AN APPROACH TO UNDER-DETERMINED SPEECH SEPARATION BASED ON A NON-LINEAR MIXTURE OF BEAMFORMERS

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 describes frequency-domain non-linear beamformers that can extract a target speech source from among multiple interfering speech sources when there are fewer microphones than sources (the under-determined case). Our approach models the data in each frequency bin via Gaussian mixture distributions, which can be learnt using the expectation maximisation (EM) algorithm. A non-linear beamformer is then developed, based on this model. The proposed non-linear beamformer is a non-linear weighted sum of linear minimum mean square error (MMSE) or minimum variance distortionless response (MVDR) beamformers. The resulting beamformer requires the direction of arrival of the target speech source to be known in advance, but the number of interferers does not need to be known or estimated. Simulations of the non-linear beamformers in under-determined mixtures with room reverberation confirm its capability to successfully separate speech sources.

1. INTRODUCTION

Speech separation is the process of extracting a target speech source from observations corrupted by interfering sources and noise. Speech separation is used in a wide range of applications, such as hearing aids, human-computer interaction, surveillance, and hands-free telephony. 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 (the under-determined case).

Various methods have been proposed for solving the speech separation problem. Linear multichannel filtering techniques such as independent component analysis (ICA) can attain excellent separation performance in determined mixtures. In under-determined mixtures, non-linear techniques which exploit the sparseness of speech sources and time-frequency (t-f) diversity play a vital role. One popular approach to perform under-determined speech separation is t-f masking. In the degenerate unmixing estimation technique (DUET) [9], binary masks are determined from the spatial location information contained in the short time Fourier transform (STFT) coefficients of a stereo mixture. 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 independent factor analysis [2], it was proposed to learn the source densities from the observed data. The sources were modeled as independent random variables with Gaussian mixture models (GMMs). An expectation maximisation (EM) algorithm [4] was used to learn the parameters of the model, namely the mixing matrix, noise covariance, and source density parameters. In [3], approximations were used to overcome the problem that the number of mixtures in the observation density in [2] grows exponentially with the number of sources. The observation density is written as a summation of Gaussians with decaying weights, and then the number of Gaussians is truncated in order to retain only those with reasonable size weights.

In this paper, we describe frequency-domain non-linear beamformers that can perform speech separation of under-determined mixtures, and do not require knowledge of the number of speakers. This beamformer utilises GMMs to model the data in each frequency bin. This in turn can be learnt using the EM algorithm. The signal estimator comprises of a set of minimum mean square error (MMSE) or minimum variance distortionless response (MVDR) beamformers. 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 states. This approach results in a "soft decision" filter for the observed signal. The resulting non-linear beamformer combines the benefits of non-linear time-varying separation in t-f masking with the benefits of spatial filtering in the linear beamformers.

The organisation of this paper is as follows. Section 2 reviews the linear MMSE beamformer, and then introduces the GMM-based non-linear beamformers. 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 the conclusions 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 signal. That is,

$$\mathbf{x} = s\mathbf{e} + \mathbf{v} \tag{1}$$

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 snapshots. We assume that the desired source and the interference are uncorrelated. The interference has spatial correlation according to the angles of the contributing interferers.

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, 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, $\mathbf{v} \sim \mathbb{N}(0, R_v)$. Additionally, it is assumed that the desired source and the interference 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 $\mathbf{x}$:

$$\hat{s}_{MMSE} = E\left[s|\mathbf{x}\right] = \int p(s|\mathbf{x}).s\,ds \tag{2}$$

This mean is referred to as the conditional mean. It can be shown that the conditional mean can be expressed as [7]:

$$E[s|\mathbf{x}] = \frac{\mathbf{e}^H R_v^{-1}\mathbf{x}}{\mathbf{e}^H R_v^{-1}\mathbf{e}} \cdot \frac{\sigma_s^2}{\sigma_s^2 + \left(\mathbf{e}^H R_v^{-1}\mathbf{e}\right)^{-1}} \tag{3}$$

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 linear MMSE estimator is just a shrinkage of the MVDR beamformer.

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 [6], 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 [6], 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 [6]. 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 intervals (rapid response).

2.3 GMM-based non-linear beamformers

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 a previous paper [5], a non-linear beamformer was developed assuming a distortionless response in the direction of the desired source, and a mixture of k zero-mean Gaussians $q = 1,...,k$ with covariances $R_{x,q}$ and mixing proportions c_q were used to model the observed mixture $\mathbf{x}$ (the desired source and interference together). This leads to a simple learning algorithm and the desired signal can be estimated using this mixture of MVDR beamformers:

$$\mathbf{w}_1^H = \sum_{q=1}^{k} \tau_q \frac{\mathbf{e}^H R_{x,q}^{-1}}{\mathbf{e}^H R_{x,q}^{-1}\mathbf{e}} \tag{4}$$

where τ_q is the relative contribution for each linear MVDR beamformer, and is calculated as the posterior probability (specific to each time-frequency point) of its corresponding Gaussian component. This beamformer is a non-linear weighted sum of distortionless MVDR beamformers, where the weights sum to unity, therefore it is distortionless in the look-direction. However, since we have a distortionless constraint, we cannot exploit the sparsity of the desired source signal.

In this section, we shall describe the density of the desired source signal s as a mixture of k_s zero-mean complex-valued 1-dimensional Gaussians $q_s = 1,...,k_s$ with variances σ_{s,q_s}^2 and mixing proportions c_{s,q_s}:

$$p(s|\theta_s) = \sum_{q_s=1}^{k_s} c_{s,q_s} \frac{1}{\pi\sigma_{s,q_s}^2} \exp\left(\frac{-|s|^2}{\sigma_{s,q_s}^2}\right) \tag{5}$$

where $\theta_s = (c_{s,1},...,c_{s,k_s},\sigma_{s,1}^2,...,\sigma_{s,k_s}^2)$, and the mixing proportions $c_{s,q_s} = p(q_s)$ are constrained to sum to one. In addition, we shall describe the density of the interference signal $\mathbf{v}$ as a mixture of k_v zero-mean complex-valued N-dimensional Gaussians $q_v = 1,...,k_v$ with covariances R_{v,q_v} and mixing proportions c_{v,q_v}:

$$p(\mathbf{v}|\theta_v) = \sum_{q_v=1}^{k_v} c_{v,q_v} \frac{1}{\pi^N \left|R_{v,q_v}\right|} \exp\left(-\mathbf{v}^H R_{v,q_v}^{-1}\mathbf{v}\right) \tag{6}$$

where $\theta_v = (c_{v,1},...,c_{v,k_v},R_{v,1},...,R_{v,k_v})$, and the mixing proportions $c_{v,q_v} = p(q_v)$ are constrained to sum to one. The number of components k_s and k_v controls the flexibility of the model.

The MMSE estimate of the desired signal s is the mean of the a posteriori probability density of s given $\mathbf{x}$:

$$\begin{aligned}
\hat{s}_{MMSE} &= E[s|\mathbf{x}] = \int p(s|\mathbf{x}).s\,ds \\
&= \int \sum_{q_s=1}^{k_s}\sum_{q_v=1}^{k_v} p(s,q_s,q_v|\mathbf{x}).s\,ds \\
&= \int \sum_{q_s=1}^{k_s}\sum_{q_v=1}^{k_v} p(q_s,q_v|\mathbf{x}).p(s|\mathbf{x},q_s,q_v).s\,ds \\
&= \sum_{q_s=1}^{k_s}\sum_{q_v=1}^{k_v} \tau_{q_s,q_v} \int p(s|\mathbf{x},q_s,q_v).s\,ds \\
&= \sum_{q_s=1}^{k_s}\sum_{q_v=1}^{k_v} \tau_{q_s,q_v} E[s|\mathbf{x},q_s,q_v]
\end{aligned} \tag{7}$$

where

$$\begin{aligned}
\tau_{q_s,q_v} &= p(q_s,q_v|\mathbf{x}) \\
&= \frac{p(\mathbf{x}|q_s,q_v).p(q_s).p(q_v)}{\sum_{q_s'=1}^{k_s}\sum_{q_v'=1}^{k_v} p(\mathbf{x}|q_s',q_v').p(q_s').p(q_v')}
\end{aligned} \tag{8}$$

is the a posteriori probability that the components q_s and q_v are active in each respective GMM when observing $\mathbf{x}$, with $\sum_{q_s}\sum_{q_v}\tau_{q_s,q_v} = 1$.

We can see that the conditional mean $E[s|\mathbf{x},q_s,q_v]$ is the linear MMSE beamformer estimator in equation (3), with $R_v = R_{v,q_v}$ and $\sigma_s^2 = \sigma_{s,q_s}^2$. The desired signal estimator in equation (7) is a non-linear weighted sum of linear MMSE beamformers over all the GMM components, and the weighting coefficients are the a posteriori probabilities of the GMM components τ_{q_s,q_v}. The mixture of MMSE beamformers is given by:

$$\mathbf{w}_2^H = \sum_{q_s=1}^{k_s}\sum_{q_v=1}^{k_v} \tau_{q_s,q_v} \frac{\sigma_{s,q_s}^2}{\sigma_{s,q_s}^2 + \left(\mathbf{e}^H R_{v,q_v}^{-1}\mathbf{e}\right)^{-1}} \cdot \frac{\mathbf{e}^H R_{v,q_v}^{-1}}{\mathbf{e}^H R_{v,q_v}^{-1}\mathbf{e}} \tag{9}$$

In comparison to independent factor analysis [2], where all sources were modeled with a mixture of Gaussians, the mixture of MMSE beamformers models all the interfering sources using one mixture of Gaussians in the observation (microphones) domain. Consequently, the number of sources in the mixture is not required. This also avoids the exponential growth of the number of Gaussian components in the observation density with the number of sources.

In Section 4, we compare the performance of these two beamformers. Also, we use the interference Gaussian mixture model to implement a distortionless response mixture of beamformers, which uses the interference model covariances R_{v,q_v} instead of $R_{x,q}$:

$$\mathbf{w}_3^H = \sum_{q_s=1}^{k_s} \sum_{q_v=1}^{k_v} \tau_{q_s,q_v} \frac{\mathbf{e}^H R_{v,q_v}^{-1}}{\mathbf{e}^H R_{v,q_v}^{-1} \mathbf{e}} \tag{10}$$

3. MODEL LEARNING

Using the EM algorithm, we can estimate the model density parameters $\theta = (\theta_s, \theta_v) = (c_{s,1}, ..., c_{s,k_s}, \sigma_{s,1}^2, ..., \sigma_{s,k_s}^2, c_{v,1}, ..., c_{v,k_v}, R_{v,1}, ..., R_{v,k_v})$ from a set of observations $D = \{\mathbf{x}(n) : n = 1, ..., \eta\}$. 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, evaluate for $q_v = 1, ..., k_v$, $q_s = 1, ..., k_s$ and every received vector $\mathbf{x}(n)$:

$$p(q_s, q_v|\mathbf{x}(n)) = \tau_{q_s,q_v}(n) = \frac{c_{s,q_s} c_{v,q_v} p(\mathbf{x}(n)|q_s, q_v)}{\sum_{q_s'=1}^{k_s} \sum_{q_v'=1}^{k_v} c_{s,q_s'} c_{v,q_v'} p\left(\mathbf{x}(n)|q_s', q_v'\right)} \tag{11}$$

where

$$\begin{aligned} p(\mathbf{x}|q_s, q_v) &= \int p(\mathbf{x}, s|q_s, q_v) ds \\ &= \int p(\mathbf{x}|s, q_v).p(s|q_s) ds \\ &= \int \mathbb{N}\left(\mathbf{x} - \mathbf{e}s, R_{v,q_v}\right).\mathbb{N}\left(s, \sigma_{s,q_s}^2\right) ds \\ &= \mathbb{N}\left(\mathbf{x}, R_{v,q_v} + \sigma_{s,q_s}^2 \mathbf{e}\mathbf{e}^H\right) \end{aligned} \tag{12}$$

and evaluate the conditional mean and variance of the desired source given both the observed mixture and the hidden states, which are denoted by $\langle s|\mathbf{x}(n), q_s, q_v\rangle$ and $\langle ss^*|\mathbf{x}(n), q_s, q_v\rangle$ respectively. Given the hidden states and the mixture, the likelihood of s is Gaussian:

$$\begin{aligned} p(s|\mathbf{x}, q_s, q_v) &= \frac{p(\mathbf{x}, s, q_s, q_v)}{p(\mathbf{x}, q_s, q_v)} \\ &= \frac{p(s|q_s).p(\mathbf{x}|s, q_v).p(q_s).p(q_v)}{p(\mathbf{x}|q_s, q_v).p(q_s).p(q_v)} \\ &= \frac{\mathbb{N}(s, \sigma_{s,q_s}^2).\mathbb{N}(\mathbf{x} - \mathbf{e}s, R_{v,q_v})}{\mathbb{N}\left(\mathbf{x}, R_{v,q_v} + \sigma_{s,q_s}^2 \mathbf{e}\mathbf{e}^H\right)} \\ &= \mathbb{N}\left(s - \alpha_{q_s,q_v}, \beta_{q_s,q_v}\right) \end{aligned} \tag{13}$$

where

$$\beta_{q_s,q_v} = \left(\sigma_{s,q_s}^{-2} + \mathbf{e}^H R_{v,q_v}^{-1} \mathbf{e}\right)^{-1} \tag{14}$$

$$\alpha_{q_s,q_v} = \left(\sigma_{s,q_s}^{-2} + \mathbf{e}^H R_{v,q_v}^{-1} \mathbf{e}\right)^{-1} \mathbf{e}^H R_{v,q_v}^{-1} \mathbf{x} \tag{15}$$

In the M-step, evaluate for $q_v = 1, ..., k_v$ and $q_s = 1, ..., k_s$:

$$c_{v,q_v} = \frac{1}{\eta} \sum_{n=1}^{\eta} \sum_{q_s=1}^{k_s} p(q_s, q_v|\mathbf{x}(n)) \tag{16}$$

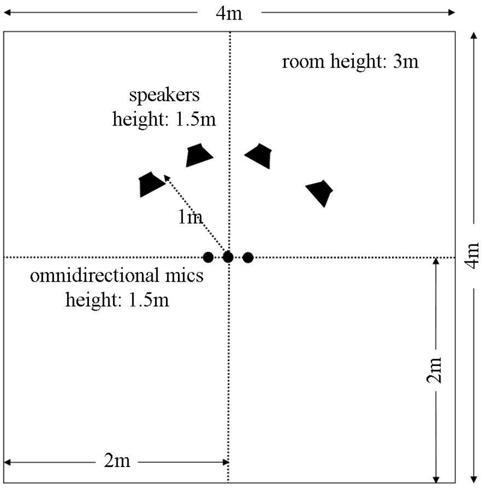

Figure 1: Layout of room used in simulations.

$$c_{s,q_s} = \frac{1}{\eta} \sum_{n=1}^{\eta} \sum_{q_v=1}^{k_v} p(q_s, q_v|\mathbf{x}(n)) \tag{17}$$

$$\sigma_{s,q_s}^2 = \frac{\sum_{n=1}^{\eta} \sum_{q_v=1}^{k_v} p(q_s, q_v|\mathbf{x}(n)) \langle ss^*|\mathbf{x}(n), q_s, q_v\rangle}{\sum_{n=1}^{\eta} \sum_{q_v=1}^{k_v} p(q_s, q_v|\mathbf{x}(n))} \tag{18}$$

$$R_{v,q_v} = \frac{\sum_{n=1}^{\eta} \sum_{q_s=1}^{k_s} p(q_s, q_v|\mathbf{x}(n)) \Lambda_{q_s,q_v}(n)}{\sum_{n=1}^{\eta} \sum_{q_s=1}^{k_s} p(q_s, q_v|\mathbf{x}(n))} \tag{19}$$

where

$$\begin{aligned} \Lambda_{q_s,q_v}(n) = \; & \mathbf{x}(n)\mathbf{x}(n)^H - \mathbf{x}(n) \langle s^*|\mathbf{x}(n), q_s, q_v\rangle \mathbf{e}^H \\ & - \mathbf{e} \langle s|\mathbf{x}(n), q_s, q_v\rangle \mathbf{x}(n)^H \\ & + \mathbf{e} \langle ss^*|\mathbf{x}(n), q_s, q_v\rangle \mathbf{e}^H \end{aligned} \tag{20}$$

In this model, there is an ambiguity in associating variance between the desired source and the interference. It is possible to incorporate some of the source signal in the interference. To avoid this, updating the desired source component variances is not performed in the first few iterations. This prevents the source components shrinking to zero variance.

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:

$$\hat{s}_f(n) = \mathbf{w}_f^H(n) \mathbf{x}_f(n) \tag{21}$$

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]. The positions of the microphones and 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. . We use speech files taken from the TIMIT speech corpus to create five mixtures of male sources, and five mixtures of female sources. The speech signals were of a duration equal to 10 s, and were sampled at 16 kHz. The number of

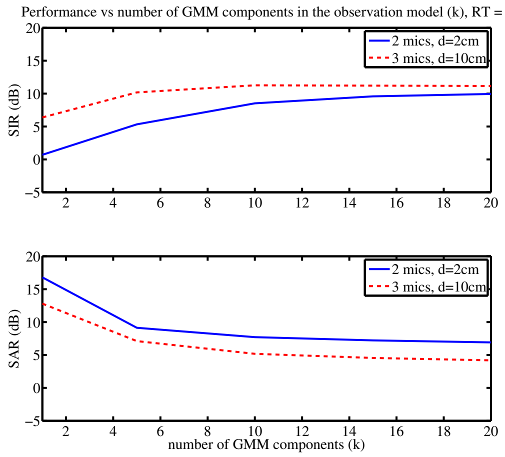

Figure 2: Average performance of the non-linear beamformer $\mathbf{w}_1$ in equation (4) as a function of the number of Gaussian components k in the GMM model.

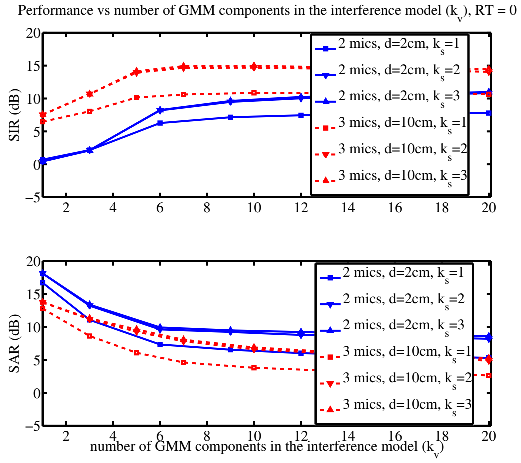

Figure 3: Average performance of the non-linear beamformer $\mathbf{w}_3$ in equation (10) as a function of the number of Gaussian components k_v and k_s in the GMM model.

the sources in each mixture was four. The sources were placed in a semi-circle of radius 1 m around the microphone arrays at angles $\phi = \{45, 75, 100, 140\}^\circ$.

To measure the quality of the signal estimate $\hat{s}$ with respect to the original signal s, we use the source to interference ratio (SIR) and the sources to artifacts ratio (SAR) calculated as defined in [8]. In our results, the 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 of equation (4) 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, but the improvement is insignificant at $k > 10$. 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 mixture of MVDR beamformers is by definition distortionless in the look-direction.

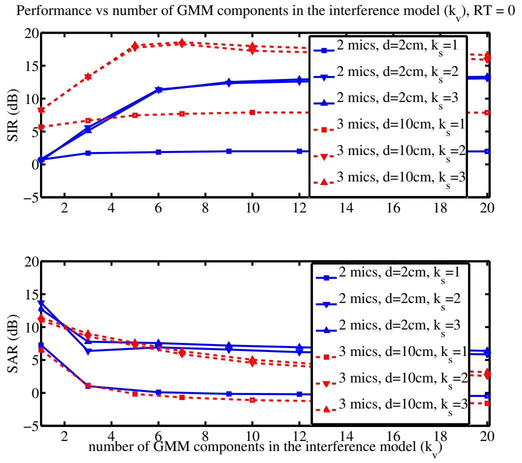

Figure 4: Average performance of the non-linear beamformer $\mathbf{w}_2$ in equation (9) as a function of the number of Gaussian components k_v and k_s in the GMM model.

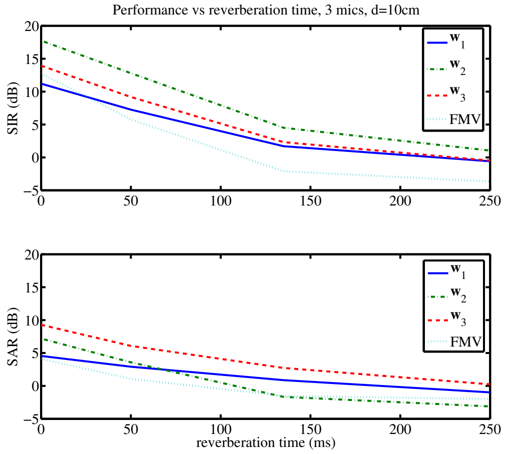

Figure 5: Separation using three microphones: average performance as a function of reverberation time.

Figure 3 shows the average performance at the output of the non-linear beamformer of equation (10) in the anechoic case as a function of the number of Gaussian components in the interference model k_v and the number of Gaussian components in the source model k_s. We can see that there is little gain for increasing the number of source Gaussian components k_s to more than two. In the two microphones case, The SIR increases with k_v, but the improvement is insignificant at $k_v > 10$. In the three microphones case, The SIR peaks around $k_v = 7$, and then levels off at higher k_v. The non-linear beamformer can attain a SIR of 10 dB in the two microphones case, and 15 dB using three microphones.

Figure 4 shows the average performance at the output of the mixture of MMSE beamformers of equation (9) in the anechoic case as a function of the number of Gaussian components in the interference model k_v and the number of Gaussian components in the source model k_s. The non-linear beamformer can attain a SIR of 13 dB in the two microphones case, and 18 dB using three microphones. However, the SAR was decreased in comparison to Figure 3 because the distortionless constraint is no longer held.

Figure 5 shows the average performance as a function of the

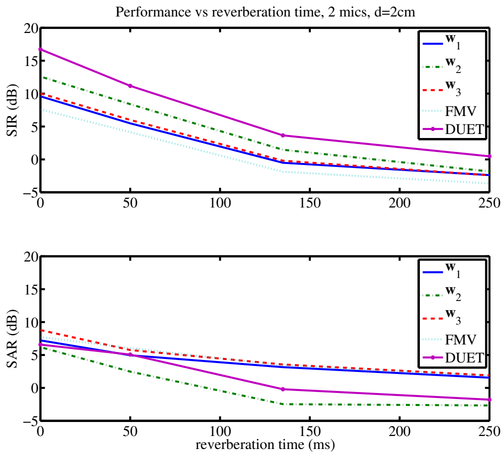

Figure 6: Separation using two microphones: average performance as a function of reverberation time.

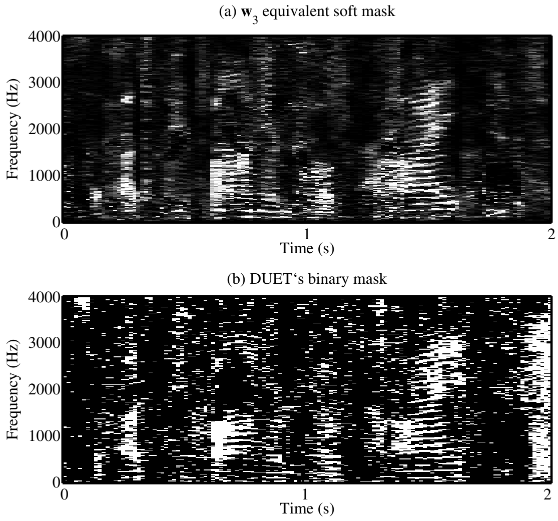

Figure 7: (a) $\mathbf{w}_3$ equivalent soft mask (b) DUET's binary masks

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 the three non-linear beamformers (equations (4), (10), and (9)) with the performance of the FMV algorithm. $k = 15$ was used in the beamformer of equation (4), and $k_s = 2, k_v = 5$ was used in the two other beamformers. A STFT of frame size 1024 samples is used. In the FMV algorithm, a small step size of 16 samples is required, while a step size of 256 samples is sufficient in the non-linear beamformers.

Figure 6 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. $k = 15$ was used in the beamformer of equation (4), and $k_s = 2, k_v = 12$ was used in the two other beamformers. We compare the performance of the three non-linear beamformers with the performance of the DUET and FMV algorithms. The DUET algorithm and the mixture of MMSE beamformers ($\mathbf{w}_2$) gives a high SIR, but suffers from a very low SAR at higher reverberation times. The non-linear beamformers of equations (4) and (10) have significantly lower artifacts in higher reverberation times.

Figure 7 compares the equivalent mask of the non-linear beamformer $\mathbf{w}_3$ with the t-f mask of DUET. The equivalent mask is computed at each t-f point as the ratio of the energy of the desired signal estimate to the energy of the observed mixture. The non-linear beamformers approach results in a "soft decision" mask for the observed signal.

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 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 does not need to know or estimate the number of interfering sources. The number of components in the mixture of Gaussians distributions 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. 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 would like to investigate the use of other linear constrained minimum variance (LCMV) beamformers and the use of auditory filter banks instead of the STFT. Through this, we aim to improve the performance of the beamformers in higher reverberation times.

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] H. Attias. Independent factor analysis. *Neural Computation*, 11(4):803–851, 1999.

[3] M. Davies and N. Mitianoudis. Simple mixture model for sparse overcomplete ICA. In *Vision, Image and Signal Processing, IEE Proceedings*, volume 151, pages 35–43, feb 2004.

[4] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

[5] M. Dmour and M.E. Davies. Under-determined speech separation using gmm-based non-linear beamforming. In *Proceedings of the sixteenth European Conference on Signal Processing (EUSIPCO 2008)*, 2008.

[6] 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.

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

[8] E. Vincent, R. Gribonval, and C. Fevotte. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech and Language Processing*, 14(4):1462–1469, jul 2006.

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