Data-driven mask generation for source separation

Nilesh Madhu

ExpORL, Dept. Neurosciences, Katholieke Universiteit Leuven, B-3000 Leuven, Belgium

Presented is a microphone-array based approach for the extraction of a target signal from a mixture of compteting sources and background noise. The approach builds upon a recent proposal for source localization and tracking in the general *M*-microphone *Q*-source case, and extends it to a versatile framework to perform source separation using data-driven soft– or hard–masks. The proposed approach is applicable to any arbitrary array – allowing for its integration into binaural hearing aids. The advantage of the proposed mask generation, in contrast to current algorithms, is the implicit scalability with respect to *M, Q*, source spread and the amount of reverberation – obviating the need for a heuristic adaptation of the mask generation algorithm in different acoustical scenarios. Further, the individual signals extracted using these soft-masks evince low amounts of musical noise. Additional mask smoothing may be performed to further reduce the musical noise phenomenon, thereby improving the listening experience.

INTRODUCTION

Separation of speech signals in day-to-day multi-speaker environments is a feasible task for humans with normal hearing, even under rather severe conditions such as many interfering speakers and background noise. For hearing-impaired people that dependent on hearing aids, this task is more difficult. To improve their lot, current research is focussed on the use of microphone arrays for extracting the sources of interest from the input medley of signals by means of *spatial* filtering.

Spatial filtering algorithms may be broadly divided into two categories: linear and non-linear. Algorithms belonging to the first category perform target source enhancement or interference[1] cancellation by a *linear* combination of the signals at the different sensors. When the interferers are highly directive, this amounts to steering a spatial *null* along their direction. The gain along the target direction may be additionally constrained, depending upon the combination philosophy being used. Such a process is known as *beamforming* and the resultant set of linear combination filters are known as *beamformers*. The interested reader is referred to van Trees (2002) for further details on beamforming philosophies and their realisations.

Alternatively, in the second category of approaches, interference *suppression*[2] may be obtained by damping those components of the mixture signals that do not belong to

1 Note that unless explicitly specified otherwise, we shall use the term interference and noise interchangeably to include the effect of both *directive* sources other than the target and background *noise*

2 Note that we speak of suppression of interference and not cancellation

Proceedings of ISAAR 2009: Binaural Processing and Spatial Hearing. 2nd International Symposium on Auditory and Audiological Research. August 2009, Elsinore, Denmark. Edited by J. M. Buchholz, T. Dau, J. Cristensen-Dalsgaard, and T. Poulsen. ISBN: 87-990013-2-2. The Danavox Jubilee Foundation, 2010.

the target source. Such an operation is usually performed on the short-time frequency representation of the signals, where only time-frequency (T-F) regions *dominated* by the target are preserved. We term such approaches as *masking*, and the corresponding filters as *masks*. Masking algorithms exploit the sparsity and disjointness of speech spectra (see next section and Rickard and Yilmaz (2002); Yilmaz *et al*. (2000)) for their function. Such approaches have been realized in a wide variety of ways such as the Wiener-filter approach of Bodden (1992) applied on the outputs of a gammatone filterbank, or the localisation based algorithms of Liu *et al*. (2001) and Roman *et al*. (2003), for example. An overview of the various mask-based approaches is given inWang (2008).

A combination of the above categories is also possible, where the masks are usually implemented as *post-filters* on the output of the beamforming algorithms. The methods proposed by Breithaupt *et al*. (2005) and Tashev and Acero (2006) are examples of this kind. Both advocate the use of a bank of fixed beamformers generated over a discrete localisation grid for the initial beamforming, followed by a single channel post-filter on the output of the beamformer for suppressing residual interference. However, while Breithaupt *et al*. (2005) uses a noise suppression filter based on classical single-channel algorithms for *stationary* noise floors, Tashev and Acero (2006) and Yoon *et al*. (2007) do the post-filtering using a T-F mask based on narrowband localisation information. In conjunction with the sparsity and disjointness properties, this approach is more suitable for non-stationary noise fields. The beamformers in the above cases are based on data-*independent* optimisation principles and assumptions on the underlying noise field statistics.

While these approaches indicate the right direction to proceed in, they suffer from a few drawbacks when it comes to the realisation of the post-filter. The approach of Breithaupt *et al*. (2005) will not work when the noise field is non-stationary, as is the case for a competing speaker scenario. The approaches of Tashev and Acero (2006) and Yoon *et al*. (2007) works better here, but the system parameters for mask realisation depend upon the acoustic scenario. In general, parameter tuning is a problem of most mask-based systems, especially in reverberant environments, and as such the system may need to be tuned for the particular combination of array, reverberation time, target source location, spread, position mismatch and interference location. As these tuning parameters arise from heuristic considerations, there is no closed-form solution for them, and they are usually set using arbitrary thresholds as in Bodden (1992), Tashev and Acero (2006), Yoon *et al*. (2007), etc.

While the approach we propose here is fundamentally similar to that of Tashev and Acero (2006), we show how the selection of a proper localisation model in the first stage allows for the generation of a wealth of parameters which may be used, in turn, to generate masks adaptive to target position mismatch, requiring no arbitrary, heuristic threshold setting for reverberant environments. Further, the method is self-scaling to M and Q. This approach may also be seen as a generalised version of the approach of Bodden (1992) and Tashev and Acero (2006).

SIGNAL MODEL

The proposed approach for target speaker separation is applied on the short-time Fourier transform (STFT) domain representation of the signals. Such a representation is obtained, e.g., by the discrete Fourier transform (DFT) on windowed and overlapped time samples of the signals. The advantage of the STFT representation is that it allows us to approximate the convolutive mixing of the signals due to the room impulse response by a simple multiplication in each frequency *bin*. This yields the signal model:

$$X_m(k,b) = \sum_q A_{mq}(k)\, S_q(k,b) + V_m(k,b)\,, \qquad \text{(Eq. 1)}$$

where k represents the frequency bin, b the time-frame under consideration, $A_{mq}(k)$ is the transfer function from the source q to the microphone m, $S_q(k, b)$ is the signal from the q th source, and $V_m(k, b)$ represents the *sum* of the diffuse and uncorrelated noise components at the mth microphone. We further assume that the transfer function contains principally the direct path from the source to the microphone and the late reverberation component is treated as stochastic in nature and uncorrelated with the direct path signal. It is subsequently subsumed into the definition of $V_m(k, b)$.

Another advantage to using the STFT representation is that despite the broadband nature of speech signals, they demonstrate considerable sparsity in the STFT domain. Further, if we consider the STFT representations of two speaker signals, we see that they are approximately disjoint. This means that if the corresponding spectra $S_1(k, b)$ and $S_2(k, b)$ are overlaid, there are very few T-F points (k, b) at which the spectra overlap. Mathematically this may be expressed as:

$$S_q(k,b)S_{q'}(k,b) \approx 0 \quad \forall q' \neq q\,. \qquad \text{(Eq. 2)}$$

The degree of disjointness depends upon the resolution of the DFT and has been examined in Yilmaz *et al*. (2004) for a sampling frequency of 16 kHz, in which case it is maximum for $K \in \{512, 1024, 2048\}$. Correspondingly, we fix our DFT resolution to lie in this range. Additionally, the degree of reverberation in the environment has a weak effect on disjointness, but in a manner similar to that due to background noise, and requires no special consideration.

Combining Eq. (1) and Eq. (2), we may write our final signal model as:

$$X_m(k,b) \approx A_{mq'}(k)\, S_{q'}(k,b) + V_m(k,b)\,, \qquad \text{(Eq. 3)}$$

i.e., any T-F point is dominated by a *single* active source.

SOURCE LOCALISATION MODEL

From (3) we see that if we perform localisation in each *bin k* at each time-frame *b*, we localise the *dominant* source at that T-F point, assuming that no spatial aliasing occurs. Considering, here, a linear array and localisation along the azimuth direction of arrival ($\theta \in [0,\pi]$), we have:

$$\underset{\theta}{\operatorname{argmin}}\, \mathcal{J}_\theta(k,b) = \widehat{\theta}_{q'}(k,b)\,, \qquad \text{(Eq. 4)}$$

where $J_\theta(k, b)$ is any generic cost-function parametrised by the source location. Given the disjointness of speech spectra, the narrowband estimates over the set K' of frequency bins where localisation is performed should yield enough information to localise all the *active* sources. This multi-source localisation is done by *clustering* the elements of the vector:

$$\widehat{\boldsymbol{\theta}}(b) = (\widehat{\theta}(1,b), \ldots, \widehat{\theta}(K',b))^T\,. \qquad \text{(Eq. 5)}$$

We further assume that we know the *approximate* location of the target source and allow for a *maximum* deviation of ± $\Delta\theta$ about this location. This value – $\Delta\theta$ – is a design parameter, and can be set independently of room reverberation and interference position.

The clustering may be done by several methods such as *k*-means (see, e.g., Faber (1994)) or parametrised mixture models (c.f. McLachlan and Peel (2007)). The disdvantage of approaches such as *k*-means is that they perform *hard* clustering, allocating each sample to only one centriod. However, when clustering data as in Eq. (5), such a hard clustering would yield biased estimates of the source locations – especially when the sources lie close together – as a particular data-point could belong to more than one cluster. Therefore we adopt the second method and model the $\theta(k, b)$ as realisations of a mixture of Gaussians (MoG) process:

$$\widehat{\theta}(k,b) \sim \sum_{i=1}^{\mathcal{I}} P_i \mathcal{N}\left(\theta_i, \sigma_i^2\right)\,. \qquad \text{(Eq. 6)}$$

In Eq. (6), the θ_i represent source locations, the σ_i^2 provide an indication of the spatial *spread* and the P_i describe the *a priori* probability that source i is present in frame b. I is framedependent and indicates the number of active sources localised in the particular time-frame under consideration. The model parameters are estimated by the expectation-maximization (EM) algorithm (c.f. Bilmes (1998)). An example of such a fitting is shown in Fig. 1.

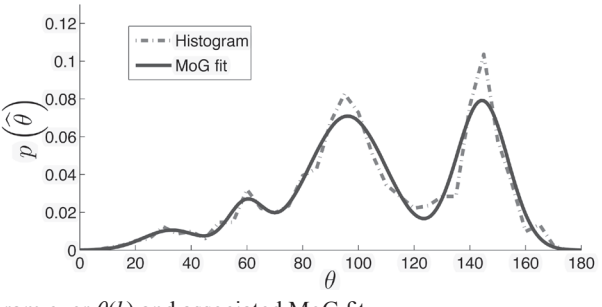


Fig. 1: Histogram over $\theta(b)$ and associated MoG fit.

MASK GENERATION

The model of Eq. (6) was used byMadhu andMartin (2008), along with a non-linear tracking and smoothing framework, as a means to localize multiple simultaneously-active sources. However, as can be seen, this richly-parametrised model yields a wealth of useful information, which can be used for source extraction using masks.

The masks are generated *independently* for each T-F point (*k, b*), using the estimated MoG model for frame *b*. The rationale behind mask generation is the computation of the *a posteriori* probability that a particular T-F point belongs to the target source, given the estimated MoG model for that frame and the $\theta(k, b)$ for the particular T-F point:

$$\mathcal{M}_{\mathrm{t}}(k,b) = P_{\mathrm{t}|\widehat{\theta}}(k,b). \qquad \text{(Eq. 7)}$$

Using Baye's rule, this can be expanded as:

$$\mathcal{M}_{\mathrm{t}}(k,b) = \frac{\frac{1}{\sigma_{\mathrm{t}}} P_{\mathrm{t}} \exp\left(-\frac{\left(\widehat{\theta}(k,b) - \widehat{\theta}_{\mathrm{t}}\right)^2}{2\sigma_{\mathrm{t}}^2} \right)}{\sum_{q'=1}^{\mathcal{I}} \frac{1}{\sigma_{q'}} P_{q'} \exp\left(-\frac{\left(\widehat{\theta}(k,b) - \theta_{q'}\right)^2}{2\sigma_{q'}^2} \right)}, \qquad \text{(Eq. 8)}$$

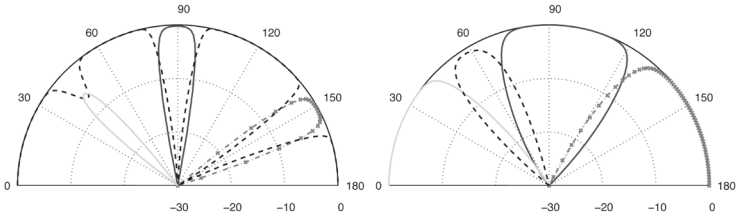
where we use the target position *estimate* from the MoG. This has the effect of increasing robustness against any errors in the knowledge of the target position. Once this mask has been computed, we can generate the target signal as

$$Y_{\mathrm{t,Msk}}(k,b) = \mathcal{M}_{\mathrm{t}}(k,b) Y_{\mathrm{t,DSB}}(k,b), \qquad \text{(Eq. 9)}$$

where $Y_{\mathrm{t,DSB}}$ is the delay-and-sum beamformed signal along θ_{t},

$$Y_{\mathrm{t,DSB}}(k,b) = \mathbf{H}^H(\theta_{\mathrm{t}}, k, b)\mathbf{X}(k,b) \qquad \text{(Eq. 10)}$$

with H(θ_{t}, *k, b*) being the filter corresponding to the delay-and-sumbeamformer (DSB) directed towards the *estimate* of the target source position at the T-F point (*k, b*). When the number of microphones available is large, using Eq. (10) yields better results due to the additional advantage of the linear combination before masking. While it is possible to use other beamforming philosophies, the delay-and-sum beamformer is simple to implement and has the advantage of being relatively robust to sensor gain mismatch and sensor noise, which factors compromise the performance of other beamformers.



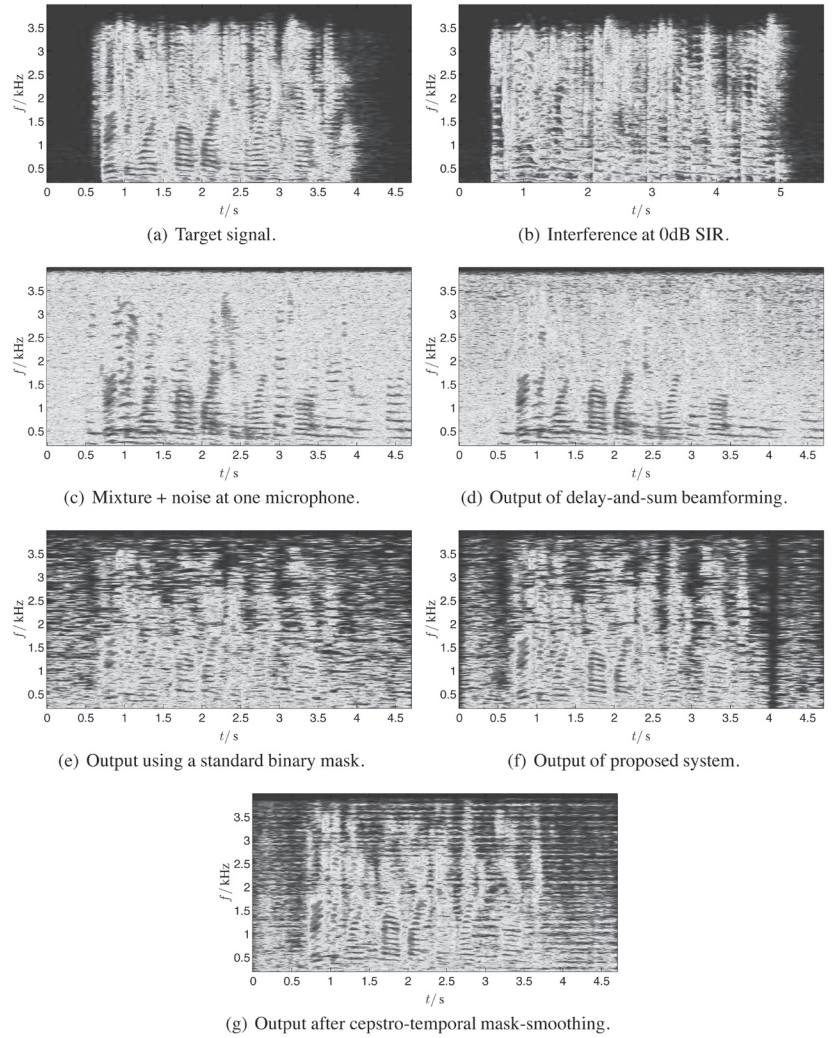
(a) Mask-generation in anechoic environents (b) Mask-generation in a highly reverberant environment ($T_{60} = 0.6$ s)

Fig. 2: Spatial effect of mask generated according to $M_t(k, b) = P_{t|b}(k, b)$. The target beam is directed towards broadside. The other beams indicate the division of the spatial region among the othercomponents. Also clearly visible is the beam towards the interferer (around $\theta = 150^\circ$). Note theadaptability of mask to target spread, reverberation and position mismatch.

An example of the effect of such masks is illustrated in Fig. 2. Note how the masks adjust to the amount of reverberation, becoming broader as reverberation increases, thus preserving the target signal, and to the changes in the source position (the target source is at broadside ($\theta = 90^\circ$) and a competing speaker is present at $\theta = 150^\circ$). The masks generated according to Eq. (8) take values in the range of [0, 1]. Due to this soft nature of the masks, the reconstructed signals show low amounts of musical tones. Further improvement in the listening experience (further reduction in musical noise) may be obtained by a temporal smoothing of the *cepstral* representations of the masks, as proposed by Madhu *et al*. (2008). The performance of the algorithm is illustrated on a sample, noisy mixture in Fig. 3. With respect to Fig. 3, the effect of the DSB is clearly visible in higher frequencies, where the noise is spatially uncorrelated. Here the DSB is the optimal beamforming solution. However, the competing speaker has strong directionality and, though damped, is still clearly audible. The target signal is undistorted. Using the binary mask the interference is suppressed to a large extent, but with a corresponding increase in target signal distortion. Further, the resulting spectrum is rich in artefacts, making for an uncomfortable listening experience. Using the proposed soft-mask, we obtain an equivalent amount of interference suppression, but see that the target is preserved. This also sounds more natural. There are still some artefacts, specifically gaps in the signal spectrum where the target is absent. Such gaps can be removed by the cepstro-temporal smoothing of the masks, yielding a resultant natural-sounding signal of a clearly-dominant speaker in a noisy environment.

EXPERIMENTAL RESULTS

This section presents an instrumental evaluation of the performance of proposed source enhancement algorithm. The performance measures computed are the long-term, intelligibilityweighted signal to interference-plus-noise *improvement* (ΔIW-SINR) and the log-spectral distortion ($\log_{SD}$). The intelligibility-weighted long-term SINR is directly related to the standard speech intelligibility index (SII) (c.f. ANSI-S3.5 (1997)).



(a) Target signal.

(b) Interference at 0dB SIR.

(c) Mixture + noise at one microphone.

(d) Output of delay-and-sum beamforming.

(e) Output using a standard binary mask.

(f) Output of proposed system.

(g) Output after cepstro-temporal mask-smoothing.

Fig. 3: Various masking algorithms for separation. Computed on data from all the channels of the array.

	5-mic. array				2-mic. array (d=15 cm)			
	DSB	Binary	Soft	Smoothed	DSB	Binary	Soft	Smoothed
ΔIW-SINR	2.77	5.60	4.09	2.82	1.12	4.14	2.75	1.76
ΔSINR	3.52	13.29	10.43	7.04	2.25	12.56	11.18	7.87
$\log_{SD}$	N/A	8.47	6.18	3.79	N/A	20.54	14.42	6.15

Table. 1: Experimental evaluation of the proposed approach. All values are in dB.

It should be noted that while a more complete test suite should incorporate subjective listening tests in order to obtain a better perspective on the algorithm performance – especially regarding the interference suppression and target distortion trade-offs involved – the instrumental measures selected are established physical measures that are related to important aspects of user-perceived signal quality and are capable of illustrating the extant trends in the algorithm performance. Therefore one may use these measures as a guideline for algorithm selection. The 'fine-tuning' of the algorithms however, would benefit from subjective listening tests.

Note also, that the achievable performance of multi-channel systems is limited by a number of factors such as the number of competing speakers, the spatial distance between them, the amount of background noise, the absolute position of the target source to be extracted etc. Consequently it is difficult to obtain a single figure-of-merit for any multi-channel separation algorithm that takes all these variables into account. In our evaluation, therefore, we shall constrain ourselves to the case where we have a single interfering talker at 150°, the target at broadside, and diffuse noise corrupting the microphone signals. The signal to interference ratio (SIR) is fixed to 0 dB and the signal to noise ratio is 10 dB. The target and interfering sources are drawn from a pool of 5 male and 5 female speakers from the TIMIT database, each uttering a single sentence. This gives us a total combination of $10 \cdot 9$ mixing scenarios, over which we shall average our performance measures. The recordings are made with a 5-microphone linear array, in a reverberant room with $T_{60} = 0.6$ s. The sources are at a distance of 1.0 m from the array, *outside* the critical distance for the room.

CONCLUSIONS

In this contribution, we have presented a mask-based approach for the extraction of a target speaker from the input mixture at a microphone array. We have utilised the property of disjointness of speech for performing the source extraction. The approach consists of applying a single-channel post-filter to the output of the DSB along the direction of the target. The postfilter gain at each T-F point is proportional to the probability of target source presence at that point. This information is obtained as a natural by-product of the localisation algorithm. Using such a soft-decision has the advantage of keeping the target signal sounding natural, whilst suppressing the unwanted T-F points. The approach is implicitly scalable with M, Q, errors in the target source position, room reverberation, etc., and thus provides robustness against imperfect knowledge, which is the case in practical situations. As expected, increasing the number of microphones in the array improves performance. Nevertheless, even in the simulated binaural case the performance is still good, and sounds better than the binary mask. A further advantage of this method for the binaural case is the preservation of the source cues, if the mask is applied to each hearing-aid output. Further development of this method is envisaged, based on recordings using an artifical head. Computational complexity is another issue which rightfully deserves consideration. However, this might be more a question of efficient programming.

REFERENCES

ANSI-S3.5 (**1997**), *American national standard methods for the calculation of the speech intelligibility index* (American National Standards Institute, New York).

Bilmes, J. A. (**1998**), "A gentle tutorial of the EM algorithm and its application to parameter estimation for gaussian mixture and hidden markov models," Tech. Rep. TR-97-021, U.C. Berkeley.

Bodden, M. (**1992**), "Binaurale Signalverarbeitung: Modellierung der Richtungserkennung und des Cocktail-Party-Effektes," Ph.D. thesis, Institute of Communication Acoustics, Ruhr-Universität Bochum.

Breithaupt, C., Madhu, N., Hummes, F., and Martin, R. (**2005**), "A robust steerable realtime multichannel noise reduction system," in *2005 Joint Workshop on Hands-Free Speech Communication and Microphone Arrays (HSCMA2005).*

Faber, V. (**1994**), "Clustering and the continuous *k*-means algorithm," URL http://www.fas.org/sgp/othergov/doe/lanl/pubs/00412967.pdf.

Liu, C.,Wheeler, B. C., O'Brien, Jr.,W. D., Bilger, R. C., Lansing, C. R., Jones, D. L., and Feng, A. S. (**2001**), "A two-microphone dual delay-line approach for extraction of a speech sound in the presence of multiple interferers," J. Acoust. Soc. Am. **110**(6), 3218–3231.

Madhu, N., Breithaupt, C., andMartin, R. (**2008**), "Temporal smoothing of spectral masks in the cepstral domain for speech separation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).*

Madhu, N. and Martin, R. (**2008**), "A scalable framework for multiple speaker localization and tracking," in *Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC)* (Seattle, USA).

McLachlan, G. and Peel, D. (**2007**), "Mixture models and neural networks for clustering," URL http://en.scientificcommons.org/43159010.

Rickard, S. and Yilmaz, ¨O. (**2002**), "On the approximate W-Disjoint orthogonality of speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).*

Roman, N., Wang, D., and Brown, G. (**2003**), "Speech segregation based on sound localization," J. Acoust. Soc. Am. **114,** (4), 2236 – 2252.

Tashev, I. and Acero, A. (**2006**), "Microphone array post-processing using instantaneous direction of arrival," in *Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC).*

van Trees, H. L. (**2002**), *Detection, Estimation and Modulation Theory, Part IV* (John Wiley and Sons).

Wang, D. (**2008**), "Time–frequency masking for speech separation and its potential for hearing aid design," Trends in amplification , 332–353.

Yilmaz, ¨O., Jourjine, A., and Rickard, S. (**2000**), "Blind separation of disjoint orthogonal signals: Demixing N sources from two mixtures," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).*

Yilmaz, ¨O., Jourjine, A., and Rickard, S. (**2004**), "Blind separation of speech mixtures via time-frequency masking," IEEE Transactions on Signal Processing **52**.

Yoon, B.-J., Tashev, I., and Acero, A. (**2007**), "Robust adaptive beamforming algorithm using instantaneous direction of arrival with enhanced noise suppression capability," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).*