Blind estimation of the number of speech source in reverberant multisource scenarios based on binaural signals

May, Tobias; van de Par, Steven

Published in:
Proceedings of IWAENC 2012

Publication date:
2012

Citation (APA):
May, T., & van de Par, S. (2012). Blind estimation of the number of speech source in reverberant multisource scenarios based on binaural signals. In Proceedings of IWAENC 2012

BLIND ESTIMATION OF THE NUMBER OF SPEECH SOURCES IN REVERBERANT MULTISOURCE SCENARIOS BASED ON BINAURAL SIGNALS

Tobias May and Steven van de Par

University of Oldenburg, Institute of Physics, Acoustics Group, Oldenburg, Germany

{tobias.may, steven.van.de.par}@uni-oldenburg.de

ABSTRACT

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.

Index Terms— binaural processing, binary mask, computational auditory scene analysis (CASA)

1. INTRODUCTION

The automatic detection of target sources is required for a wide range of applications, among them self-steering hearing aids [1] and teleconferencing systems. In such applications it is important to detect only the speech sources and distinguish them from interfering noise.

If it were assumed that only target sources are acoustically active, which is a strong limitation for practical applications, it might be appropriate to cluster direction of arrival (DOA) estimates to make inferences about the number of target sources [2, 3, 4]. In more realistic scenarios, however, two target speakers might be involved in a conversation while noise sources and room reverberation interfere, making the problem much more difficult. Thus, evidence about the spatial activity of sound sources needs to be combined with a distinction between speech and noise sources.

Recently we have presented a speech detection module (SDM) which is able to select an *a priori* known number of speech sources from a set of candidate positions by exploiting the distinct spectral characteristics of speech and noise sources with a missing data (MD) classifier [5]. The current study extends the approach by removing the *a priori* knowledge about the number of target sources. To achieve a blind estimation of the number of active speech sources, each detected candidate source is characterized by a set of features. A decision stage based on an SVM classifier is employed to automatically select the candidate sources that are most likely speech. In addition, the present study investigates to what extent the pattern of an estimated binary mask is specific to speech sources and whether features describing the mask pattern itself can be exploited to further improve the detection of speech sources in adverse acoustic scenarios. Therefore, a new set of mask-based features is proposed which aims at capturing the specific patterns of reliable T-F units in the estimated binary mask of speech and noise sources.

2. SYSTEM DESCRIPTION

Given a binaural mixture, the proposed algorithm aims at determining the number and the spatial locations of all active sources and subsequently identifying the ones that correspond to speech sources. The system consists of three main stages, namely a binaural front-end for robust localization of active sound sources, a feature extraction stage, and an SVM-based decision stage. In the following the individual blocks are explained in detail.

2.1. Binaural front-end for robust localization

The localization stage is based on a binaural front-end for robust sound source localization [4]. The assumed input is a binaural signal sampled at 16 kHz. First, the acoustic signal is split into $Q = 32$ Gammatone filter channels with center frequencies equally spaced on the equivalent rectangular bandwidth (ERB) scale between 80 and 5000 Hz. Then, interaural time differences (ITDs) and interaural level differences (ILDs) are extracted in individual frequency channels by analyzing 20-ms frames (B samples) with a shift of 10 ms (O samples). Both interaural cues are combined in a two-dimensional (2D) binaural feature space $\vec{x}_{t,f} = \{\hat{\mathrm{itd}}_{t,f}, \hat{\mathrm{ild}}_{t,f}\}$, where t and f indicate time frames and Gammatone filter channels, respectively. To achieve robust localization for a set of $K = 37$ sound source directions $\{\varphi_1, \ldots, \varphi_K\}$ spaced by 5° within the range of $[-90^\circ, 90^\circ]$, the joint distribution of both ITDs and ILDs is approximated by a set of frequency- and azimuth-dependent diagonal Gaussian mixture models (GMMs) $\{\lambda_{f,\varphi_1}, \ldots, \lambda_{f,\varphi_K}\}$ with 15 Gaussian components [4]. Multi-conditional training is performed to incorporate the uncertainty of binaural cues resulting from multiple sound sources, changes in the source-receiver configuration, and reverberation [4]. Given the binaural feature vector $\vec{x}_{t,f}$, a three-dimensional spatial log-likelihood can be computed that the k-th source direction is active at frame t and frequency channel f:

$$\mathcal{L}(t, f, k) = \log p\left(\vec{x}_{t,f} | \lambda_{f,\varphi_k}\right). \tag{1}$$

To obtain a robust estimation of the spatial positions of all active sources, the log-likelihood about a source location is first accumulated across all frequency channels, and the most probable location is used to reflect the frame-based azimuth estimate:

$$\hat{P}(t) = \arg\max_{k} \sum_{f=1}^{Q} \mathcal{L}(t, f, k). \tag{2}$$

Then, all frame-based azimuth estimates $\hat{P}(t)$ are pooled together over the entire mixture to form an azimuth histogram H, where $H[k]$ represents the number of azimuth estimates that are assigned to the k-th sound source direction. Peaks in this histogram indicate relevant sound source activity, and the corresponding histogram bin indices are used to form a set of M candidate positions $L = \{\ell_1, \ldots, \ell_M\}$.

Furthermore, the spatial log-likelihood about sound source locations is used to estimate a binary mask $\mathcal{M}$ for each candidate position by determining the most dominant source direction among all candidate positions for each individual T-F unit:

$$\mathcal{M}_m\left(t,f\right) = \begin{cases} 1 & \text{if } m = \underset{k\in L}{\arg\max}\, \mathcal{L}\left(t,f,k\right) \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

2.2. Feature extraction

All detected candidate positions are characterized by a set of features $\mathbf{X}$. The first three features are based on the previously developed speech detection module [5]. Furthermore, a new set of mask-based features is proposed to describe the distribution of reliable T-F units in the estimated binary mask.

2.2.1. Speech detection module features

As proposed in [5], the estimated binary mask is used to exploit the distinct spectral characteristics of the speech and noise signal with a missing data (MD) classifier. Based on a smoothed envelope e_f obtained by low-pass filtering the half-wave rectified output of the f-th Gammatone channel with a time constant of 10 ms, the mean absolute deviation of the envelope over B time samples with a shift of O samples is calculated as $\mathcal{F}\left(t,f\right) = \frac{1}{B}\sum_{b=0}^{B-1}\left|e_f\left(t\,O+b\right) - \bar{e}_f\right|$, where $\bar{e}_f$ reflects the mean envelope of the t-th frame. Note that the left and right ear signals are averaged prior to envelope extraction. The distribution of this spectral feature $\mathcal{F}$ is approximated by two diagonal GMMs, namely λ_{Speech} and λ_{Noise}, for about 30 minutes of clean speech and noise files. Based on the estimated binary mask $\mathcal{M}_m$, the two GMM models λ_{Speech} and λ_{Noise}, and the spectral feature space $\mathcal{F}$, the first feature computes the log-likelihood ratio that the m-th candidate corresponds to a speech source:

$$\mathbf{X}_{m,1} = \log\left(\frac{p\left(\mathcal{F}|\lambda_{\text{Speech}}\right)}{p\left(\mathcal{F}|\lambda_{\text{Noise}}\right)}\right). \tag{4}$$

The second feature uses a normalized histogram of the frame-based azimuth estimates to approximate the probability that the m-th candidate was acoustically active over the entire mixture:

$$\mathbf{X}_{m,2} = H\left[\ell_m\right] / \sum_k H\left[k\right] \tag{5}$$

The third feature is a combination of the first two features $\mathbf{X}_{m,3} = \mathbf{X}_{m,1} + \log(\mathbf{X}_{m,2})$. This azimuth-weighted log-likelihood ratio performed best in ranking candidate sources according to their likelihood of being speech [5]. Therefore, $\mathbf{X}_{m,3}$ will be also used as a decision criterion for the first baseline system (see Section 3.2).

2.2.2. Mask-based features

By visual inspection it often seems possible to identify the mask pattern of a speech source when comparing it with mask patterns that correspond to noise sources. An illustration of typical mask patterns estimated by the binaural front-end is given in Fig. 1 where the masks in panels (a) and (b) depict noise sources and the mask in panel (c) corresponds to a speech source. Although spectro-temporal regions of speech-dominated T-F units are sparsely distributed in the presence of noise [6], they still tend to occupy contiguous groups of neighboring T-F units, resulting in coherent patches, so-called *fragments*. In contrast, the patterns of noise-dominated T-F units often appear to be more diffuse and less compact. This discernability of mask patterns related to speech and noise sources motivated the design of a new feature set, which aims at characterizing these distinct differences in the distribution of reliable T-F units between speech and noise sources. Furthermore, in the context of speech recognition, a recent study reported that mask patterns contain sufficient information to be used for isolated digit recognition [7].

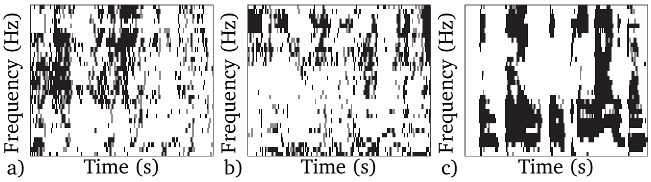


Fig. 1. Typical mask patterns corresponding to 3 competing sources: (a)-(b) factory noise sources located at $-85\,^\circ$ and $-45\,^\circ$, and (c) a speech source at $35\,^\circ$. Black pixels indicate reliable T-F units.

To describe the sparsity of a particular binary mask pattern $\mathcal{M}_m$, we first extract a set of t, f indices $\mathcal{R}_m = \{t_1, f_1, \ldots, t_R, f_R\}$ describing the position of all reliable T-F units. Furthermore, a set of fragments $\mathcal{P}_m = \{\mathcal{P}_m^1, \ldots, \mathcal{P}_m^P\}$ is created. A fragment itself is defined as a set of at least two t, f indices which refer to the position of reliable T-F units in the binary mask that are connected either horizontally or vertically (4-neighbors connectivity). Then, we extract the following fragment-related features: the number of fragments related to the total number of reliable T-F units $\mathbf{X}_{m,4} = |\mathcal{P}_m|/|\mathcal{R}_m|$, where $|.|$ refers to the cardinality of a set; the average fragment size $\mathbf{X}_{m,5} = \frac{1}{P}\sum_{p=1}^{P}|\mathcal{P}_m^p|$; and the percentage of reliable T-F units that are covered by fragments $\mathbf{X}_{m,6} = \frac{1}{|\mathcal{R}_m|}\sum_{p=1}^{P}|\mathcal{P}_m^p|$. Likewise, the percentage of reliable T-F units that are covered by fragments larger than 10 and 20 T-F units is computed, resulting in feature $\mathbf{X}_{m,7}$ and $\mathbf{X}_{m,8}$, respectively.

To measure the amount of randomness that is associated with the distribution of reliable T-F units, a 2D entropy filter is applied to the binary mask $\mathcal{M}_m$, resulting in a 2D entropy map E_m

$$E_m(t,f) = \text{entropy}\left\{\mathcal{M}_m\left(t',f'\right) : \left(t',f'\right) \in W_{(t,f)}\right\}, \tag{6}$$

where $W_{(t,f)}$ is chosen to be a plus sign-shaped neighborhood function [8] of dimensions 3 x 3 which is centered around t, f

$$\begin{aligned} W_{(t,f)} := \{(u,v) : \max\{|u-t|, |v-f|\} \leq 1 \;\wedge \\ \min\{|u-t|, |v-f|\} = 0\}. \end{aligned} \tag{7}$$

Note that the mask pattern $\mathcal{M}_m$ is symmetrically mirrored for T-F units that are located at the boarders of the mask. Afterwards, the mean of the entropy map E_m is calculated over the set of reliable T-F units

$$\mathbf{X}_{m,9} = \frac{1}{|\mathcal{R}_m|}\sum_{t,f\in\mathcal{R}_m} E_m(t,f). \tag{8}$$

We hypothesize that lower entropy values will be observed for speech-specific mask patterns where a consistent labeling of T-F units over a wider range of connected T-F units is expected. In contrast, higher values are conceivable for noisy mask patterns with a more sparse distribution of reliable T-F units as shown in Fig. 1.

Speech-dominated T-F units are often located in frequency regions of formants. Therefore, the centroid is used to describe the center of gravity of the m-th mask pattern

$$\mathbf{X}_{m,10} = \frac{1}{|\mathcal{R}_m|}\sum_{t,f\in\mathcal{R}_m} \mathcal{M}_m(t,f)\cdot c_f, \tag{9}$$

where c_f refers to the center frequency of the f-th Gammatone channel in Hz.

Table 1. Acoustic conditions for training and evaluation.

	Training	Testing
# speech sources $\mathcal{S}$	1, 2	0, 1, 2, 3
# noise sources $\mathcal{N}$	1	1, 2, 3
SNR (dBA)	$\{-\infty, 10, 20, \infty\}$	$\{0, 5, 10, 15, 20\}$
noise type	babble, factor 1 and factory 2	babble, factory 1, factory 2, destroyer engine, cockpit, car
T_{60} (s)	0.5	0, 0.29, 0.48
d_{rad} (m)	$0.5, 1.0, 2.0$	1.5

Speech signals are broadband and tend to simultaneously excite a number of neighboring frequency channels (see Fig. 1(c)), which is not generally true for noise. Therefore, onsets are detected by computing the first order difference of the binary mask pattern $\mathcal{M}_m$ across adjacent time frames. Then, the number of onsets are integrated across frequency channels and averaged over all frames to measure the average onset strength $\mathbf{X}_{m,11}$.

2.3. SVM classifier

The decision stage is based on an SVM classifier. A single SVM is constructed with a radial basis function using the LIBSVM toolbox [9]. The classifier is designed to distinguish between two classes, namely speech sources and background. The background class is used to summarize acoustic activity that is not related to speech sources but caused by interfering noise or reflections.

During training, a set of training files is processed by the binaural front-end. According to the number of local peaks in the azimuth histogram $H[k]$, the binary masks and the set of features are computed for all M candidate positions. To assign the extracted features to one of the two classes (speech or background), *a priori* knowledge about the spatial position of speech and noise sources is employed. First, the *a priori* known positions of all speech sources are compared to all candidate positions. Only if all speech sources in the training mixture are detected within an absolute error margin of $5\,^\circ$, the file is used for training, and the corresponding features are used to train the speech class. This prevents leakage of speech files that are not properly detected into the background class. Secondly, all remaining candidate sources are assigned to the background class.

The SVM parameters are optimized on the training set by 5-fold cross validation. Due to an unbalanced distribution of training samples for the two classes (20% speech and 80% background), we find that optimizing the hit rate minus false alarm rate leads to better results rather than optimizing the overall accuracy of the SVM.

3. EVALUATION

3.1. Acoustic mixtures

Binaural sound sources are created by convolving monaural speech and noise signals with binaural room impulse responses (BRIRs). Speech and noise files are randomly positioned within the azimuth range of $[-90^\circ, 90^\circ]$ while having an angular distance of at least $15\,^\circ$. The corresponding BRIRs are simulated according to the image-source model [10] where the receiver (KEMAR) was placed in a simulated room of dimensions 6.6 x 8.6 x $3\,$m. Speech files, randomly selected from the speech separation challenge (SSC) database [11], are mixed with various noise types from the NOISEX-92 database [12]. The signal-to-noise ratio (SNR) is 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.

Table 2. Accuracy in $\%$ of estimating the number of up to three active speech sources in the presence of interfering noise averaged over all six types of background noise.

T_{60} (s)	Method	SNR (dBA)				
		0	5	10	15	20
0	SDM	50.3	68.2	78.5	83.0	83.9
	SVM SDM	46.3	66.1	81.8	91.8	93.9
	SVM Prop	**61.3**	**81.0**	**92.9**	**96.2**	**96.0**
0.29	SDM	45.4	61.0	70.9	69.9	67.9
	SVM SDM	43.5	63.1	76.2	85.3	90.4
	SVM Prop	**54.0**	**72.0**	**87.5**	**89.7**	**92.0**
0.48	SDM	36.6	50.6	59.6	56.5	53.8
	SVM SDM	37.6	55.1	68.4	74.4	78.6
	SVM Prop	**44.2**	**61.0**	**76.6**	**83.0**	**88.8**

The binaural localization model and the SVM classifier are both trained with BRIRs corresponding to three radial distances (different from the one used for evaluation). Whereas the reverberation characteristic is intentionally simplified during training by using a frequency-independent reverberation time of $T_{60} = 0.5\,$s, a realistic frequency-dependent reverberation time with a low-pass characteristic is used for evaluation. Further, the SVM classifier is partly evaluated with noise types that are not used for training. This mismatch between training and testing conditions is incorporated to analyze to what extend the proposed approach is able to generalize to *unknown* acoustic conditions. An overview about the acoustic conditions during training and testing is given in Tab. 1. To train the SVM classifier, 50 binaural mixtures are created for each training condition, resulting in a total number of 3600 training files. For testing, 10 files are created for each testing condition, resulting in 10800 test files. Mixtures had an average length of $1.95\,$s. During testing, a binaural mixture is correctly classified if the number and the azimuth of all detected speech sources is correct (within $\pm 5\,^\circ$ of the true azimuth).

3.2. Algorithms

We consider three variations to estimate the number of active speech sources. The first system, denoted as *SDM*, is solely based on the azimuth-weighted log-likelihood ratio (feature $\mathbf{X}_{m,3}$) as proposed in [5]. A speech source is detected if $\mathbf{X}_{m,3} \geq \theta$, where the optimal decision threshold θ is found by maximizing the hit rate minus false alarm rate on the training set using 5-fold cross validation. The second system, referred to as *SVM SDM*, uses the first three features supplied by the speech detection module in combination with an SVM classifier. The proposed method *SVM Prop* uses the complete feature vector $\mathbf{X}$, which additionally includes the newly developed mask-based features.

4. RESULTS

The ability of all three methods to estimate the number and the spatial position of up to three competing speech sources in the presence of up to three interfering noise sources is shown in Tab. 2 depending on the SNR and the reverberation time T_{60}. The first baseline system *SDM* achieves the overall lowest performance and seems to be particularly affected by reverberation. By using the first three features

Table 3. Performance improvement in $\%$ of the proposed system *SVM Prop* over *SVM SDM* when incorporating mask-based features.

Types of noise	SNR (dBA)				
	0	5	10	15	20
babble, factory 1, factory 2	10.3	10.3	11.3	5.8	4.7
destroyer engine, cockpit, car	11.0	9.4	9.1	5.8	4.4
All noise types	10.7	9.9	10.2	5.8	4.6

in combination with an SVM classifier, a substantial improvement as high as 22% at higher SNRs is observed, which can be mainly attributed to the SVM classifier. The system *SVM Prop*, which additionally includes the mask-based features, achieves the highest performance and significantly outperforms the other two systems over all experimental conditions. Especially at lower SNRs, the additional use of mask-based features provides a performance gain of up to 15%. These results confirm that there are distinct differences in binary mask patterns of speech and noise sources and that the proposed mask-based features which exploit these differences can supply complementary information to the SVM classifier, allowing for an improved discrimination between speech and noise sources.

To analyze the influence of the background noise on the mask-based features, the improvement of the proposed system *SVM Prop* with mask-based features in comparison to *SVM SDM* is shown in Tab. 3 for known and unknown noise types averaged over the three T_{60}'s. The consistent benefit of the proposed system for both known and unknown types of background noise validates that the proposed system is able to generalize to unknown noise conditions.

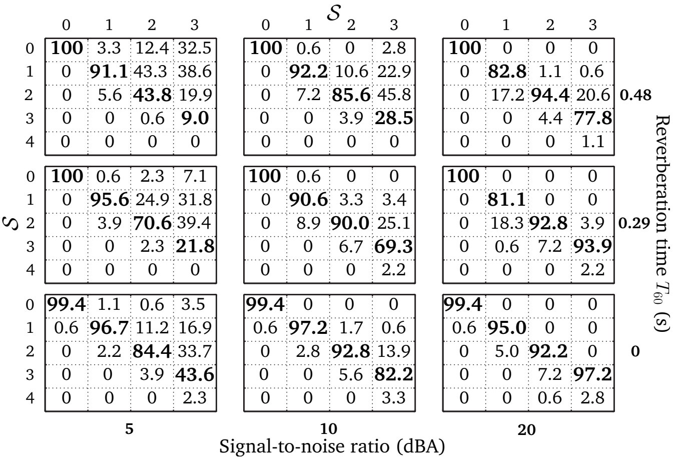
Finally, the confusion matrices of the proposed SVM system are presented in Fig. 2 averaged over all six types of background noise, where $\mathcal{S}$ represents the true number of speech sources and $\hat{\mathcal{S}}$ indicates the estimate. It's worth mentioning that although the system was only trained for acoustic scenarios with one and two speech sources in the presence of one interfering noise source, the system is fairly well able to generalize to scenarios with three competing speech sources for a wide range of acoustic conditions.

5. CONCLUSIONS

In this paper, we have presented a new method to automatically estimate the number and the spatial positions of active speech sources in reverberant multisource scenarios. The system first detects the spatial position of active sound sources and then extracts a set of features for each candidate position. An SVM-based decision stage distinguishes between speech and noise sources. A new set of mask-based features was introduced and substantially improved the discernibility between speech and noise sources in complex scenarios. Furthermore, experimental results indicate that the proposed system is able to generalize to unknown acoustic conditions, which corroborates the relevance of the presented approach for practical applications.

6. ACKNOWLEDGMENTS

The helpful discussions with J.-H. Bach and H. Kayser about SVM classifiers are gratefully acknowledged.



Reverberation time T_{60} (s)	$\hat{\mathcal{S}}$	SNR 5: $\mathcal{S}$ = 0	1	2	3	SNR 10: $\mathcal{S}$ = 0	1	2	3	SNR 20: $\mathcal{S}$ = 0	1	2	3
0.48	0	**100**	3.3	12.4	32.5	**100**	0.6	0	2.8	**100**	0	0	0
	1	0	**91.1**	43.3	38.6	0	**92.2**	10.6	22.9	0	**82.8**	1.1	0.6
	2	0	5.6	**43.8**	19.9	0	7.2	**85.6**	45.8	0	17.2	**94.4**	20.6
	3	0	0	0.6	**9.0**	0	0	3.9	**28.5**	0	0	4.4	**77.8**
	4	0	0	0	0	0	0	0	0	0	0	0	1.1
0.29	0	**100**	0.6	2.3	7.1	**100**	0.6	0	0	**100**	0	0	0
	1	0	**95.6**	24.9	31.8	0	**90.6**	3.3	3.4	0	**81.1**	0	0
	2	0	3.9	**70.6**	39.4	0	8.9	**90.0**	25.1	0	18.3	**92.8**	3.9
	3	0	0	2.3	**21.8**	0	0	6.7	**69.3**	0	0.6	7.2	**93.9**
	4	0	0	0	0	0	0	0	2.2	0	0	0	2.2
0	0	**99.4**	1.1	0.6	3.5	**99.4**	0	0	0	**99.4**	0	0	0
	1	0.6	**96.7**	11.2	16.9	0.6	**97.2**	1.7	0.6	0.6	**95.0**	0	0
	2	0	2.2	**84.4**	33.7	0	2.8	**92.8**	13.9	0	5.0	**92.2**	0
	3	0	0	3.9	**43.6**	0	0	5.6	**82.2**	0	0	7.2	**97.2**
	4	0	0	0	2.3	0	0	0	3.3	0	0	0.6	2.8

Fig. 2. Confusion matrices showing the accuracy in $\%$ of estimating the number of speech sources $\hat{\mathcal{S}}$ in the presence of 1, 2 and 3 noise sources as a function of the SNR and the reverberation time T_{60}.

7. REFERENCES

[1] T. Rohdenburg, S. Goetze, V. Hohmann, K.-D. Kammeyer, and B. Kollmeier, "Objective preceptual quality assessment for self-steering binaural hearing aid microphone arrays," in *Proc. ICASSP*, Las Vegas, Nevada, USA, 2008, pp. 2449–2452.

[2] B. Loesch and B. Yang, "Source number estimation and clustering for underdetermined blind source separation," in *Proc. IWAENC*, Seattle, WA, USA, 2008.

[3] N. Madhu and R. Martin, "A scalable framework for multiple speaker localization and tracking," in *Proc. IWAENC*, Seattle, WA, USA, 2008.

[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] ——, "Binaural detection of speech sources in complex acoustic scenes," in *Proc. WASPAA*, New Paltz, NY, USA, 2011, pp. 241–244.

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

[7] A. Narayanan and D. L. Wang, "Robust speech recognition from binary masks," *J. Acoust. Soc. Amer.*, vol. 128, no. 5, pp. EL217–EL222, 2010.

[8] M. Kühne, R. Togneri, and S. Nordholm, "Time-frequency masking: Linking blind source separation and robust speech recognition," in *Speech Recognition, Technologies and Applications*, F. Mihelič and J. Žibert, Eds. Vienna, Austria: I-Tech, 2008.

[9] C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," Software is available at www.csie.ntu.edu.tw/~cjlin/libsvm, 2001.

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

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

[12] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–251, 1993.