N}$ log-likelihood ratios $\{p_1^{\text{Speech}}, \ldots, p_{\hat{N}}^{\text{Speech}}\}$ is available that specifies the evidence of all detected speech sources. Based on this, a new set of histogram bin indices $L^{\text{Speech}} = \{\ell_1^{\text{Speech}}, \ldots, \ell_{\hat{N}}^{\text{Speech}}\}$ is available, which is a subset of L, and reflects the individual positions of all detected speech sources. Assuming that N represents the number of *a priori* known speech sources, two cases need to be considered:

- $\hat{N} \geq N$: Reflections and the interaction of multiple overlapping sources can cause the azimuth histogram to have numerous local peaks. As a consequence, the number of detected speech sources $\hat{N}$ might be larger than N.
- $\hat{N} < N$: In conditions where the signal-to-noise ratio (SNR) between the speech sources and the noise sources is close to zero, or even negative, the locations of the noise sources will dominate the azimuth histogram.

Assuming that $\hat{N} \geq N$, the N most likely speech sources need to be selected. We consider two selection strategies and both methods are evaluated in Section 4. The first approach is using the evidence from the missing data classifier to determine the most likely speech sources. Therefore, the set of log-likelihood ratios of all detected speech sources is rearranged in descending order

$$\{p_1^{\text{Speech}} \geq p_2^{\text{Speech}} \geq \cdots \geq p_{\hat{N}}^{\text{Speech}}\} \qquad (7)$$

and the azimuth locations corresponding to the highest N values are selected to represent the estimated speech source positions. This selection, however, does not account for the fact that speech sources that are more prominently represented in the azimuth histogram are more likely to reflect the real position of the speech sources. Therefore, the second method applies a weight to the log-likelihood ratio of each detected speech source reflecting the *a priori* probability that the corresponding source was active. This probability is approximated by the normalized azimuth histogram and subsequently used as a weight. The weighted log-likelihood ratio for the n-th speech source is given by

$$p_n^{\text{Speech,W}} = p_n^{\text{Speech}} + \underbrace{\log\left(H[\ell_n^{\text{Speech}}] / \sum\nolimits_k H[k]\right)}_{\text{azimuth weight}}. \qquad (8)$$

Similar to the first method, the resulting likelihood values are ranked in descending order, and the azimuth positions corresponding to the N highest values reflect the most likely speech sources.

Assuming that $\hat{N} < N$, the histogram of the frame-based azimuth estimates $\hat{P}(t)$ was potentially dominated by locations corresponding to noise sources, therefore the histogram did not reflect the locations of speech sources. Indeed, spectro-temporal regions that are dominated by speech tend to be sparse in the presence of noise [8]. Thus, whenever $\hat{N} < N$, the azimuth histogram $H[k]$ is recomputed using the azimuth estimates on a T-F basis

$$\hat{P}_{\text{TF}}(t, f) = \arg\max_{k} \mathcal{L}(t, f, k). \qquad (9)$$

The rationale behind (9) is that speech source positions that were not resolved on a frame-by-frame basis can potentially be recovered on a T-F level. Again, all peaks within this histogram are considered as speech source candidates, and the missing data masks corresponding to these locations are fed to a missing data classifier to determine the most likely speech source positions.

The proposed system is demonstrated in Fig. 2, where two speech sources are detected in the presence of three noise sources and reverberation ($T_{60} = 0.29\,\text{s}$). The estimated binary masks of all candidate positions are shown on the right hand side.

2.3. Feature extraction

In the context of speech recognition, missing data classification is usually performed with spectral features which reflect the energy of individual frequency channels [5]. As this study aims at discriminating between speech and noise, also, two alternative features are evaluated in the framework of missing data classification. Features

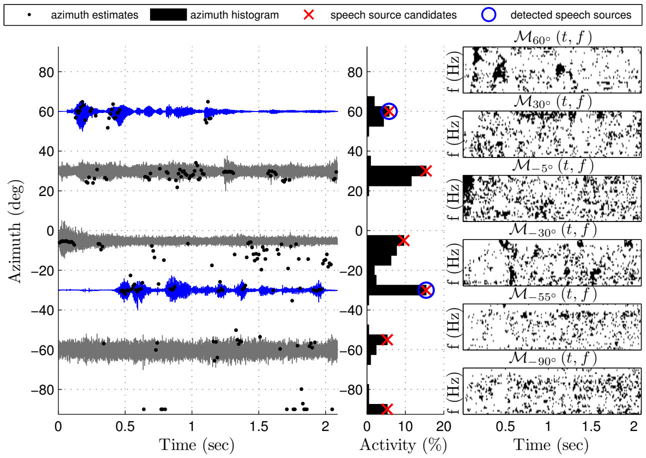

Figure 2: Detection of two speech sources (60° and $-30°$) in the presence of three factory noise sources (30°, $-5°$ and $-60°$) and reverberation ($T_{60} = 0.29$ s, SNR = 0 dBA).

are based on 20 ms frames using a shift of 10 ms. Note that signals of the left and the right ear are averaged prior to feature extraction.

First, a smoothed envelope e_f is obtained by low-pass filtering the half-wave rectified output of the f-th gammatone channel with a time constant of 10 ms. A map of auditory nerve firing rates, a so called *ratemap*, is computed by averaging the smoothed envelope e_f over B adjacent samples with a shift of R samples and subsequent cube-root compression

$$\mathcal{F}_1(t,f) = \left(\frac{1}{B}\sum_{b=0}^{B-1} e_f(tR+b)\right)^{1/3}. \qquad (10)$$

As an alternative, the mean absolute deviation of the envelope within a frame is used to reflect the amount of fluctuation

$$\mathcal{F}_2(t,f) = \frac{1}{B}\sum_{b=0}^{B-1} \left|e_f(tR+b) - \bar{e}_f\right|, \qquad (11)$$

where $\bar{e}_f$ refers to the mean envelope of the t-th frame. Similar to $\mathcal{F}_1$, the feature magnitude of $\mathcal{F}_2$ is lower for speech-dominant T-F units compared to units that are corrupted by noise. Thus, for feature $\mathcal{F}_1$ and $\mathcal{F}_2$ it is possible to use *bounded marginalization* [5] where the true value of the unreliable feature components is constrained to be between zero and the observed feature magnitude.

A third feature is based on the observation that speech-dominant areas typically show a higher amount of periodicity compared to noise [9]. The harmonic structure of speech tends to excite a similar auto-correlation pattern across neighboring frequency channels. This synchrony is expected to be reduced due to the influence of noise. Following [10], the synchrony is computed by correlating the normalized auto-correlation pattern $\hat{A}$ of the hair cell response h_f across neighboring frequency channels

$$\mathcal{F}_3(t,f) = \frac{1}{T}\sum_{\tau=0}^{T-1} \hat{A}(t,f,\tau)\,\hat{A}(t,f+1,\tau+1), \qquad (12)$$

where τ indexes the time lag and T refers to the maximum delay. Time lags corresponding to frequencies as low as 80 Hz are considered. Because no useful bounds can be defined for feature $\mathcal{F}_3$, *unbounded marginalization* [5] is performed for this feature.

Note that both GMMs λ_{Speech} and λ_{Noise} are trained with features based on monaural and anechoic signals. To compensate for the mismatch caused by reverberation, interfering noise and HRTF filtering, spectral normalization [11] is performed for feature $\mathcal{F}_1$ and $\mathcal{F}_2$. As the synchrony feature $\mathcal{F}_3$ is based on normalized auto-correlation patterns, its magnitude is limited to values between $[-1, 1]$. Therefore, no additional normalization is required.

3. EVALUATION SETUP

3.1. Acoustic mixtures

Acoustic sources were simulated by convolving monaural audio files with binaural room impulse responses (BRIRs). BRIRs were constructed by combining head related transfer functions (HRTFs) of a KEMAR artificial head taken from the MIT database [12] with room impulse responses (RIRs) that were simulated according to the image-source model [13]. The receiver (KEMAR) was placed at various positions in a simulated room of dimensions 6.6 x 8.6 x 3 m at 1.75 m above the ground. The localization model was trained with BRIRs corresponding to eight training positions (different from those in evaluation) using three different radial distances (0.5 , 1 and 2 m) between the source and the receiver. A reverberation time of $T_{60} = 0.5$ sec was used for all training positions. A different set of absorption coefficients was used for evaluation.

For evaluation the receiver was randomly placed at seven evaluation positions using a radial distance of 1.5 m. Speech and noise sources (different from the material used to train the speech detection module) were randomly positioned within the azimuth range of $[-90°, 90°]$, while having an angular distance of at least 15° to the nearest source. A set of 600, 4-source mixtures (one speech source) and 600, 5-source mixtures (two speech sources) are generated for each SNR condition. Mixtures had an average length of 1.83 s. The SNR was adjusted by comparing the broadband energy of all binaural speech sources with the energy of all binaural noise sources. The level between multiple speech or noise sources was always set equal. For a given multi-source mixture, performance is evaluated by comparing the positions of the detected speech sources to their real positions. A speech source is correctly detected only if the deviation from the real position is within an error margin of 5 °.

3.2. Baseline systems

To demonstrate the added value of the speech detection module, the first baseline system solely relies on the estimated azimuth information. Speech sources are classified by selecting the N most prominent peaks in the azimuth histogram. Obviously, with decreasing SNR, the locations of the noise sources will dominate the azimuth histogram, and, consequently, the speech sources will not be seen.

Secondly, a classifier based on 13 mel frequency cepstral coefficients (MFCCs) and their first-order dynamic coefficients with cepstral mean normalization is trained to discriminate between speech and noise. The classifier is trained with the same material that was used for the speech detection module. Based on a frame-by-frame decision, a modified azimuth histogram is computed by only using frames that were classified as being dominated by speech. Again, the N most dominant peaks reflect the detected speech sources.

4. EXPERIMENTAL RESULTS

First, the performance of the proposed system is evaluated using three different features. The speech detection accuracy of the MD-recognizer based on (7) for mixtures with one and two speech sources is shown in Tab. 1. The synchrony feature $\mathcal{F}_3$ showed the overall lowest performance. One possible explanation might be that the characteristic periodicity of speech is not only reduced by interfering noise but also due to the presence of multiple speech sources.

Table 1: Detection accuracy in % of one and two speech sources in the presence of reverberation ($T_{60} = 0.29\,\mathrm{s}$) and three interfering noise sources for different features using the MD recognizer.

1 speech source	SNR in dBA (factory noise)				
3 noise sources	−5	0	5	10	20
MD $\mathcal{F}_1$	68.86	86.21	94.14	97.57	98.86
MD $\mathcal{F}_2$	74.86	94.79	98.79	99.21	98.93
MD $\mathcal{F}_3$	40.07	63.5	76.21	82.5	94.21

2 speech sources	SNR in dBA (factory noise)				
3 noise sources	−5	0	5	10	20
MD $\mathcal{F}_1$	55.92	66.33	78.58	85	89.25
MD $\mathcal{F}_2$	56.5	69	82.75	88	87.42
MD $\mathcal{F}_3$	43	52.33	62.5	70.33	79.67

Since all classifiers are trained with single-source mixtures, it seems that the classifier trained with the synchrony feature is least capable of generalizing to more complex acoustic scenes. The performance of feature $\mathcal{F}_1$ and $\mathcal{F}_2$ is comparable at high SNRs. However, with decreasing SNR, feature $\mathcal{F}_2$ consistently outperformed the ratemap feature $\mathcal{F}_1$. This advantage is observed for both scenarios with one and two target speakers and might be related to the fact that the mean average deviation is invariant to the overall signal level. In the following the MD-based classifier is based on feature $\mathcal{F}_2$.

In the second experiment, the MD-based system is compared to two baseline systems. The accuracy of detecting two simultaneously active speech sources in the presence of three non-stationary factory noise sources and reverberation is shown in Fig. 3. As expected the first baseline system is only able to reflect the positions of both speech sources in conditions with high SNR, but performance rapidly drops with decreasing SNR. The MFCC-based selection of speech-dominant frames substantially improved the speech detection accuracy. However, the MD-based system is significantly more robust, especially at low SNRs and conditions with strong reverberation. Furthermore, the selection of the most likely speech sources based on the azimuth-weighted likelihood values using (8) provides a distinct performance gain for all experimental conditions.

5. CONCLUSION

In this paper, we proposed a novel method to robustly localize and detect a predefined number of speech sources in adverse acoustic conditions based on binaural signals. This is achieved by combining a front-end for binaural localization with a speech detection module that is using missing data classification to select source positions from a set of candidate positions that are most likely speech. In addition, it was shown that in the context of speech detection the feature based on the mean absolute deviation of the envelope is superior to the ratemap feature.

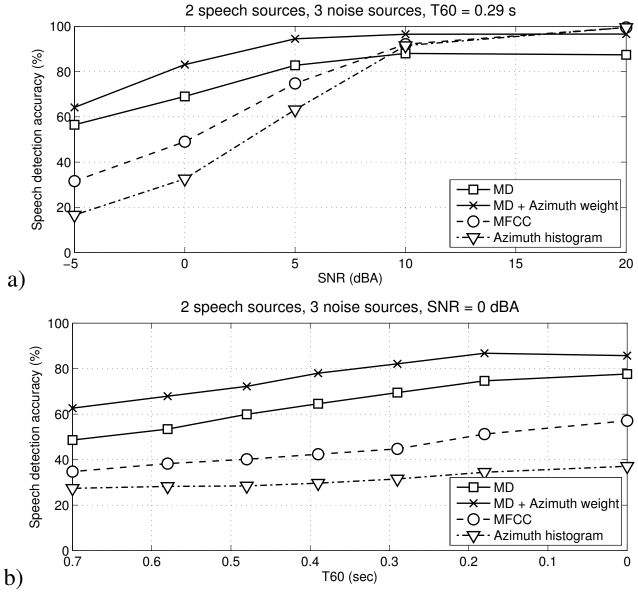

Figure 3: Detection accuracy in % of two speech sources in the presence of three factory noise sources and reverberation as a function of (a) the SNR and (b) the reverberation time T_{60}.

6. REFERENCES

[1] E. C. Cherry, “Some experiments on the recognition of speech, with one and two ears,” *J. Acoust. Soc. Amer.*, vol. 25, no. 5, pp. 975–979, 1953.

[2] C. Liu, B. C. Wheeler, W. D. O’Brien, R. C. Bilger, C. R. Lansing, and A. S. Feng, “Localization of multiple sound sources with two microphones,” *J. Acoust. Soc. Amer.*, vol. 108, no. 4, pp. 1888–1905, 2000.

[3] J. Woodruff and D. L. Wang, “Sequential organization of speech in reverberant environments by integrating monaural grouping and binaural localization,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1856–1866, 2010.

[4] T. May, S. van de Par, and A. Kohlrausch, “A probabilistic model for robust localization based on a binaural auditory front-end,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 1, pp. 1–13, 2011.

[5] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, “Robust automatic speech recognition with missing and unreliable acoustic data,” *Speech Commun.*, vol. 34, pp. 267–285, 2001.

[6] M. Cooke and T.-W. Lee, “Speech separation and recognition competition,” 2006. [Online]. Available: http://staffwww.dcs.shef.ac.uk/people/M.Cooke/SpeechSeparationChallenge.htm

[7] A. Varga, H. Steeneken, M. Tomlinson, and D. Jones, “The NOISEX-92 study on the effect of additive noise on automatic speaker recognition,” *Technical Report, Speech Research Unit, Defence Research Agency, Malvern, UK*, 1992.

[8] M. Cooke, “A glimpsing model of speech perception in noise,” *J. Acoust. Soc. Amer.*, vol. 199, no. 3, pp. 1562–1573, 2006.

[9] L. Atlas and L. Hengky, “Cross-channel correlation for the enhancement of noisy speech,” in *Proc. ICASSP*, 1985, pp. 724–727.

[10] Y. Shao, S. Srinivasan, Z. Jin, and D. Wang, “A computational auditory scene analysis system for speech segregation and robust speech recognition,” *Speech Commun.*, vol. 24, pp. 77–93, 2010.

[11] K. J. Palomäki, G. J. Brown, and J. P. Barker, “Techniques for handling convolutional distortion with ’missing data’ automatic speech recognition,” *Speech Commun.*, vol. 43, no. 1-2, pp. 123–142, 2004.

[12] W. Gardner and K. Martin, “HRTF measurements of a KEMAR dummy-head microphone,” *MIT Media Lab Perceptual Computing Technical Report* #280, 1994.

[13] S. M. Schimmel, M. F. Müller, and N. Dillier, “A fast and accurate ”shoebox” room acoustics simulator,” in *Proc. ICASSP*, pp. 241–244, 2009.