UNDER-DETERMINED SPEECH SEPARATION USING GMM-BASED NON-LINEAR BEAMFORMING

Mohammad A. Dmour and Michael E. Davies

Institute for Digital Communications and Joint Research Institute for Signal and Image Processing,
University of Edinburgh, Edinburgh, EH9 3JL, UK
{M.Dmour, Mike.Davies}@ed.ac.uk

ABSTRACT

This paper introduces a frequency-domain non-linear beamformer that can perform speech source separation of under-determined mixtures, is reasonably artifact-free and does not require prior knowledge of the number of speakers. This beamformer utilises a Gaussian mixture distribution to model the observation probability density in each frequency bin, which can be learnt using the expectation maximisation (EM) algorithm. A linear minimum-variance distortionless response (MVDR) beamformer is determined for each of the Gaussian components. The proposed non-linear beamformer is then a weighted sum of these linear MVDR beamformers and is therefore also distortionless. The relative contribution for each linear MVDR beamformer is calculated as the posterior probability (specific to each time-frequency point) of its corresponding Gaussian component. Simulation results of the non-linear beamformer in under-determined mixtures with room reverberation confirm its ability to successfully separate speech sources with virtually no artifacts.

1. INTRODUCTION

Speech separation is the problem of extracting a target speech signal from observations corrupted by interfering signals such as other speech signals and background noise. Speech separation is used in a wide range of applications, such as hearing aids, human-computer interaction, surveillance, and hands-free telephony. In general, observations are obtained at the output of a set of microphones, each receiving different combinations of the source signals. The use of microphone arrays gives one the opportunity to exploit the fact that the desired source and the interfering sources originate at different points in space. The difficulty of the speech separation task depends on the way in which the signals are mixed within the acoustic environment. Speech separation is more difficult when the reverberation time of the acoustic environment is large, and when there are fewer microphones than sources.

Suppose that M source signals are mixed and observed at N microphones. The signal at microphone j can be modeled as:

$$x_j(t) = \sum_{i=1}^{M} \sum_{p=0}^{P-1} a_{ji}(p) s_i(t-p) \tag{1}$$

where a_{ji} represents the impulse response from source i to microphone j, and P is the length of the impulse response between each source-microphone pair. A mixture is termed a determined mixture when the number of microphones is equal to the number of sources, over-determined when the number of microphones is larger than the number of sources, and under-determined when it is smaller.

One approach to speech separation is to use statistical modeling of source signals. Independent component analysis (ICA) is one of the major statistical tools for solving the problem of speech separation. In ICA, separation is performed using the assumption that the source signals are statistically independent with no information on the direction of arrival of source signals, or microphone array configuration. To perform source separation, we process the mixture channels by a set of time-invariant demixing filters and sum the filtered channels together. ICA implicitly estimates the source directions by maximising the independence of the sources, and acts as an adaptive null beamformer that reduces the undesired sources.

However, some aspects limit the application of ICA in real-world environments. Most ICA methods assume the number of sources is given a priori. In general, classical ICA techniques cannot perform source separation when spatially spread sources are involved, or in the under-determined mixtures case.

Another approach to speech separation is to use adaptive beamforming techniques. In adaptive beamforming, the microphone array is used to form a spatial filter which can extract a signal from a specific direction and reduce signals from other directions. For example, in minimum-variance distortionless response (MVDR) beamforming, the beamformer response is constrained so that signals from the direction of interest are passed with no distortion, while it suppresses noise and interference at the output of an array of microphones. In [2, 3], beamforming weights were calculated using time-domain recursive algorithms. It was shown recently in [4] that a frequency-domain MVDR (FMV) beamformer which performs sample matrix inversion using statistics estimated from a short sample support gives better performance than time-domain recursive algorithms in non-stationary acoustic environments. Compared to ICA, adaptive beamforming can utilise the available information about source signals and the microphone array configuration. In addition, there is no need to model the source signals or determine their number. Adaptive beamforming can attain excellent separation performance in determined or over-determined time-invariant mixtures involving point sources. However, when spatially spread sources are involved, or in under-determined mixtures, perfect attenuation of all interferers becomes impossible and only partial interference attenuation is possible. This, in turn, leads to performance degradation.

In the under-determined mixing case, the assumption of spatial diversity is insufficient to perform source separation, thereby necessitating additional assumptions. One increasingly popular and very useful assumption is that the sources have a sparse representation in a given basis. The advantage of sparse signal representation is that the probability of more than one active source is low. A sparse representation of a speech signal can be achieved by a short term Fourier transform (STFT). One popular approach to perform under-determined speech separation is time-frequency (t-f) masking. This approach is a special case of non-linear time-varying filtering that estimates the desired source signal by:

$$\hat{s}(n,f) = M(n,f) x_j(n,f) \tag{2}$$

where $M(n,f)$ is a t-f mask containing positive gains which must be adapted to extract the desired source from the observed mixtures. A popular method used to perform speech separation of under-determined mixtures using only two microphones is the degenerate unmixing estimation technique (DUET) [8, 5]. In DUET, binary masks are determined from the spatial location information contained in the STFT coefficients of the mixture channels. DUET is capable of performing separation of two or more sources using just two channels, and without significant computational complexity. However, this method suffers from the so-called musical noise or burbling artifacts due to binary masking of t-f points where the sources overlap.

In this paper, we introduce a frequency-domain non-linear beamformer that can perform speech separation of under-determined mixtures and is distortionless. This beamformer utilises Gaussian mixture models (GMMs) to model the observation probability density in each frequency bin. This in turn can be learnt using the expectation maximisation (EM) algorithm. The signal estimator comprises of a set of MVDR beamformers, one for each component of the GMM. In order to estimate the signal, all beamformers are concurrently applied to the observed signal, and the weighted sum of the beamformers' outputs is used as the signal estimator, where the weights are the posterior probabilities of the GMM components. This approach results in a "soft decision" filter for the observed signal. The resulting non-linear beamformer has low computational costs, and does not need to know or estimate the number of sources. It combines the benefits of non-linear time-varying separation in t-f masking with the benefits of spatial filtering and distortionless response in the linear MVDR beamformer.

The organisation of this paper is as follows. Section 2 reviews the linear minimum mean square error (MMSE) beamformer, and then introduces the GMM-based non-linear beamformer. In Section 3, the EM algorithm is used to learn the GMM parameters. The experimental conditions and simulation results are presented in Section 4, followed by a discussion in Section 5.

2. OPTIMUM BEAMFORMERS

Consider a narrow band array signal $\mathbf{x} = [x_1, ..., x_N]^T$ that consists of the desired signal arriving at the array from a known direction, and an interference-plus-noise signal. That is,

$$\mathbf{x} = s\mathbf{e} + \mathbf{v} \tag{3}$$

where $\mathbf{e}$ is the known $N \times 1$ array response vector in the direction of the desired source signal (the array manifold), and $\mathbf{v}$ is the $N \times 1$ complex vector of interference-plus-noise snapshots. We assume that the signal and interference-plus-noise snapshots are uncorrelated. The interference has spatial correlation according to the angles of the contributing interferers. The ultimate goal is to combine the received signals in such as way that the interference-plus-noise signal is reduced while the desired signal is preserved.

2.1 Linear MMSE beamformer

We first consider the optimum estimator whose output is the MMSE estimate of the desired signal s in the presence of Gaussian interference and noise, assuming known desired signal direction. We assume that the desired source signal is a sample function from a zero-mean complex-valued Gaussian random process, $s \sim \mathbb{N}(0, \sigma_s^2)$. We also assume a zero-mean complex-valued Gaussian interference-plus-noise, $\mathbf{v} \sim \mathbb{N}(0, R_v)$. Additionally, it is assumed that the signal and interference-plus-noise snapshots are uncorrelated. Hence, $\mathbf{x} \sim \mathbb{N}(0, R_v + \sigma_s^2 \mathbf{e}\mathbf{e}^H)$, and $\mathbf{x}|s \sim \mathbb{N}(s\mathbf{e}, R_v)$, where $(.)^H$ denotes the Hermitian transpose operator. The MMSE estimate of the desired signal s is the mean of the a posteriori probability density of s given $\mathbf{x}$:

$$\hat{s}_{MMSE} = E\left[s|\mathbf{x}\right] = \int p(s|\mathbf{x}).s\, ds \tag{4}$$

This mean is referred to as the conditional mean. It can be shown that the conditional mean can be expressed as [6]:

$$E\left[s|\mathbf{x}\right] = \frac{\mathbf{e}^H R_v^{-1}\mathbf{x}}{\mathbf{e}^H R_v^{-1}\mathbf{e}} . \frac{\sigma_s^2}{\sigma_s^2 + \left(\mathbf{e}^H R_v^{-1}\mathbf{e}\right)^{-1}} \tag{5}$$

The first term is an MVDR spatial filter, which suppresses the interfering signals and noise without distorting the signal propagating along the desired source direction. The second term is a single-channel Wiener post-filter. We see that the MMSE estimator is just a shrinkage of the MVDR beamformer. Unfortunately, the MMSE beamformer depends explicitly on σ_s^2 which is typically unknown. Therefore, we cannot implement the MMSE beamformer in practice. However, we can obtain a beamformer that does not depend on σ_s^2 by assuming a distortionless response in the specified direction. The result is the MVDR beamformer. However, since we have a distortionless response, we cannot exploit the sparsity of the desired source signal. The MVDR beamforming process can be written as:

$$\begin{aligned} \hat{s} &= \mathbf{w}^H \mathbf{x} \\ &= \frac{\mathbf{e}^H R_v^{-1}}{\mathbf{e}^H R_v^{-1}\mathbf{e}} \mathbf{x} \end{aligned} \tag{6}$$

In practice, the desired signal may either be present all the time, or it is difficult to estimate its activity periods. As a result of this, the estimation of the signal-free interference-plus-noise covariance matrix R_v is not possible. It can be shown, however, that if there is no mismatch between the vector $\mathbf{e}$ used in the MVDR beamformer and the true array manifold, then the estimator which uses the observed signal covariance matrix R_x is identical to the estimator which uses the signal-free interference-plus-noise covariance matrix R_v [6].

In general, the conditional mean estimator is not linear. The MMSE estimator is linear if either the estimator is constrained to be linear or the signals are Gaussian. Speech sources are generally non-stationary and non-Gaussian. This suggests extending the optimum beamformers to exploit the non-stationarity and non-Gaussianity of speech signals.

2.2 Frequency-domain MVDR (FMV) beamformer

Speech is a non-stationary process, but over short durations speech signals can be considered stationary. In the FMV algorithm [4], it is assumed that source activity patterns are constant over small time intervals of speech signals in each frequency band, but could change over longer time spans. In the FMV algorithm [4], frequency-domain signals are stored in a buffer, and a correlation matrix is calculated for each frequency bin using the 32 most recent STFT values . MVDR weights are then calculated using the correlation matrix. Therefore, in the FMV algorithm, new beamformer weights are calculated every small time interval in order to reduce the contribution to the extracted signal of interfering sources active during that time interval, while having a distortionless response in the desired source DOA. Only statistics gathered over a very short period of time are used in the calculation of weights.

The quick adaptation of the beamformer weights can substantially reduce a large number of non-stationary interferences while utilising few microphones [4]. But the computational load is high due to recurrent matrix inversions in each frequency band and the need to have a very small step size in the STFT. In practice, however, source activity patterns can change abruptly between samples, and the FMV will perform spatial filtering based on the average power of the interfering sources active in the time interval during which the beamformer weights are calculated. On the other hand, the spatial distribution of the sources does not change very quickly, and we can gather statistics for the desired signal estimator over a longer time span. Thus the FMV beamformer is forced to compromise between long intervals (good statistics) and short interval (rapid response).

2.3 GMM-based non-linear beamformer

In the frequency-domain, speech signals have a super-Gaussian (sparse) distribution, due to a combination of the non-stationarity and harmonic content of speech. Therefore, even if sources might overlap at some t-f points, not all speech sources in a mixture are active at the same t-f points. It is therefore advantageous to exploit the sparsity property of speech signals in the frequency-domain in order to perform separation in under-determined environments. In order to model the speech non-Gaussianity, we propose to apply

GMMs, which are widely used for modeling highly complex probability densities.

In this section, we use a Gaussian mixture interference-plus-noise model and find the optimum estimator whose output is the MMSE estimate of the desired signal s assuming a known desired signal direction. We shall describe the density of the interference-plus-noise signal $\mathbf{v}$ as a mixture of k zero-mean Gaussians $q = 1,...,k$ with covariances $R_{v,q}$ and mixing proportions c_q:

$$p(\mathbf{v}|\theta) = \sum_{q=1}^{k} c_q \frac{1}{\pi^N |R_{v,q}|} \exp\{-\mathbf{v}^H R_{v,q}^{-1}\mathbf{v}\} \tag{7}$$

where $\theta = (c_1,...,c_k,R_{v,1},...,R_{v,k})$, and the mixing proportions c_q are constrained to sum to one. The number of components k controls the flexibility of the GMM. When dealing with mixture models, it is useful to consider that there exists a hidden random variable z, taking its values in a set $Z = [1,...,k]$ with probability $P(z = q) = c_q$, $1 \leq q \leq k$. Therefore we have $\mathbf{v}|z = q \sim \mathbb{N}(0,R_{v,q})$. The MMSE estimate of the desired signal s is the mean of the a posteriori probability density of s given $\mathbf{x}$:

$$\begin{aligned}
\hat{s}_{MMSE} &= E[s|\mathbf{x}] \\
&= \int p(s|\mathbf{x}).s\,ds \\
&= \int \sum_q p(s,z=q|\mathbf{x}).s\,ds \\
&= \int \sum_q p(s|z=q,\mathbf{x}).p(z=q|\mathbf{x}).s\,ds \\
&= \sum_q p(z=q|\mathbf{x}) \int p(s|z=q,\mathbf{x}).s\,ds \\
&= \sum_q \tau_q \int p(s|z=q,\mathbf{x}).s\,ds \\
&= \sum_q \tau_q E[s|\mathbf{x},q] \qquad (8)
\end{aligned}$$

where $\tau_q = p(z = q|\mathbf{x})$ is the a posteriori probability that the component q is active in the Gaussian mixture, when observing $\mathbf{x}$.

We can see that the conditional mean $E[s|\mathbf{x},q]$ is the MMSE beamformer estimator derived in the previous section, with $R_v = R_{v,q}$. In practice, modelling the signal-free interference-plus-noise signal $\mathbf{v}$ is not possible, and therefore we model the observed signal $\mathbf{x}$ instead. The desired signal estimator in equation (8) is a weighted sum of linear beamformers $\mathbf{w}_q$ over all the GMM components, and the weighted coefficients are the a posteriori probabilities of the GMM components τ_q. The mixture of beamformers (MOB) is given by:

$$\mathbf{w} = \sum_{q=1}^{k} p(z=q|\mathbf{x})\mathbf{w}_q \tag{9}$$

The resulting MOB is a weighted sum of distortionless MVDR beamformers, where the weights sum to unity, therefore it is distortionless in the look-direction.

3. MODEL LEARNING

Using the EM algorithm, we can estimate the observation model density parameters $\theta = (c_1,...,c_k,R_1,...,R_k)$ from a set of observations $D = \{\mathbf{x}(n) : n = 1,...,\eta\}$. The EM algorithm is used to find a ML estimate of parameters in probabilistic models with latent variables. The EM algorithm is an iterative algorithm with two steps: (1) an expectation step (E-step), and (2) a maximisation step (M-step). In the E-step, we calculate the probability of the latent variables, given the observed variables and the current estimates of the parameters. In the M-step, the new estimates of the parameters are calculated to maximise the conditional expectation of the complete data likelihood $p(\mathbf{x},\mathbf{z}|\theta^l)$ given the observed data under the previous parameter value. For the estimation of the parameters of the observation model, the EM algorithm may be performed as follows: At each iteration l:

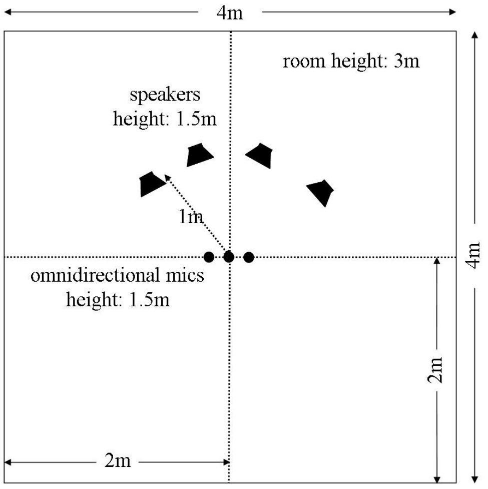

Figure 1: Layout of room used in simulations.

In the E-step, compute:

$$\tau_q^{(l)}(n) = \frac{c_q^{(l)}\mathbb{N}\left(\mathbf{x}(n)|R_q^{(l)}\right)}{\sum_{j=1}^{k} c_j^{(l)}\mathbb{N}\left(\mathbf{x}(n)|R_j^{(l)}\right)} \tag{10}$$

where $\mathbb{N}$ is the complex Gaussian distribution.

In the M-step, compute:

$$R_q^{(l+1)} = \frac{\sum_{n=1}^{\eta} \tau_q^{(l)}(n)\mathbf{x}(n)\mathbf{x}(n)^H}{\sum_{n=1}^{\eta} \tau_q^{(l)}(n)} \tag{11}$$

$$c_q^{(l+1)} = \frac{1}{\eta}\sum_{n=1}^{\eta} \tau_q^{(l)}(n) \tag{12}$$

In order to perform frequency-domain beamforming, the signal received by each microphone is separated into narrow-band frequency bins using the STFT. The EM algorithm is then applied separately in each frequency bin. For each t-f point (n, f), the output of the non-linear beamformer is given by:

$$\hat{s}(n,f) = \sum_{q=1}^{k} \tau_{q,f}\mathbf{w}_{q,f}^H\mathbf{x}(n,f) \tag{13}$$

where:

$$\mathbf{w}_{q,f}^H = \frac{\mathbf{e}^H R_{q,f}^{-1}}{\mathbf{e}^H R_{q,f}^{-1}\mathbf{e}} \tag{14}$$

4. EXPERIMENTAL EVALUATION

In order to illustrate the performance of the non-linear beamformer, multichannel recordings of several speech sources were simulated using impulse responses determined by the room image method [1] using the *rir.m*[1] function. The positions of the microphones and

[1] http://2pi.us/code/rir.m

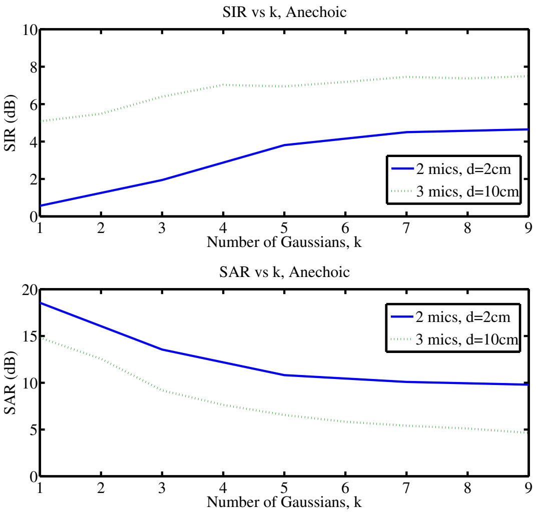

Figure 2: Average performance as a function of the number of Gaussian components k in the GMM model.

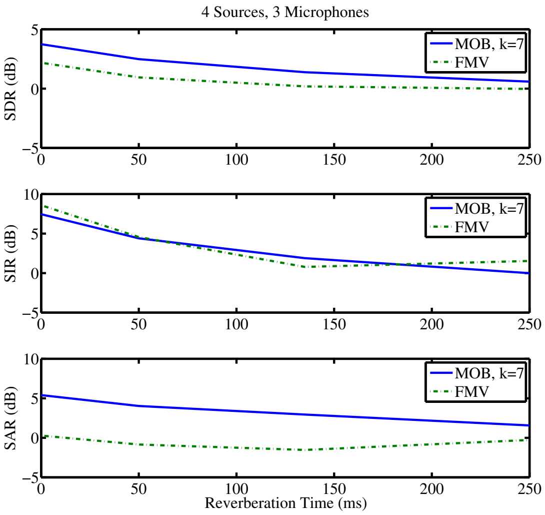

Figure 3: Separation using three microphones: average performance as a function of reverberation time.

the sources are illustrated in Figure 1. Two microphone arrays were used. The first has three microphones with a 10 cm spacing, and the second has two microphones with a 2 cm spacing. The number of the sources was four. The sources were placed in a semi-circle of radius 1 m around the microphone arrays at angles $\phi = \{45, 75, 100, 140\}^{\circ}$. We use the speech files used in the development data in [7], where eight speech files were grouped into two mixtures. The speech signals were of a duration equal to 10 s, and were sampled at 16 kHz.

To measure the quality of the signal estimate $\hat{s}$ with respect to the original signal s, we use the signal to distortion ratio (SDR), source to interference ratio (SIR) and the sources to artifacts ratio (SAR) calculated as defined in [7]. Though we note that the SAR measure does not fully capture the nature of the distortion in the output and recommend that the reader also listens to the output signals. The speech files used in the simulations and the outputs can be found online [2].In our results, the SDR, SIR and SAR values were averaged over all the sources and mixtures.

Figure 2 shows the average performance at the output of the non-linear beamformer in the anechoic case as a function of the number of Gaussian components k in the GMM model. In this experiment, four sources were operating in an anechoic environment. The case of $k = 1$ is equivalent to a time-invariant MVDR beamformer. The SIR increases with k, and then stays constant when $k \geq 7$. The increase in the SIR is more pronounced in the two microphone case, where the separation using a time-invariant beamformer ($k = 1$) gives bad results. Although there is a unity-gain response in the direction of the desired source signal, the SAR decreases with k. The decrease in the SAR can be attributed to the non-linear attenuation of the interfering sources. These artifacts therefore introduce distortion only into the residual interfering signals. We stress that the MOB is by definition distortionless in the look-direction.

Figure 3 shows the average performance as a function of the room reverberation time when four sources are operating, and the microphone array used has three microphones with a 10 cm microphone spacing. We compare the performance of a mixture of beamformers with the performance of the FMV algorithm. A STFT of frame size 1024 samples is used. In the FMV algorithm, the STFT step size is 16 samples, while a step size of 256 samples is used in the MOB algorithm. The MOB ($k = 7$) can attain an SIR of 7.5 dB in anechoic rooms.

Figure 4 shows the average performance as a function of the room reverberation time when four sources are operating, and the microphone array used has two microphones with a 2 cm microphone spacing. We compare the performance of a mixture of beamformers with the performance of the DUET and FMV algorithms. The DUET algorithm gives a high SIR, but suffers from a low SAR. The low SAR can be attributed to the binary masking of t-f points where the sources overlap. In contrast to the MOB, this distorts the desired signal itself. In DUET, when the desired source is dominant, we attribute all the received signal to the source, and when it is not dominant, we null the output. This generates musical noise due to spectro-temporal discontinuities in the source estimates.

Figures 5 and 6 show the average performance as a function of the room reverberation time when 20 dB i.i.d. additive Gaussian noise is added at the microphones. Both the MOB and FMV were robust to the additive noise and achieved good separation performance.

5. CONCLUSION

A frequency-domain non-linear beamformer was introduced and applied to source separation for under-determined speech mixtures. The beamformer is derived assuming non-Gaussian interference-plus-noise signals modelled using a mixture of Gaussians distribution. This estimator introduces additional degrees of freedom to the beamformer by exploiting the super-Gaussianity (sparsity) of the interferers.

The non-linear beamformer has low computational costs, and does not need to know or estimate the number of interfering sources. The number of components in the mixture of Gaussians distribution controls the flexibility of the model and can be used to trade-off complexity with performance. The non-linear beamformer can be applied to microphone arrays with two or more microphones. The unity gain constraint on the direction of arrival of the desired source signal results in a clear desired signal output, and avoids any permutation ambiguities. Simulation results in under-determined mixtures with room reverberation confirmed the non-linear beamformer's ability to successfully separate speech sources.

In the future, we plan to investigate the use of an on-line EM algorithm - instead of the batch EM algorithm used herein - that allows for the observation model parameters to be updated in real-time. Furthermore, we would like to compare the MOB against other speech separation techniques.

[2] http://www.see.ed.ac.uk/~s0565920/EUSIPCO08/

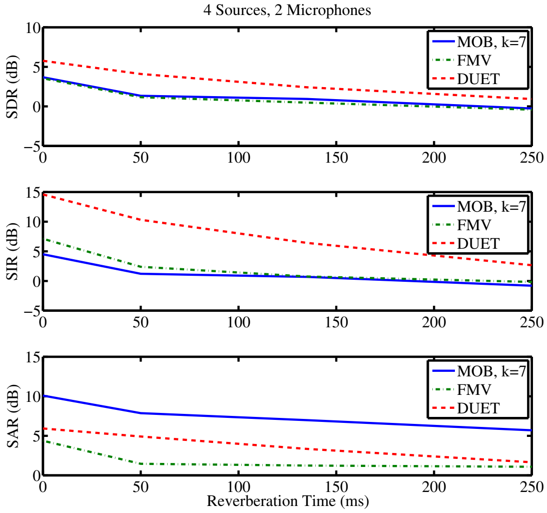

Figure 4: Separation using two microphones: average performance as a function of reverberation time.

Acknowledgements

The authors wish to thank Scott Rickard for providing the implementation of the DUET algorithm. Michael E. Davies acknowledges support of his position from the Scottish Funding Council and their support of the Joint Research Institute with the Heriot-Watt University as a component part of the Edinburgh Research Partnership. Mohammad A. Dmour also acknowledges support of the Wolfson Microelectronics Scholarship.

REFERENCES

[1] J. Allen and D. Berkley. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65(4):943–950, 1979.

[2] O.L. Frost. An algorithm for linearly constrained adaptive array processing. *Proceedings of the IEEE*, 60(8):926–935, 1972.

[3] L. Griffiths and C. Jim. An alternative approach to linearly constrained adaptive beamforming. *IEEE Transactions on Antennas and Propagation*, 30(1):27–34, 1982.

[4] M. Lockwood, D. Jones, R. Bilger, C. Lansing, W. O'Brien Jr., B. Wheeler, and A. Feng. Performance of time- and frequency-domain binaural beamformers based on recorded signals from real rooms. *The Journal of the Acoustical Society of America*, 115(1):379–391, 2004.

[5] S. Rickard. The DUET blind source separation algorithm. In S. Makino, T.-W. Lee, and H. Sawada, editors, *Blind Speech Separation*. Springer Netherlands, 2007.

[6] H. L. Van Trees. *Optimum Array Processing*. John Wiley & Sons, Inc., 2002.

[7] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. Rosca. First stereo audio source separation evaluation campaign: data, algorithms and results. In *7th International Conference on Independent Component Analysis and Source Separation*, 2007.

[8] O. Yilmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing*, 52(7):1830–1847, 2004.

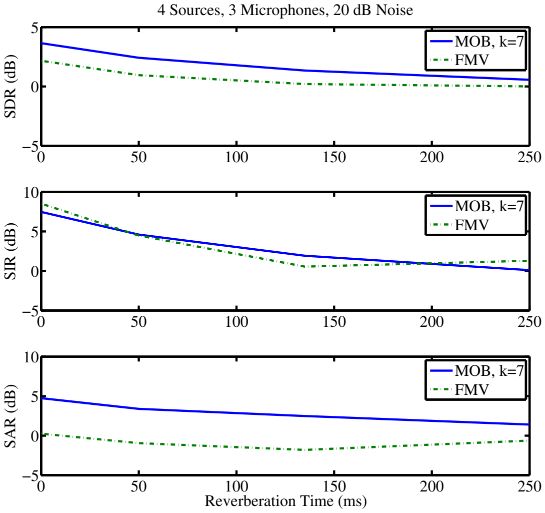

Figure 5: Separation using three microphones, with 20 dB noise: average performance as a function of reverberation time.

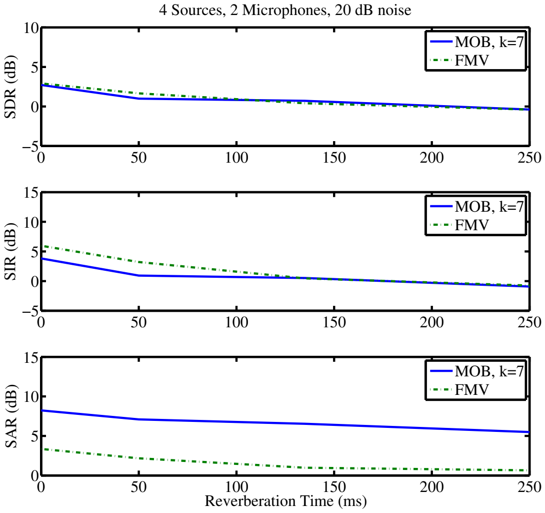

Figure 6: Separation using two microphones, with 20 dB noise: average performance as a function of reverberation time.