Multi-Channel Speech Separation with Soft Time-Frequency Masking

Rahil Mahdian Toroghi, Friedrich Faubel, Dietrich Klakow

Spoken Language Systems, Saarland University, Saarbrücken, Germany
{Rahil.Mahdian, Friedrich.Faubel}@lsv.uni-saarland.de

Abstract

This paper addresses the problem of separating concurrent speech through a spatial filtering stage and a subsequent time-frequency masking stage. These stages complement each other by first exploiting the spatial diversity and then making use of the fact that different speech signals rarely occupy the same frequency bins at a time. The novelty of the paper consists in the use of auditory-motivated log-sigmoid masks, whose scale parameters are optimized to maximize the kurtosis of the separated speech. Experiments on the Pascal Speech Separation Challenge II show significant improvements compared to previous approaches with binary masks.

Index Terms: speech recognition, microphone arrays, time-frequency masking, kurtosis maximization

1. Introduction

Early sound capturing systems for hands-free speech recognition [1, 2] aimed at acquiring a high-quality speech signal from a single, distant speaker. These systems typically employed a spatial filter in order to extract the signal originating from the direction of the speaker. Noise and reverberation were suppressed, provided they came from other directions. When the research focus shifted towards meeting recognition [3], it turned out, however, that overlapping speech is far more difficult to handle. This motivated the combination of classical beamforming techniques with blind source separation approaches such as time-frequency masking [4, 5] and minimization of the mutual information (MMI) [6, 7]. The resulting systems eventually lined up for competition at the PASCAL Speech Separation Challenge II (SSC2) [5, 7]. In this work, we continue along the lines of these approaches. But we propose to replace the binary time-frequency masks in [5, 7] by their sigmoid counterparts. This is based on the rationale that (1) soft masks can more accurately account for uncertainties [8] and (2) they include binary masks as a special case, with the scale parameter approaching infinity. The latter raises the question which choice of a scale parameter would give a good separation performance. To answer this question, we here avail ourselves of the work of Li and Lutman [9] who have shown that the kurtosis (a) decreases in dependency of the number of speakers as well as (b) that it highly correlates with the speech recognition performance. Consequently, we here use the kurtosis as a criterion for optimizing the scale parameter. In addition to this, we investigate the use of a linear constrained minimum variance (LCMV) beamformer, which separates concurrent speech by putting a distortionless constraint on the desired speaker and a zero constraint on the interfering speaker. This gave significantly better results than a superdirective beamformer [4], while being only slightly inferior to the MMI beamformer [6].

The remaining part of the paper is organized as follows. In Section 2, we give an overview of the used speech separation system. This is followed by a brief review of the considered spatial filtering techniques in Section 3 as well as postfiltering in Section 4. The proposed log-sigmoid masking approach is explained in Section 5. Experimental results on the speech separation challenge [10] are finally reported in Section 6.

2. System Overview

Following the design, which prevailed in the source separation challenge [5, 7], we here use a system which consists of three components: a spatial filtering stage, a postfiltering stage and a time-frequency (T/F) masking stage. These components are combined as shown in Figure 1.

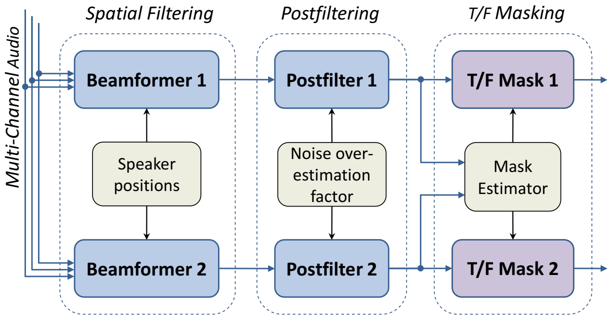

Figure 1: *Block diagram of the speech separation system used in this work.*

This work was supported by the European Union through the Marie-Curie Initial Training Network (ITN) SCALE (Speech Communication with Adaptive LEarning, FP7 grant agreement number 213850); and by the Federal Republic of Germany, through the Cluster of Excellence for Multimodal Computing and Interaction (MMCI).

The spatial filtering stage exploits the fact that speakers tend to reside at different positions. This allows us to separate their speech by directing one beamformer at each speaker (which in case of the speech separation challenge necessitates two beamformers to be operated in parallel). The beamformers we consider here are a delay-and-sum beamformer, a superdirective beamformer [2, 11, 4, 5], the MMI beamformer from [6, 7] as well as an interference canceling LCMV (see Section 3.2). Following [7], we use an additional postfiltering stage, which aims at further reducing the noise. This is achieved under the assumption of spatially uncorrelated noise at the sensor pairs [12], which does not really hold[1] in the presence of a strong directed interference such as a second speaker. For this reason, Maganti et al. [4, 5] proposed a specialized crosstalk cancellation postfilter. It makes use of the *W-disjoint orthogonality* of speech [13], which states that different speakers tend to excite different frequency bands at a time. Consequently, Maganti et al. use a binary mask, which selects the stronger speaker in each time-frequency unit (based on the relative power at the beamformer outputs), and then sets the weaker speaker to zero (in this unit). This is what essentially happens in the T/F masking stage in Figure 1. Deviating from [4, 5], we here however explore the use of soft masks that account for uncertainty of one speaker being stronger than the other.

3. Spatial Filtering Stage

The aim of spatial filtering is to extract the signal coming from the direction of the desired speaker while suppressing noise and reverberation coming from other directions. This is done based on the fact that sound sources introduce time differences of arrival (TDOAs) in dependency of their position in relation to the array. Considering an array of L microphones with locations $\mathbf{m}_i$, $i = 1, \ldots, L$, the possibly easiest way to achieve spatial filtering is to (1) calculate the TDOAs that would be expected for a given source position and array geometry, (2) invert the TDOAs in order to align the signals coming from the desired position and then (3) sum the aligned signals in order to extract the desired signal while attenuating other sources (i.e. the non-aligned part) through destructive interference. This approach is called delay-and-sum beamforming and it is described in more detail in the following.

3.1. Delay-and-Sum Beamforming

Let us consider a plane wave approaching the array aperture from direction

$$\mathbf{a} = \begin{bmatrix}\cos\theta\sin\phi & \sin\theta\sin\phi & \cos\phi\end{bmatrix}^T \tag{1}$$

with azimuth θ and elevation ϕ. Then, using the far field assumption, the delay which is introduced at the i-th microphone in relation to the center of the array (i.e. $\mathbf{0}$) is $\tau_i = -\mathbf{a}^T\mathbf{m}_i/c$ where c denotes the speed of sound. Translating these delays to phase shifts in the frequency domain leads to the so-called *array manifold vector* $\mathbf{v}$:

$$\mathbf{v}(\omega) = \begin{bmatrix}e^{-j\omega\tau_1} & \cdots & e^{-j\omega\tau_L}\end{bmatrix}^T \tag{2}$$

where ω is the angular frequency. Now, denoting the frequency spectrum of the signals $x_i(t)$, $i = 1, \ldots, L$, at the microphones by

$$\mathbf{X}(\omega) = \begin{bmatrix}X_1(\omega) & \cdots & X_L(\omega)\end{bmatrix}^T,$$

the frequency spectrum $S(\omega)$ of sound coming from direction $\mathbf{a}$ can be extracted as the scalar product of $\mathbf{X}(\omega)$ with the weight vector $\mathbf{w}(\omega) \triangleq \frac{1}{N}\mathbf{v}(\omega)$:

$$\begin{aligned} Y(\omega) &= \mathbf{w}^H(\omega)\cdot\mathbf{X}(\omega) \\ &= \underbrace{\mathbf{w}^H(\omega)\cdot S(\omega)\mathbf{v}}_{S(\omega)} + \mathbf{w}^H(\omega)\cdot\mathbf{N}(\omega) \end{aligned} \tag{3}$$

where $S(\omega)$ denotes the spectrum of the desired signal and where $\mathbf{N}(\omega)$ denotes the noise vector. In particular note that the multiplication by $\mathbf{w}^H(\omega)$ is nothing else but the frequency domain equivalent to inverting the delays and summing. The noise remaining after spatial filtering is $\mathbf{w}^H(\omega)\mathbf{N}(\omega)$. The dependency on ω will be dropped in most of the following in order to improve the readability.

3.2. LCMV Beamforming

Delay-and-sum beamforming may be generalized to using arbitrary weight vectors $\mathbf{w}$ that are optimized according to certain criteria. For the particular case of *Linear Constraint Minimum Variance* (LCMV) beamforming, the optimization criterion consists of minimizing the noise power at the output of the beamformer while at the same time maintaining a set of linear constraints. The latter typically includes the *distortionless constraint*, which ensures that the signal from the desired direction is neither attenuated nor amplified, i.e. $\mathbf{w}^H\mathbf{v} = 1$. This leads to the following optimization problem:

$$\min_{\mathbf{w}} \mathbf{w}^H\Sigma_{nn}\mathbf{w} \quad \text{subject to} \quad \mathbf{w}^H\mathbf{v} = 1 \tag{4}$$

where Σ_{nn} denotes the power spectral density (PSD) matrix of the noise. Now, solving (4) with a Lagrange multiplier, we obtain the *Minimum Variance Distortionless Response* (MVDR) beamformer whose weight vector is given by:

$$\mathbf{w}_{\text{mvdr}} = \frac{\Sigma_{nn}^{-1}\mathbf{v}}{\mathbf{v}^H\Sigma_{nn}^{-1}\mathbf{v}}. \tag{5}$$

The most problematic aspect is getting a reliable estimate of the PSD matrix. But that can be avoided by making the

[1] We would like to emphasize that, though suboptimal, postfiltering led to a significant reduction of the word error rate in practice (see Section 6).

assumption of a homogeneous noise field, for which Σ_{nn} can be written [11]: $\Sigma_{nn} = \Phi_{nn}\Gamma_{nn}$ where Φ_{nn} denotes the noise power and where Γ_{nn} is the noise coherence matrix. With this factorization, the MVDR solution devolves to:

$$\mathbf{w}_{\text{sdb}}^{H} = \frac{\Phi_{nn}^{-1}\Gamma_{nn}^{-1}\mathbf{v}}{\Phi_{nn}^{-1}\mathbf{v}^{H}\Gamma_{nn}^{-1}\mathbf{v}} = \frac{\Gamma_{nn}^{-1}\mathbf{v}}{\mathbf{v}^{H}\Gamma_{nn}^{-1}\mathbf{v}}. \tag{6}$$

Following [11], we in particular assume a spherically isotropic noise field, i.e. a particular choice of Γ_{nn}, which is optimal for reverberant environments [14]:

$$(\Gamma_{nn})_{i,j}(\omega) = \text{sinc}\left(\omega\frac{\|m_i - m_j\|}{c}\right). \tag{7}$$

This results in the *Superdirective Beamformer* (SDB) [11]. In the case of the speech separation scenario, we would actually like to "listen" to one speaker while suppressing the other. This can be achieved by using the general LCMV solution [15] in order to apply a 1 in the direction of the desired speaker and a 0 in the direction of the interference, i.e. $\mathbf{w}^H\mathbf{v}_1 = 1$ and $\mathbf{w}^H\mathbf{v}_2 = 0$ with $\mathbf{v}_1$ and $\mathbf{v}_2$ denoting the location of the desired and interfering speaker, respectively. This leads to the following weight vector:

$$\mathbf{w}_{\text{lcmv}} = \Gamma_{nn}^{-1}\mathbf{C}\left(\mathbf{C}^H\Gamma_{nn}^{-1}\mathbf{C}\right)^{-1}\mathbf{f} \tag{8}$$

with $\mathbf{C} = [\mathbf{v}_1 \quad \mathbf{v}_2]$, $f = [1 \quad 0]^T$ and with Γ_{nn} again denoting the noise coherence matrix of the spherically isotropic noise field. As a result we have a simple extension of the superdirective beamformer to the speech separation case.

3.3. MMI Beamforming

In contrast to the above approach, which could be considered to be a bit more of a standard solution, the authors of [6] proposed to solve the speech separation problem by using two beamformers whose weights are jointly optimized to minimize the mutual information at the beamformer outputs. In order to ensure that the distortionless constraint is maintained, the authors used the *Generalized Sidelobe Canceler* (GSC) configuration

$$\mathbf{Y}_i = (\mathbf{w}_{q,i} - \mathbf{B}_i\mathbf{w}_{a,i})^H\,\mathbf{X}(K) \tag{9}$$

with the quiescent weight vector being chosen as $\mathbf{w}_{q,i} = \mathbf{v}_i$ for the i-th speaker. The blocking matrix $\mathbf{B}_i$ by definition projects to the subspace, which is orthogonal to $\mathbf{w}_{q,i}$. The active weight vectors $\mathbf{w}_{a,i}$ are optimized to individually minimize the mutual information

$$I(Y_1, Y_2) = \mathcal{E}\left\{\log\frac{p(Y_1, Y_2)}{p(Y_1)p(Y_2)}\right\} \tag{10}$$

in each frequency bin. Due to the lack of an analytical solution, the minimization is performed in an iterative fashion, as explained in more detail in [6]. The resulting beamformer is referred to as *Minimum Mutual Information* (MMI) beamformer.

4. Postfiltering Stage

As shown in [16], the minimum mean square error (MMSE) solution[2] to spatial filtering consists of an MVDR beamformer plus a Wiener postfilter:

$$\mathbf{w}_{\text{mmse}}(\omega) = \underbrace{\left(\frac{\Phi_{ss}(\omega)}{\Phi_{ss}(\omega) + \Phi_{nn}(\omega)}\right)}_{H(\omega)}\mathbf{w}_{\text{mvdr}}(\omega) \tag{11}$$

The $\Phi_{ss}(\omega)$ and $\Phi_{nn}(\omega)$ denote the speech and noise power at the output of the array. Following Zelinski [12], they are here estimated as follows:

$$\Phi_{ss} \approx \frac{2}{L(L-1)}\Re\left\{\sum_{i=1}^{L-1}\sum_{j=i+1}^{L} v_i^*\Phi_{x_ix_j}v_j\right\} \tag{12}$$

$$\Phi_{nn} \approx \frac{1}{L}\sum_{i=1}^{L}\Phi_{x_ix_i} - \Phi_{ss} \tag{13}$$

with $\Phi_{x_ix_j}$ and $\Phi_{x_ix_i}$ denoting the cross and power spectral densities of the sensor signals and with v_i denoting the i-th coefficient of the array manifold vector. The dependency on ω has again been dropped for the sake of readability. Equations (12) and (13) are based on the assumption that the noise is incoherent. But this may actually not be the case in practice. Hence, we here use a noise overestimation factor β in order to compensate for possible systematic errors. This is achieved by changing the frequency response $H(\omega)$ in (11) to:

$$H(\omega) = \frac{\Phi_{ss}(\omega)}{\Phi_{ss}(\omega) + \beta\Phi_{nn}(\omega)} \tag{14}$$

Early speech recognition experiments indicated that a value β of 0.5 gives reasonable results – at least on the MC-WSJ-AV corpus [10] which has been used in the speech separation challenge. This value compares to a theoretical optimum of $1/L = 0.125$ for delay-and-sum beamforming with an 8-sensor array [17] (under assumption of incoherent noise).

5. Time/Frequency Masking

Although the use of postfiltering led[3] to a significant reduction of the word error rate, it is suboptimal for speech separation where we are mainly facing a strong directive interference (the second speaker) [5]. Hence, Maganti et al. [4] proposed to replace the postfiltering stage by a specialized cross-talk cancellation postfilter, which makes use of the fact that different speakers tend to excite different frequency bands at a time. Consequently, the authors of [4, 5] use binary masking of the beamformer

[2] This is also called the multi-channel Wiener filter.

[3] see the experimental results in Section 6

outputs Y_i, $i \in \{1,2\}$, in order to extract the estimated clean speech spectra $\hat{S}_i(\omega,t)$ at time t:

$$\hat{S}_i(\omega,t) = M_i(\omega,t) \cdot Y_i(\omega,t) \tag{15}$$

where $M_i(\omega,t)$ denotes a binary mask, which would optimally be set to 1 if the T/F unit (ω,t) is used by that (the i-th) speaker and which is set to 0 otherwise. This approach is justified by Rickard and Yilmaz's approximate W-disjoint orthogonality (WDO) of speech [13], which states that

$$S_1(\omega,t)S_2(\omega,t) \approx 0 \quad \forall \omega, t. \tag{16}$$

Hence, perfect demixing via binary T/F masks is possible if the time-frequency representations of the sources do not overlap (i.e. the WDO condition holds for all T-F points) [18, 13]. This, however, requires the masks $M_i(\omega,t)$ to be known. As that is not the case in practice, Maganti et al. [4] proposed to use the following estimate:

$$M_i(\omega,t) = \begin{cases} 1, & |Y_i(\omega,t)| \geq |Y_j(\omega,t)| \quad \forall j \\ 0, & otherwise \end{cases} \tag{17}$$

which is based on the assumption that the spatial filtering stage has already suppressed the interfering speaker, such that $|Y_i(\omega,t)| > |Y_j(\omega,t)|$ if the i-th speaker is using the (ω,t)-th frequency unit while the j-th speaker is not. The same approach has been used in [7].

5.1. Log-Sigmoid Masks

Although binary masking is optimal in theory, it has certain deficiencies in practice. First of all, the mask estimates may be erroneous if the interfering speaker is not sufficiently suppressed through spatial filtering. Secondly, the approach may not be optimal in reverberant environments, such as the SSC task [10], where the spectral energy is smeared in time. Hence, we here propose the use of soft masks, which can more appropriately treat the arising uncertainties. The use of sigmoid masks is motivated by

1. the work of Barker et al. [8] where it has been shown that sigmoid masks give better results in the presence of mask uncertainties.
2. the work of Araki et al. [19] where its has been shown (a) that soft-masks can perform better in convolutive mixtures and (b) that a simple sigmoid mask can perform comparably to other sophisticated soft masks or even better.

In case of the speech separation scenario, we may use sigmoid masks in order to apply a weight to each of the sources, based on the difference of their magnitudes:

$$M_i(\omega,t) = \frac{1}{1+\exp[-\alpha\left(|Y_i(\omega,t)| - |Y_j(\omega,t)|\right)]} \tag{18}$$

with $i \in 1,2$, $j = 3-i$ and with α being a scale parameter, which specifies the sharpness of the mask. Instead of directly applying this mask, we here use the fact that the human auditory system perceives the intensity of sound in a logarithmic scale. This can be incorporated into (18) by replacing the magnitudes $|Y_i(\omega,t)|$ by logarithmic magnitudes $\log|Y_i(\omega,t)|$:

$$\widetilde{M}_{i,\alpha}(\omega,t) = \frac{1}{1+\left(\dfrac{|Y_j(\omega,t)|}{|Y_i(\omega,t)|}\right)^{\alpha}} \tag{19}$$

where the logarithms have been pulled out of the exponential function. Although the scale parameter α may be chosen individually for each frequency bin, it is here jointly optimized once for each utterance, as described in Section 5.2. In the particular case where $\alpha = 2$, the log-sigmoid mask is identical to a Wiener filter

$$\widetilde{M}_{i,2}(\omega,t) = \frac{|Y_i(\omega,t)|^2}{|Y_i(\omega,t)|^2 + |Y_j(\omega,t)|^2} \tag{20}$$

with $|Y_i(\omega,t)|$ being the magnitude of clean speech and with $|Y_j(\omega,t)|$ being the magnitude of the noise.

5.2. Kurtosis Optimization

Motivated by Li and Lutman's work[4] we here use the subband-kurtosis as a measure for judging the separation quality of concurrent speech. Consequently, the quality of a separated utterance $\hat{S}_{i,\alpha}$ is determined as the average kurtosis over all frequencies:

$$\text{kurt}\left\{\hat{S}_{i,\alpha}\right\} = \frac{1}{|\Omega|}\sum_{\omega\in\Omega}\frac{\frac{1}{T}\sum_{t=1}^{T}|\hat{S}_{i,\alpha}(\omega,t)|^4}{\left(\frac{1}{T}\sum_{t=1}^{T}|\hat{S}_{i,\alpha}(\omega,t)|^2\right)^2} \tag{21}$$

where $\omega \in \Omega$ and $t \in \{1,\ldots,T\}$ denote the angular frequency and the discrete time index, respectively. The $\hat{S}_{i,\alpha}(\omega,t)$ are the separated subband samples after time-frequency masking, i.e.

$$\hat{S}_{i,\alpha}(\omega,t) = \widetilde{M}_{i,\alpha}(\omega,t)Y_i(\omega,t),$$

with scale parameter α. Now, α may be optimized by running a search over a range $\mathcal{R}_\alpha$ of possible values and then selecting that α for which kurt$\{\hat{S}_{i,\alpha}\}$ is maximized. To get a good coverage of different mask shapes with few parameters to test, we use:

$$\mathcal{R}_\alpha = \{\exp(a/10), \quad a = -50,\ldots,50\}. \tag{22}$$

This optimization is done once for each utterance and individually for each speaker. Also note that soft-masking can easily be extended to more than 2 speakers by using $j = \text{argmax}_{k\neq i}|Y_k(\omega,t)|$ instead of $j = 3-i$ in (19).

[4] which clearly shows that the subband kurtosis decreases with the number of simultaneous speakers [9]

6. Experiments

The performance of the proposed system has been evaluated on the two speaker condition of the Multi-Channel Wall Street Journal Audio-Visual (MC-WSJ-AV) corpus [10]. This dataset has been used in the PASCAL Speech Separation Challenge II [7, 5] and it consists of two concurrent speakers who are simultaneously reading sentences form the Wall Street Journal. The total number of utterances is 356 (or 178, respectively, if we consider the fact that two sentences are read at a time [5]). The speech recognition system used in the experiments is identical to the one in [7], except that we use three passes only (instead of four): a first, unadapted pass; a second pass with unsupervised MLLR feature space adaptation; and a third pass with full MLLR adaptation. The estimated speaker positions are the ones used in [7].

6.1. Results

Table 1 shows the word error rates (WERs) we obtained with different configurations of the speech separation system from Figure 1. For spatial filtering, we used either a delay-and-sum beamformer (DSB), a superdirective beamformer (SDB), the LCMV beamformer from Section 3.2 or the minimum mutual information (MMI) beamformer from [6, 7]. The first row of table 1 reveals that the WER of the plain SDB is 9% lower than that of the DSB. LCMV and MMI beamforming give a further reduction of 10% and therewith perform comparably (58.6% versus 57.6%). The second row of table 1 shows the combination of spatial filtering with binary masking. This combination gives a significant improvement over the plain beamformers: almost 20% for the DSB, 10% for the SDB and still 8% and 5% for the LCMV and MMI beamformers. The use of kurtosis optimized log-sigmoid masks (row 3) results in similar improvements, except for the SDB where we have a further reduction of 13% compared to binary masking.

Mask	PF	Beamformer WER(%)			
		DSB	SDB	LCMV	MMI
None	no	77.87	68.73	58.56	57.58
Binary	no	58.89	57.20	49.97	52.15
log-sigm.	no	58.65	44.63	48.56	52.39
None	yes	69.07	61.65	56.77	56.98
Binary	yes	51.03	45.33	51.06	49.99
log-sigm.	yes	48.09	42.73	43.47	46.83
Headset		23.44			

Table 1: *Word error rates for different beamformers with and without postfiltering (PF) and time-frequeny masking. The last row gives a comparison to the headset data which has been recorded in parallel to the array [10].*

These results changed dramatically when a postfilter[5] was applied between spatial filtering and masking. In this case, the combination with log-sigmoid masks gave the best speech recognition results obtained in this paper, with a word error rate of approximately 43% for the SDB and LCMV beamformers. The MMI and DSB beamformer were only slightly inferior, with a performance of 47% and 48%. Binary masking was between 3% and 6% worse. These results demonstrate that the right choice of post-processing can have a tremendous effect. The best WER is not necessarily achieved with the most sophisticated beamformer.

6.2. Some Analysis

Due to the large improvements obtained with log-sigmoid masks, we considered it worth investigating how the kurtosis optimization affects the mask shape. For this purpose, we first selected some utterances which seemed to be well separated (after processing with the SDB) and then plotted their kurtosis (after T/F-masking) in dependency of the scale parameter. An example of such a plot is shown in Figure 2, along with a plot for an utterance where the separation was poor.

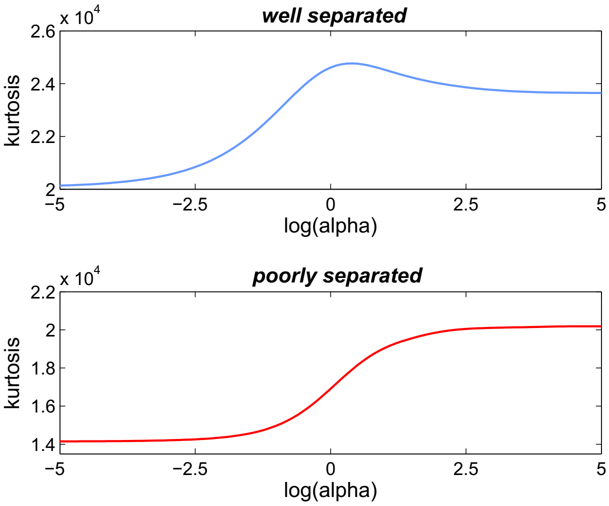

Figure 2: *Kurtosis for well (upper plot) and poorly (lower plot) separated speech, in dependence of α.*

Motivated by the strong differences in these plots, we divided the corpus into a set of utterances for which the speech seemed to be well-separated and a set of utterances for which the separation seemed to be poor. Subsequently, we plotted the average mask shape for each of the sets, as shown in Figure 3. This reveals that the kurtosis maximization selects harder (closer to binary) masks when the separation through spatial filtering is poor. It selects softer (i.e. less strongly reacting) masks when the separation quality is good.

[5] the one from Section 4

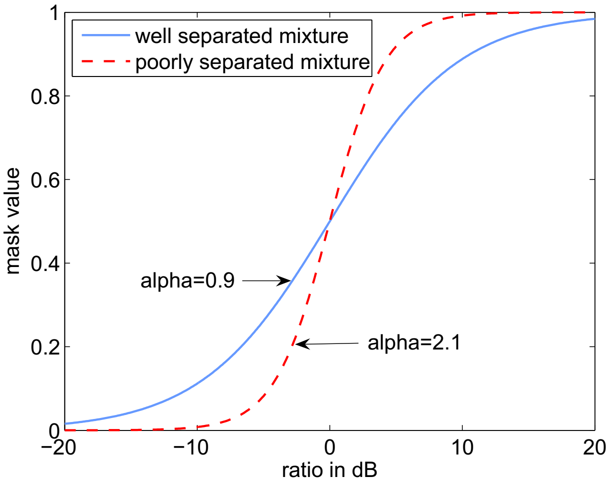

Figure 3: *Mask shape for well and poorly separated mixtures, with the x-axis representing the ratio of* $|Y_2|/|Y_1|$.

7. Conclusions

We have investigated the use of a speech separation system which consists of a spatial filtering stage, a postfiltering stage and a subsequent T/F-masking stage. We have shown in particular (1) that log-sigmoid masks can significantly outperform binary masks and (2) that kurtosis optimization chooses the mask shape in dependence of the separation quality after spatial filtering. In total, the proposed approach gave a reduction of over 25% in WER over a plain SDB (in combination with postfiltering). Apart from the above, we have shown that an interference canceling LCMV with a diffuse noise field design gives almost the same performance as a minimum mutual information beamformer.

8. Acknowledgments

We would like to thank Kenichi Kumatani from Disney Research for kindly providing the MMI beamformed data of the speech separation challenge. Apart from him, we are also indebted to John McDonough who provided the speaker positions and the ASR framework used in the experiments, as well as to Tobias Wolff and Markus Buck from Nuance who advised us about the use of a noise over-estimation factor in the postfilter implementation.

9. References

[1] M. Omologo, M. Matassoni, and P. Svaizer, "Speech recognition with microphone arrays," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds., pp. 331–353. Springer, 2001.

[2] I. A. McCowan, C. Marro, and L. Mauuary, "Robust speech recognition using near-field superdirective beamforming with post-filtering," *Proc. ICASSP*, vol. 3, pp. 1723–1726, June 2000.

[3] D. Moore and I. McCowan, "Microphone array speech recognition: Experiments on overlapping speech in meetings," *Proc. ICASSP*, vol. 5, pp. 497–500, Apr. 2003.

[4] H. K. Maganti, D. Gatica-Perez, and I. McCowan, "Speech enhancement and recognition in meetings with an audio-visual sensor array," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 8, pp. 2257–2269, Nov. 2007.

[5] I. Himawan, I. McCowan, and M. Lincoln, "Microphone array beamforming approach to blind speech separation," *Proc. MLMI*, pp. 295–305, June 2007.

[6] K. Kumatani, T. Gehrig, U. Mayer, E. Stoimenov, J. McDonough, and M. Wölfel, "Adaptive beamforming with a minimum mutual information criterion," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 8, pp. 2527–2541, Nov. 2007.

[7] J. McDonough, K. Kumatani, T. Gehrig, E. Stoimenov, U. Mayer, S. Schacht, M. Wölfel, and D. Klakow, "To separate speech - a system for recognizing simultaneous speech," *Proc. MLMI*, pp. 283–294, June 2007.

[8] J. Barker, L. Josifovsky, M. Cooke, and P. Green, "Soft decisions in missing data techniques for robust automatic speech recognition," *Proc. ICSLP*, vol. 1, pp. 373–376, Oct. 2000.

[9] G. Li and M. E. Lutman, "Sparseness and speech perception in noise," *Proc. SAPA*, pp. 7–11, Sept. 2006.

[10] M. Lincoln, I. McCowan, I. Vepa, and H. K. Maganti, "The multi-channel wall street journal audio visual corpus (MC-WSJ-AV): Specification and initial experiments," *Proc. ASRU*, pp. 357–362, Nov. 2005.

[11] M. Bitzer and K. U. Simmer, "Superdirective microphone arrays," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds., pp. 19–38. Springer, 2001.

[12] R. Zelinski, "A microphone array with adaptive postfiltering for noise reduction in reverberant rooms," *Proc. ICASSP*, vol. 5, pp. 2578–2581, Apr. 1988.

[13] S. Rickard and O. Yilmaz, "On the approximate w-disjoint orthogonality of speech," *Proc. ICASSP*, vol. 1, pp. 529–532, May 2002.

[14] R. K. Cook, R. V. Waterhouse, R. D. Berendt, S. Edelman, and M. C. Thompson, "Measurement of correlation coefficients in reverberant sound fields," *Journal of the Acoustic Society of America*, vol. 27, pp. 1072–1077, Nov. 1955.

[15] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Magazine*, pp. 4–24, Apr. 1988.

[16] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds., pp. 39–62. Springer, 2001.

[17] K. U. Simmer and A. Wasiljeff, "Adaptive microphone arrays for noise suppression in the frequency domain," *Workshop on Adaptive Algorithms in Communications*, Sept. 1992.

[18] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, July 2004.

[19] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Blind sparse source separation with spatially smoothed time-frequency masking," *Proc. IWAENC*, Sept. 2006.