Soft Mask Estimation for Single Channel Speaker Separation

Aarthi M. Reddy, Bhiksha Raj

Mitsubishi Electric Research Laboratories
Cambridge, MA, 02139

aarthi@lantana.cs.iitm.ernet.in, bhiksha@merl.com

1. Abstract

The problem of single channel speaker separation, attempts to extract a speech signal uttered by the speaker of interest from a signal containing a mixture of auditory signals. Most algorithms that deal with this problem, are based on *masking*, where reliable components from the mixed signal spectrogram are inversed to obtain the speech signal from speaker of interest. As of now, most techniques, estimate this mask in a binary fashion, resulting in a *hard mask*. We present a technique to estimate a soft mask that weights the frequency sub-bands of the mixed signal. The speech signal can then be reconstructed from the estimated power spectrum of the speaker of interest. Experimental results shown in this paper, prove that the results are better than those obtained by estimating the hard mask.

2. Introduction

In a natural scenario, speech signals are usually perceived against a background of many sounds. The human ear has the ability to efficiently separate necessary speech signals from a plethora of other auditory signals, even if these signals have similar overall frequency characteristics, and are perfectly coincident in time.

However, it has not been possible to achieve similar results through automatic techniques. The problem of source separation – separation of one or more desired signals from mixed recordings of multiple signals – has traditionally been approached by using multiple microphones, in order to obtain sufficient information about the incoming speech signals to perform effective separation. Typically, no prior information about the speech signals is assumed, other than that the multiple signals that have been combined are statistically independent, or uncorrelated with each other. The problem is treated as one of Blind Source Separation (BSS), which can be performed by techniques such as deconvolution [1], decorrelation [2] and Independent Component Analysis (ICA) [3]. This approach works best when the number of recording channels (microphones) are at least as many as the number of signal sources (speakers).

A more challenging, and potentially far more interesting problem is that of separating speech signals from a single channel recording, i.e. when the multiple concurrent speakers/sources have been recorded by only a single microphone. Since the problem is inherently underspecified, prior knowledge, either of the physical nature, or the signal or statistical properties of the signals, must be assumed. Computational auditory scene analysis (CASA) based solutions (e.g. [4], [5]), are based on the premise that human-like performance is achievable through processing that models the mechanisms of human perception, e.g. via signal representations that are based on models of the human auditory system [6], the grouping of related phenomena in the signal, and the ability of humans to comprehend speech even when several components of the signal have been removed. Jang *et. al.* [7] present a signal-based approach where basis functions extracted from training instances of the signals from the individual sources are used to identify and separate the component signals in mixtures.

A third approach, and one that is related to the subject matter of this paper, uses a combination of detailed statistical models and Weiner filtering to separate the component speech signals in a mixture. The methods are largely founded on two assumptions: 1. any time-frequency component of a mixed recording is dominated by only one of the components of the independent signals (an assumption that is sometimes termed as the *log-max* assumption), 2. perceptually acceptable signals for any speaker can be reconstructed from only a subset of the time-frequency components, suppressing others to a floor value. Roweis [8] models the distributions of short-time Fourier transform (STFT) representations of the signals from the individual speakers by HMMs. Mixed signals are modeled by *factorial* HMMs, that combine the HMMs for the individual speakers. Speaker separation proceeds by first identifying the most likely combination of states to have generated each short-time spectral vector from the mixed signal. The means of the states are used to construct spectral *masks* that identify the time-frequency components that are estimated as belonging to each of the speakers. The time-frequency components identified by the masks are used to reconstruct the separated signals, a procedure Rowies terms *re-filtering*.

Hershey *et. al.* [9] extend the above technique by modeling narrow and wide-band spectral representations separately for the speakers. The overall statistical model for each speaker is thus a factorial HMM that combines the two spectral representations. The mixed speech signal is further augmented by visual features representing the speakers' lip and facial movements. Reconstruction is performed by estimating a target spectrum for the individual speakers from the factorial HMM apparatus, estimating a Weiner filter that suppresses undesired time-frequency components in the mixed signal, and reconstructing the signal from the remaining spectral components.

Reyes-Gomez *et. al.* [10] decompose the signal into multiple frequency bands. The overall distribution for any speaker is a *coupled* HMM in which each spectral band is separately modeled, but the permitted trajectories for each spectral band are governed by all spectral bands. The statistical model for the mixed signal is a larger factorial HMM derived from the coupled HMMs for the individual speakers. Speaker separation is performed using the re-filtering technique proposed by Roweis. Similar techniques have also been proposed by other authors.

All of the above methods feature several simplifying approximations. Roweis and Reyes *et. al.* utilize the log-max assumption to describe the relationship of the log power spectrum of the mixed signal to that of the component signals. In conjunction with the log-max assumption, it is assumed that the distribution of the log of the maximum of two log-normal random variables is well defined by a normal distribution whose mean is simply the largest of the means of the component random variables. In addition, only the most likely combination of states from the HMMs for the individual speakers is used to identify the spectral masks for the speakers. Hershey *et. al.* do not use the log-max assumption, preferring instead to more accurately model the power spectrum of the mixed signal as the sum of the power spectra of the component signals. However, in order to account for this model, they approximate the distribution of the sum of log-normal random variables as a log-normal distribution whose moments are derived as combinations of the statistical moments of the component random variables. In all of these techniques speaker separation is achieved by *suppressing* time-frequency components that are estimated as not representing the speaker, and reconstructing signals from only the remaining time-frequency components.

In this paper we present some algorithms that attempt to avoid some of the approximations in the above techniques. We continue to utilize the log-max algorithm, primarily because the approximation introduces little error, as we explain in Section 3. However, the probability distributions computed for the log spectral vectors of the mixed signal are exact, within the restrictions of the log-max model. In Section 5 we describe a minimum mean-squared error (MMSE) [1] estimation technique that attempts to *reconstruct* all spectral components of the separated signals, as opposed to the conventional technique of only retaining spectral components that are known to belong to the signal with some certainty. In Section 6 we present a *soft-mask* technique that assigns probabilities to the various spectral bands. Reconstruction is not performed by the simple re-filtering used by Roweis et. al., but by ensuring that the reconstructed signals sum back to the original mixed signal. For both techniques, we derive contributions to the separated signals from every combination of component densities from the individual speakers, rather than just the most likely combination.

We utilize simple mixture Gaussian densities to model the distributions of entire spectral vectors. In terms of the statistical models employed, the closest comparable algorithm is the MAXVQ algorithm [11], which is essentially the same as the re-filtering algorithm in [8], with the difference that mixture Gaussian densities are employed instead of HMMs. However, the algorithms presented in this paper can be easily extended to work with more detailed, better models such as HMMs, factorial HMMs, or coupled HMMs, such as those used in [8], [9] and [10], although we have not attempted to do so in this paper. The algorithms are presented in the context of separating signals from two speakers, however, as explained in Section 8 they can be extended to multiple speakers, with some modifications.

The experimental results presented in Section 7 indicate that the presented techniques can result in better reconstruction than that obtained with the MAXVQ algorithm. As explained in Section 8, this leads us to hypothesize that results obtained with techniques that use more detailed statistical models can be improved by using the extensions proposed in this paper.

3. The Mixing Model

Let $X(t)$ and $Y(t)$ be the signals generated by two speakers S_X and S_Y, speaking simultaneously into a single microphone. The mixed signal $Z(t)$ recorded by the microphone is the sum of the two speech signals:

$$Z(t) = X(t) + Y(t). \tag{1}$$

Let $X(\omega)$ represent the power spectrum of $X(t)$, i.e.

$$X(\omega) = |\mathcal{F}(X(t))|^2, \tag{2}$$

where $\mathcal{F}$ represents the Fourier transform, and the $|.|^2$ operation computes a component-wise squared magnitude. Similarly, $Y(\omega)$ and $Z(\omega)$ denote the power spectra of $Y(t)$ and $Z(t)$ respectively. If we assume that $X(t)$ and $Y(t)$ are uncorrelated with each other, we get:

$$Z(\omega) = X(\omega) + Y(\omega). \tag{3}$$

The relationship in equation 3 is strictly valid only in the long term, and is not guaranteed to hold for power spectra measured from analysis windows of finite length. In general, equation 3, becomes more valid as the length of the analysis window increases.

Let $x(\omega)$, $y(\omega)$ and $z(\omega)$ represent the logarithm of $X(\omega)$, $Y(\omega)$ and $Z(\omega)$ respectively. From equation 3 we get:

$$z(\omega) = \log(e^{x(\omega)} + e^{y(\omega)}), \tag{4}$$

which can be written as

$$z(\omega) = \max(x(\omega),\, y(\omega)) + \log(1 + e^{\min(x(\omega),\, y(\omega)) - \max(x(\omega),\, y(\omega))}). \tag{5}$$

In practice, the instantaneous spectral power in any frequency band of the mixed signal is typically dominated by one speaker. The *log-max* approximation codifies this observation by modifying equation 3 to

$$z(\omega) \approx \max(x(\omega),\, y(\omega)). \tag{6}$$

In the rest of this paper, we will drop the frequency argument ω and simply represent the logarithm of the power spectra, which we will refer to as log spectra, as x, y and z respectively.

The requirements for the log-max assumption to hold contradict those for equation 3, whose validity increases with the length of the analysis window. The analysis window used to estimate the power spectra of the signals must hence effect a compromise between the requirements for equations 3 and 6. In our experiments, we have utilized an analysis window of 25ms. This window size is quite common, and strikes a good balance between the window length requirements for both the uncorrelatedness and the log-max assumptions to hold.

For all the experiments reported in this paper, signals were sampled at 16Khz, and were divided into frames of 25ms, with an overlap of 15ms between adjacent frames. A 400 point Hanning window was applied to each frame, and a 512 point DFT computed from it. 257 point log power spectral vectors were derived from the resulting Fourier spectra.

Figure 1 shows the log spectrum of a 25ms segment of a mixed signal for two speakers, and the corresponding log spectra for the original unmixed signals for the two speakers. We observe that in general the value of the log spectrum of the mixed signal is very close to the larger of the log spectra for the two speakers, although it is not always exactly equal to the larger value. In general, the error between the true log spectrum and that predicted by the log-max approximation is very small.

Comparison of equations 5 and 6 shows that the maximum error introduced by the log-max approximation is $\log(2) = 0.69$. The typical values of log-spectral components in our experiments lay between 7 and 20, and the largest error introduced

[1] Accepted for presentation at Interspeech 2004 - ICSLP

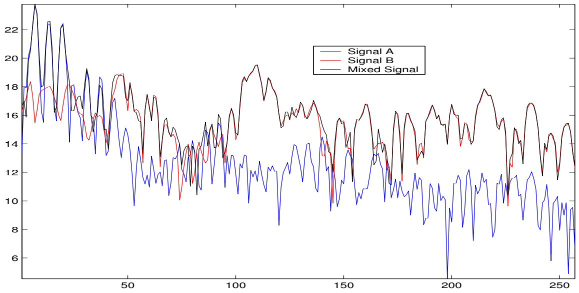


Figure 1: Log Power Spectrum for one frame of speech

by the log-max approximation was less than 10% of the value of any spectral component. More importantly, the ratio of the *average* value of the error to the standard deviation of the distribution of the log-spectral vectors is less than 0.1, (for the specific data sets used in our paper), and can be considered negligible.

4. The Statistical Model

We model the distribution of the log spectral vectors for any speaker by a mixture Gaussian density. Within each Gaussian in the mixture, the various dimensions (i.e. the frequency bands in the log spectral vector) are assumed to be independent of each other. Note that this does not imply that the frequency bands are independent of each other over the entire distribution of the speaker.

Let x and y denote log power spectral vectors for speakers S_X and S_Y respectively. According to the above model, the distribution of x for speaker S_X can be represented as

$$P(x) = \sum_{k_x=1}^{K_x} P_x(k_x) \prod_{d=1}^{D} \mathcal{N}(x_d;\ \mu^x_{k_x,d},\ \sigma^x_{k_x,d}), \tag{7}$$

where, K_x is the number of Gaussians in the mixture Gaussian, $P_x(k_x)$ represents the *a priori* probability of the k_x^{th} Gaussian, D represents the dimensionality of the the power spectral vector x, x_d represents the d^{th} dimension of x, and $\mu^x_{k_x,d}$ and $\sigma^x_{k_x,d}$, represent the mean and variance respectively of the d^{th} dimension of the k_x^{th} Gaussian in the mixture. $\mathcal{N}(x_d;\ \mu^x_{k_x,d},\ \sigma^x_{k_x,d})$ represents the value of a Gaussian density with mean $\mu^x_{k_x,d}$ and variance $\sigma^x_{k_x,d}$ at x_d.

The distribution of y for speaker S_Y can similarly be expressed as

$$P(y) = \sum_{k_y=1}^{K_y} P_y(k_y) \prod_{d=1}^{D} \mathcal{N}(y_d;\ \mu^y_{k_y,d},\ \sigma^y_{k_y,d}). \tag{8}$$

The parameters of $P(x)$ and $P(y)$ are learnt from training corpora of speech recorded independently for the two speakers.

Let z represent any log power spectral vector for the mixed signal. Let z_d denote the d^{th} dimension of z. The relationship between x_d, y_d and z_d follows the log-max approximation given in equation 6. We introduce the following notation for simplicity

$$\begin{aligned}
C_x(\omega|k_x) &= \int_{-\infty}^{\omega} \mathcal{N}(x_d; \mu^x_{k_x,d},\ \sigma^x_{k_x,d})\, dx_d && (9)\\
P_x(\omega|k_x) &= \mathcal{N}(\omega;\ \mu^x_{k_x,d},\ \sigma^x_{k_x,d}) && (10)\\
C_y(\omega|k_y) &= \int_{-\infty}^{\omega} \mathcal{N}(x_d; \mu^x_{k_y,d},\ \sigma^x_{k_y,d})\, dx_d && (11)\\
P_y(\omega|k_y) &= \mathcal{N}(\omega;\ \mu^x_{k_y,d},\ \sigma^x_{k_y,d}) && (12)
\end{aligned}$$

where, k_x and k_y represent indices in the mixture Gaussian distributions for x and y, and ω is a scalar random variable.

It can now easily be shown that

$$P(z_d|k_x, k_y) = P_x(z_d|k_x)C_y(z_d|k_y) + P_y(z_d|k_y)C_x(z_d|k_x). \tag{13}$$

i.e. under the log-max assumption, either $z = x$ and $y < z$, or $z = y$ and $x < z$. Since the dimensions of x and y are independent of each other, given the indices of their respective Gaussians, it follows that the the components of z are also independent of each other. Hence

$$P(z|k_x, k_y) = \prod_{d=1}^{D} P(z_d|k_x, k_y), \tag{14}$$

and

$$\begin{aligned}
P(z) &= \sum_{k_x,k_y} P(k_x, k_y)P(z|k_x, k_y)\\
&= \sum_{k_x,k_y} P_x(k_x)P_y(k_y) \prod_{d} P(z_d|k_x, k_y).
\end{aligned} \tag{15}$$

Note that the conditional probability of the Gaussian indices is given by

$$P(k_x, k_y|z) = \frac{P_x(k_x)P_y(k_y)P(z|k_x, k_y)}{P(z)}. \tag{16}$$

5. Minimum Mean Squared Error Estimation

The minimum-mean-squared error estimate $\hat{x}$ for a random variable x is defined as the value that has the lowest expected squared norm error, given all the conditioning factors ϕ. *i.e.*

$$\hat{x} = \text{argmin}_w E[\|\ w - x\ \|^2 |\ \phi] \tag{17}$$

It is easy to show that this estimate is given by the mean of the distribution of x, i.e. $\hat{x} = E[x \mid \phi]$.

For the problem of speaker separation, the random variables to be estimated are the log spectra of the individual speakers. Let z be the log spectrum of the mixed signal in any frame of speech. Let x and y be the log spectra of the desired unmixed signals for the frame. The MMSE estimate for x is given by

$$\begin{aligned}
\hat{x} &= E[x \mid z]\\
&= \int_{-\infty}^{\infty} xP(x|z)dx.
\end{aligned} \tag{18}$$

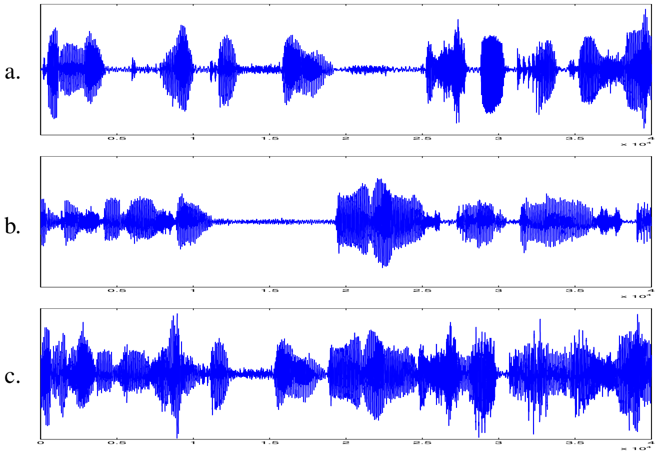
Alternately, the MMSE estimate $\hat{x}$ can be stated as a vector, whose individual components are obtained as:

$$\hat{x}_d = \int_{-\infty}^{\infty} x_d P(x_d|z)dx_d \tag{19}$$

$P(x_d|z)$ can be expanded as

$$P(x_d|z) = \sum_{k_x,k_y} P(k_x,\ k_y \mid z)P(x_d|k_x,\ k_y,\ z_d). \tag{20}$$

In this equation, $P(x_d|k_x,\ k_y,\ z_d)$ is dependent only on z_d, the d^{th} dimension of z, since individual Gaussians in the mixture Gaussians are assumed to have diagonal covariance matrices.



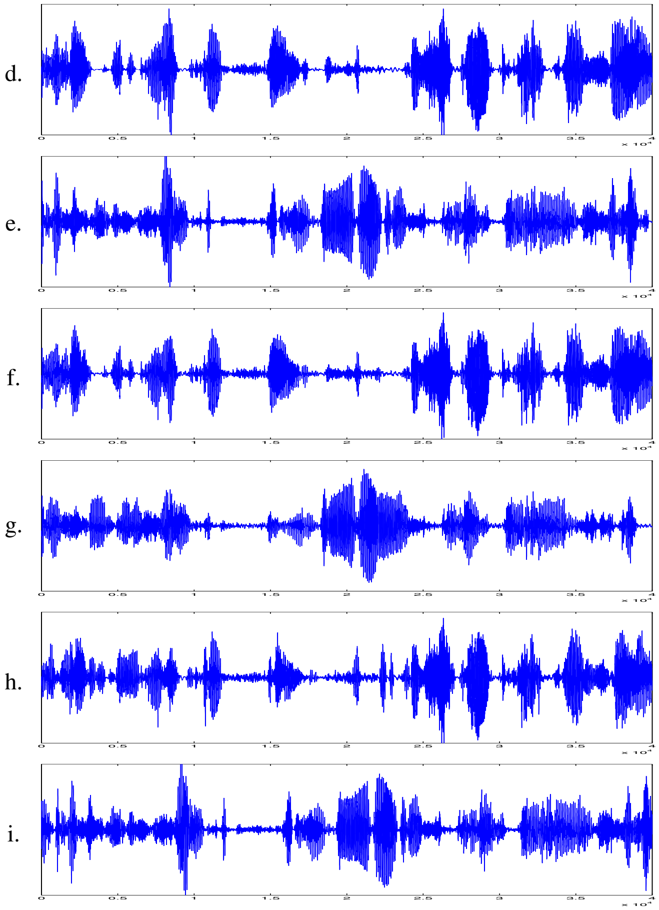


Figure 2: Waveforms for: (a) Original signal of Speaker 1; (b) Original signal of Speaker 2; (c) Digitally mixed signal; (d) Soft-Mask reconstruction of signal of Speaker 1; (e) Soft-Mask reconstruction of signal of Speaker 2; (f) MMSE reconstruction of signal of Speaker 1; (g) MMSE reconstruction of signal of Speaker 2; (h) MAXVQ reconstruction of signal of Speaker 1; (i) MAXVQ reconstruction of signal of Speaker 2

It can be shown that

$$P(x_d|k_x,k_y,z_d) = \begin{cases} \frac{P_x(x_d|k_x)P_y(z_d|k_y)}{P(z_d|k_x,k_y)} \\ \quad + \frac{P_x(z_d|k_x)C_y(z_d|k_y)\delta(x_d - z_d)}{P(z_d|k_x,k_y)} \\ \quad \text{if } x_d \leq z_d \\ 0 \quad \text{otherwise} \end{cases} \tag{21}$$

where, $\delta(x_d - z_d)$ is a Dirac delta function of x_d centered at z_d. Equation 21 has two components, one accounting for the case where x_d is less than z_d, while y_d is exactly equal to z_d, and the other for the case where y_d is less than z_d while x_d is equal to it. x_d can never be less than z_d.

Combining equations 19, 20 and 21 we get the following equation for the MMSE estimate of x_d:

$$\hat{x}_d = \sum_{k_x,k_y} \frac{P(k_x,\ k_y|z)}{P(z_d|k_x,k_y)} \Big\{ P_y(z_d|k_y)[\mu^x_{k_x,\ d}C_x(z_d|k_x) - \sigma^x_{k_x,\ d}P_x(z_d|k_x)] \tag{22}$$
$$+ C_y(z_d|k_y)P_x(z_d|k_x)z_d \Big\}.$$

The MMSE estimate for the entire vector, $\hat{x}$, is obtained by estimating each component separately using equation 22. Note that equation 22 is exact for the mixing model and the statistical distributions assumed.

Reconstructing separated speech signals: The DFT of each frame of speech from speaker S_X is computed as

$$\hat{X}(\omega) = exp(\hat{x} + i\angle Z(\omega)), \tag{23}$$

where, $\angle Z(\omega)$ represents the phase of $Z(\omega)$, the Fourier spectrum from which the log spectrum z was obtained. The estimated signal for S_X in the frame is obtained as the inverse Fourier transform of $\hat{X}(\omega)$. The estimated signals from all the frames are stitched into a continuous utterance using the overlap-add method.

6. Soft Mask Estimation

As per the log-max assumption of equation 6, z_d, the d^{th} component of any log spectral vector z computed from the mixed signal is equal to the larger of x_d and y_d, the corresponding components of the log spectral vectors for the underlying signals from the two speakers S_X and S_Y. Thus, any observed spectral component belongs completely to one of the speakers. The probability that the observed log spectral component z_d belongs to speaker S_X, and not to S_Y, conditioned on the fact that the entire observed vector is z, is given by

$$P(x_d = z_d|z) = P(x_d > y_d|z). \tag{24}$$

In other words, the probability that z_d belongs to S_X is simply the conditional probability that x_d is greater than y_d. $P(x_d > y_d|z_d)$ can be expanded as

$$P(x_d > y_d|z) = \sum_{k_x,\ k_y} P(k_x,\ k_y|z)P(x_d > y_d|z_d\ ,\ k_x,\ k_y). \tag{25}$$

Note that x_d is dependent only on z_d and not all of z, once k_x and k_y are given. Using Bayes rule, and the definition in equation 9 we obtain:

$$\begin{aligned} P(x_d > y_d|\ z_d,\ k_x,\ k_y) &= \frac{P(x_d = z_d, y_d < z_d|k_x,k_y)}{P(z_d|k_x,k_y)} \\ &= \frac{P_x(z_d|\ k_x)\ C_y(z_d|\ k_y)}{P(z_d|\ k_x,\ k_y)}. \end{aligned} \tag{26}$$

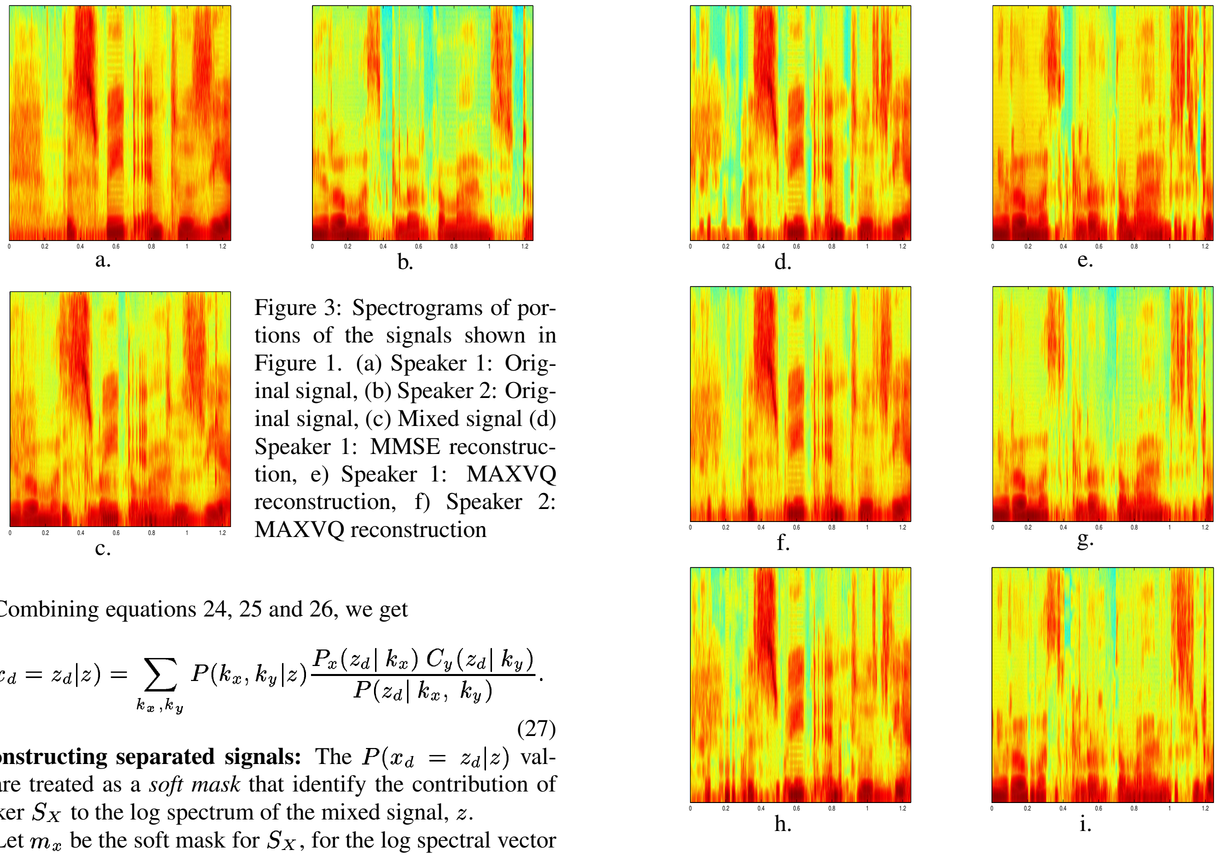


Figure 3: Spectrograms of portions of the signals shown in Figure 1. (a) Speaker 1: Original signal, (b) Speaker 2: Original signal, (c) Mixed signal (d) Speaker 1: MMSE reconstruction, e) Speaker 1: MAXVQ reconstruction, f) Speaker 2: MAXVQ reconstruction

Combining equations 24, 25 and 26, we get

$$P(x_d = z_d|z) = \sum_{k_x, k_y} P(k_x, k_y|z) \frac{P_x(z_d|\, k_x)\, C_y(z_d|\, k_y)}{P(z_d|\, k_x,\; k_y)}. \tag{27}$$

Reconstructing separated signals: The $P(x_d = z_d|z)$ values are treated as a *soft mask* that identify the contribution of speaker S_X to the log spectrum of the mixed signal, z.

Let m_x be the soft mask for S_X, for the log spectral vector z. Note that the corresponding mask for S_Y is $1 - m_x$. The estimated masked Fourier spectrum $\hat{X}(\omega)$ for S_X can be computed in one of two ways. In the first method, $\hat{X}(\omega)$ is obtained by component-wise multiplication of m_x and $Z(\omega)$, the Fourier spectrum for the mixed signal from which z was obtained.

In the second method, we apply the soft mask to the log spectrum of the mixed signal. The d^{th} component of the estimated log spectrum for S_X is given by

$$\hat{x_d} = m_{x,d}.z_d \;-\; C(z_d, m_{x,d}), \tag{28}$$

where, $m_{x,d}$ is the d^{th} component of m_x and $C(z_d, m_{x,d})$ is a normalization term that ensures that the estimated power spectra for the two speakers sum to the power spectrum for the mixed signal, and is given by

$$C(z_d, m_{x,d}) = log(e^{z_d m_{x,d}} + e^{z_d(1-m_{x,d})}). \tag{29}$$

The entire estimated log spectrum $\hat{x}$ is obtained by reconstructing each component using equation 28. The separated signals are obtained from the estimated log spectra in the manner described in Section 5.

7. Experiments and Results

Experiments were conducted to evaluate the MMSE and soft mask estimation algorithms. Approximately one hour of speech data were recorded from two speakers, one male and one female, for training mixture Gaussian distributions. Mixture Gaussian densities with 256 Gaussians were estimated for the log spectra of each of the speakers, using the Expectation Maximization algorithm.

In addition to the training data, 10 minutes of speech were recorded by each of the two speakers individually, as test data. These signals were digitally added with an SNR of of 0dB to simulate mixed recordings. Figure 2 shows an example of the signals obtained using the speaker separation algorithms presented in this paper. Figure 1(a) and 1(b) show original speech signals of Speaker 1 and Speaker 2 respectively. Figure 2(c) shows the digitally mixed signal. Waveforms of signals reconstructed using the soft mask estimation algorithm, MMSE estimator and the MAXVQ algorithm are also shown. MAXVQ uses mixture Gaussian densities as statistical models for the distributions of the log spectral vectors for the speakers, and is thus an appropriate comparator for the algorithms presented in this paper. Spectrograms of segments of the speech signals shown in figure 2 are also shown in figure 3.

In a second test, the two speakers were recorded speaking simultaneously into a single microphone, to obtain real mixed recordings. Figure 4 shows the results obtained on one such signal. The mixed signal is shown in figure 4(a). Also shown are individual speaker signals reconstructed using soft mask estimation (in figures 4(b) and 4(c)), MMSE estimator (in figures 4(d) and 4(e)), and MAXVQ approach (in figures 4(f) and 4(g)). In all cases, soft-mask based reconstruction was performed by applying soft masks to the log spectra.

8. Observation and Conclusions

In the examples shown in section 7 and other tests, the techniques proposed in this paper consistently result in cleaner speaker separation than MAXVQ. The competing speaker is significantly more suppressed than with MAXVQ. The MMSE reconstruction results in an improvement of 3-5dB in SNR for the speakers. Reconstruction with the soft masks shows a peculiar artifact: in most regions of the signals an SNR gain of 5-8dB over MAXVQ is obtained. However, in some short segments of the signal, typically about 50-200ms wide, the separated spec-

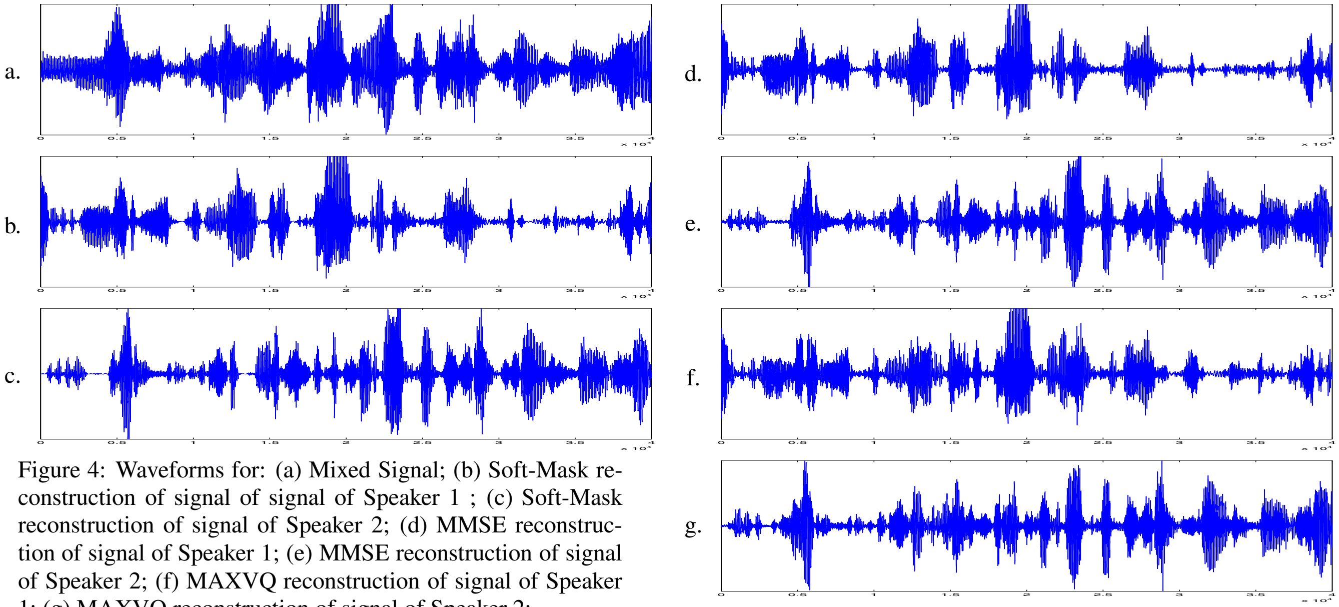


Figure 4: Waveforms for: (a) Mixed Signal; (b) Soft-Mask reconstruction of signal of signal of Speaker 1 ; (c) Soft-Mask reconstruction of signal of Speaker 2; (d) MMSE reconstruction of signal of Speaker 1; (e) MMSE reconstruction of signal of Speaker 2; (f) MAXVQ reconstruction of signal of Speaker 1; (g) MAXVQ reconstruction of signal of Speaker 2;

trum for the speaker is almost completely suppressed to a floor value. Consequently, the reconstructed signal sounds choppy.

Of the two techniques presented, only the MMSE reconstruction algorithm optimizes an objective function that is related to the SNR of the reconstructed signal, and can hence be expected to actually improve the SNR of the reconstructed signal. Any sub-optimality in the results must be attributed to inadequacies in the basic models used (*i.e.* the log-max mixing model and the mixture Gaussian statistical distributions). Additionally it must be noted that the SNR may be improved if the phase of the speaker of interest was known. However, this is not feasible as the recording is done on a single channel.

On the other hand, the soft mask methods are essentially heuristic in nature (as are all other masking based techniques) and are not guaranteed to result in improved SNR. Additionally, the technique incorrectly assigns identical phase spectra to all component signals. The fact that the latter in fact result in superior reconstruction for the most part is hence surprising.

The estimation of reconstructed spectra and the soft masks in the MMSE and the soft mask algorithms, respectively, are exact within the constraints of the mixing and statistical models used. Thus, the results obtained cannot be improved upon without improving the underlying models themselves. Both techniques are computationally intensive in their exact form, since they must explicitly compute the contribution of every combination of Gaussians for the two speakers. However, the required computation can be reduced greatly by employing the variational approximation of Ghahramani *et. al.* [12] or the simpler factorial approximation of Hershey *et. al.* [9]. These have not been implemented for this paper.

The algorithms, in their current form, employ only simple Gaussian mixture models. They can however be easily extended to employ more detailed models and processing, such as HMMs and sub-band decomposition of the signals, to obtain significantly improved separation. Also, the algorithms in this paper (and indeed most current single-channel speaker separation algorithms) assume instantaneous mixing of the signals, and ignore the effect of room response on the signals. The algorithmic formulations used in this paper enable easy incorporation of the estimation and cancellation of short-time room response. This will be the topic of future work.

9. References

[1] Bell, A. J., Sejnowski, T. J., "An Information-Maximization Approach to Blind Separation and Blind Deconvolution," Neural Computation, Vol. 7., 1129-1159, 1995

[2] Weinstein, E., Feder, M., Oppenheim, A. V., "Multi-channel Signal Separation by Decorrelation," IEEE Transactions on Speech and Audio Processing, Vol. 1, No. 4, 405-413, Oct 1993

[3] Cardoso, J-F., "Blind signal separation: statistical principles," Proceedings of the IEEE, Vol. 9, No. 10, 2009-2025, Oct 1998

[4] Weintraub, M., "A Theory and Computational Model of Auditory Monaural Sound Separation," Ph.D. dissertation, Stanford University, EE department, 1985.

[5] Lyon, R. F., "A Computational Model of Filtering, Detection and Compression in the Cochlea," Proceedings of IEEE ICASSP-82, 1282-1285, 1982

[6] Scheirer, E., Slaney, M., "Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator," Proceedings of ICASSP-97, 1997

[7] Jang, G-J, Lee, T-W, "A Maximum Likelihood Approach to Single-Channel Source Separation," Journal of Machine Learning Research, Vol. 4, 1365-1392, 2003

[8] Roweis, S. T., "Factorial Models and Re-filtering for Speech Separation and Denoising," EUROSPEECH 2003., 7(6):1009–1012, 2003.

[9] Hershey, J., Casey, M., “Audio-Visual Sound Separation Via Hidden Markov Models”, Proc. Neural Information Processing Systems 2001.

[10] Reyes-Gomez, M. J., Ellis, D. P. W., Jojic, N., "Multiband Audio Modeling for Single-Channel Acoustic Source Separation," To appear in ICASSP 2004

[11] Roweis, S. T., "One Microphone Source Separation," Advances in Neural Information Processing Systems, 13:793–799, 2001.

[12] Ghahramani, Z. , and Jordan, M. , "Factorial hidden Markov models," Machine Learning, Vol. 29, 1997