SINGLE MICROPHONE SOURCE SEPARATION USING HIGH RESOLUTION SIGNAL RECONSTRUCTION

Trausti Kristjansson, Hagai Attias

Machine Learning and Applied Statistics
Microsoft Research
{traustik,hagaia}@microsoft.com

John Hershey

University of California, San Diego
Machine Perception Lab
jhershey@cogsci.ucsd.edu

ABSTRACT

We present a method for separating two speakers from a single microphone channel. The method exploits the fine structure of male and female speech and relies on a strong high frequency resolution model for the source signals.

The algorithm is able to identify the correct combination of male and female speech that best explains an observation and is able to *reconstruct* the component signals, relying on prior knowledge to 'fill in' regions that are masked by the other speaker.

The two speaker single microphone source separation problem is one of the most challenging source separation scenarios and few quantitative results have been reported in the literature. We provide a test set based on the Aurora 2 data set and report performance numbers on a portion of this set. We achieve results of 6.59 dB average increase in SNR for female speakers and 5.51 dB for male speakers.

1. INTRODUCTION

Source separation involves recovering two or more signals that have been mixed. When multiple microphones are available the phase between the different signals can be exploited to recover the composite signals. A large body of work revolves around exploiting phase information for source separation. Source separation via Independent Component Analysis (ICA)[1, 2] relies on multiple signals as well.

The most challenging case for source separation is when only one signal is available. In this case, one has to rely exclusively on the prior knowledge of the signals to be separated.

Previous work in the area of single microphone source separation has used less accurate approximations to the mixing process[3] or sub-band representation of speech[4] which remove important correlations in the speech signal.

The core inference method used in this work has been extensively studied in the context of robust speech recognition[5], using low dimensional representations of speech. Recently we have shown [6] that statistical models of the harmonic structure of speech are of substantial value for separating speech from noise, in very noisy conditions. In this paper, we extend the method for the cross-speaker condition, where the competing signal is a second speaker.

Figures 1(a)-1(c) shows the result of running the algorithm on a single frame of the input (frame 100 from Figure 3(a)). Figure 1(a) shows the input to the algorithm (black heavy line), the female component feature vector (red dotted line) and male component feature vector (blue dashed line). Only frequencies 1200 Hz -2600 Hz are shown for clarity.

Intuitively, the algorithm has identified the best combination of male and female speech, that explains the observation. Notice that the amplitude of the male speaker is stronger in the lower half of the frequency range shown and the female speaker is stronger in the upper half of the frequency range. In the middle of the frequency range, the amplitudes are in a similar range. Notice that due to the log scale the mixed signal is effectively equal to the maximum of the two signals if one signal is considerably stronger than the other signal as happens on both ends of the frequency range. Notice also that when the values are in a similar range (e.g. at 1900 Hz) the mixed signal is not effectively equal to them maximum[1].

Figure 1(b) shows the posterior estimate for the female signal. As can be seen in Figure 1(a), the female signal is effectively masked by the male speaker in the lowest part of the frequency range and vice versa. The algorithm is able to *reconstruct* the signal in these areas based on prior knowledge of female speech encoded in the speech model. Notice that the algorithm finds a remarkably good estimate.

In the areas where the female signal is 'submerged' in the male signal, the uncertainty of the estimate is much larger than where the signal dominates. The uncertainty is quantified by the variance of the posterior and is represented by the shaded area in the figure.

Figure 1(c) shows the posterior estimate for the male signal. Similarly we see that the signal has been reconstructed where it is effectively masked, and the uncertainty is larger in these areas.

The change in uncertainty as a function of signal to noise ratio allows the model to effectively ignore frequency bands that are masked by the other signal, and attend to frequency bands that are not masked. The values in masked frequency bands are thus automatically inferred from clean ones.

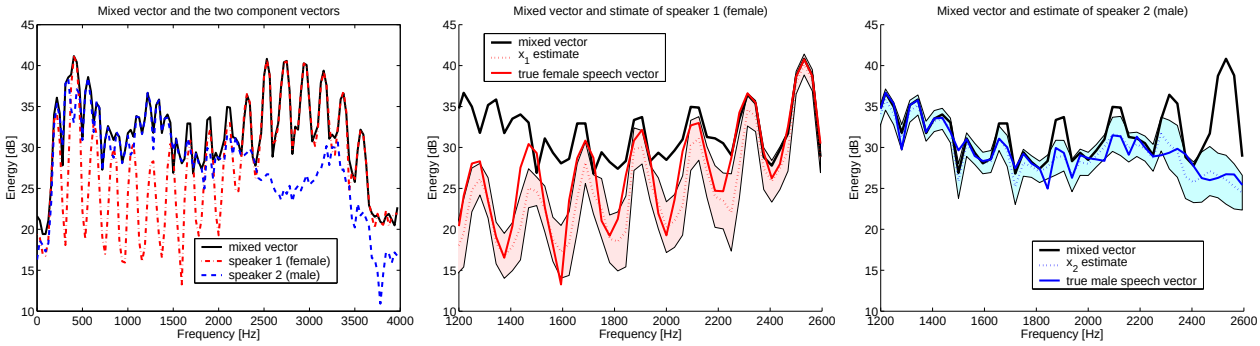
2. HIGH RESOLUTION SOURCE SEPARATION

The core of the method involves calculating posteriors for the high frequency resolution log-spectrums $p(x_1|y)$ and $p(x_2|y)$ of the two speakers, given the mixed signals. We employ the Algonquin framework [5] to calculate these posteriors. The derivation given here is exactly equivalent to the derivation when the interfering signal is noise.

The model for mixed speech in the time domain is

$$y[t] = x_1[t] + x_2[t]. \quad (1)$$

[1] in this case, the max approximation is sub-optimal[4].



(a) Mixed signal input feature vector (solid black line), and the component signals.

(b) Posterior estimate for the clean component for speaker 1 (female).

(c) Posterior estimate for the clean component for speaker 2 (male)

Fig. 1. (a) Mixed signal input feature vector (solid black line), the female component feature vector (red dotted line) and the male component feature vector (blue dashed line). (b) The posterior estimate for speaker 1 (female). Notice that signal is effectively masked in the lower portion of the frequency range. The algorithm is able to *reconstruct* these values due to the strong prior model. The shaded area represents the uncertainty of the estimate and is the first standard deviation. Notice that the uncertainty is larger for 'submerged' estimates. (c) Posterior estimate for speaker 2 (male).

where $x_1[t]$ denotes the first speaker, $x_2[t]$ denotes the second speaker, and $y[t]$ denotes the mixed signal. In the Fourier domain, the relationship becomes

$$Y(f) = X_1(f) + X_2(f) \tag{2}$$

where f designates the frequency component of the FFT. This can also be written in terms of the magnitude and the phase of each component:

$$|Y(f)|\angle Y(f) = |X_1(f)|\angle X_1(f) + |X_2(f)|\angle X_2(f) \tag{3}$$

where $|Y(f)|$ is the magnitude of $Y(f)$ and $\angle Y(f)$ is the phase.

We model only the magnitude components and do not explicitly model the phase components. The relationship between the magnitudes is

$$|Y(f)|^2 = |X_1(f)|^2 + |X_2(f)|^2 + 2|X_1(f)||X_2(f)|cos(\theta) \tag{4}$$

where θ is the angle between X_1 and X_2. For the purposes of modelling, we assume the we can model the last term as a noise term, hence we approximate this relationship between magnitudes as

$$|Y(f)|^2 = |X_1(f)|^2 + |X_2(f)|^2 + e \tag{5}$$

where the e is a random error [5]. Next we take the logarithm and arrive at the relationship in the high resolution log-magnitude-spectrum domain

$$y = x_1 + \ln(1 + \exp(x_2 - x_1)) + \varepsilon \tag{6}$$

where $y = \log(|Y(f)|^2)$, x_1 and x_2 are similarly defined and ε is assumed to be Gaussian. Hence, we can also write this relationship in terms of a distribution over the mixed speech features y as

$$p(y|x_1,x_2) = N(y; x_1 + \ln(1 + \exp(x_2 - x_1)), \psi) \tag{7}$$

where ψ is the variance of ε, and $N(y|\mu,\psi)$ denotes a normal density function in y with mean μ and variance ψ.

The transformations that we have applied to the model above are the same as the first steps in the calculation of the Mel frequency cepstrum features with the exception that we did not perform the Mel-scale warping before applying the log transform.

For the purpose of signal reconstruction, we are interested in likely values of the two composite signals, given the noisy speech. By recasting this relationship in terms of a likelihood $p(y|x_1,x_2)$, and using prior models for the two signals $p(x_1)$ and $p(x_2)$, we can arrive at a posterior distribution for the joint distribution $p(x_1,x_2|y)$ from which we can easily get the posterior distributions for the component signals $p(x_1|y)$ and $p(x_1|y)$. This will be described in the next section.

By inverting the procedure described above we can reconstruct an estimate of each signal. To do this we find the MMSE estimate for the signal $\hat{x_1}$ and calculate the inverse Fourier transform

$$\hat{x_1}[t] = IFFT(\exp(\hat{x_1}) \cdot \angle Y) \tag{8}$$

where $\hat{x_1} = \int x_1 p(x_1|y) dx_1$. The same is done for $\hat{x_2}$. In this reconstruction, we have used the original phases from the mixed signal.

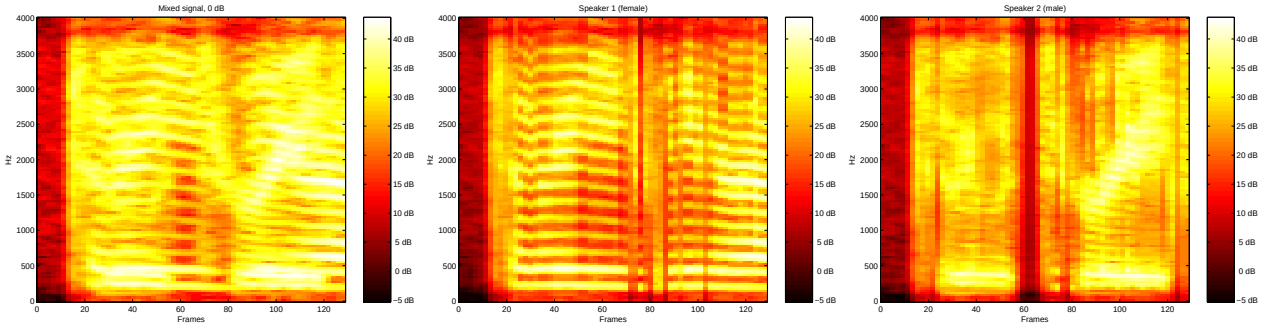
2.1. Inference

We now turn our attention to the procedure for estimating the posterior for the clean speech log-magnitudes $p(x_1|y)$. For this we employ the Algonquin method. Extensive evaluations of this framework have been performed in the context of robust speech recognition. In previous work, speech and noise models have either been in the "low-resolution" log-Mel-spectrum domain, or in the truncated cepstrum domain. Here we briefly outline the Algonquin procedure. Detailed discussions can be found in [5].

At the heart of the Algonquin method is the approximation of the posterior $p(x_1,x_2|y)$ by a Gaussian.

The true posterior

$$p(x_1,x_2|y) \propto p(y|x_1,x_2)p(x_2)p(x_1) \tag{9}$$



(a) Spectrogram of mixed signal

(b) Reconstructed spectrogram for speaker 1 (female).

(c) Reconstructed spectrogram for speaker 2 (male).

Fig. 2. (a) The spectrogram of the mixed signal. (b) The reconstructed spectrogram for signal 1 (female). (c) The reconstructed spectrogram for signal 2 (male).

is non-Gaussian, due to the non-linear relationship in Eqn. (6). In Eqn. (9) $p(x_1)$ is the model for the first speaker, $p(x_2)$ is the model for the second speaker, and $p(y|x_1,x_2)$ is the likelihood function from Eqn. (7).

We use a mixture of Gaussians to model both speech signals. Hence

$$p(x_1) = \sum_{s_1} p(s_1)p(x_1|s_1) = \sum_{s_1} \pi_{s_1} N(x_1|\mu_{s_1}^{x_1}, \Sigma_{s_1}^{x_1}) \qquad (10)$$

and similarly for $p(x_2)$. The construction of the speech models will be discussed below.

Due to the non-linear relationship between x_1 and x_2 for a given y, the true posteriors $p(x_1,x_2|y)$ is non-Gaussian. We wish to approximate this posterior with a Gaussian posterior. The first step is to linearize the relationship between y, x_1 and x_2.

For notational convenience, we write the stacked vector $z = [x_1^T\, x_2^T]^T$ and we introduce the function $g(z) = x_1 + \ln(1+\exp(x_2 - x_1))$.

If we linearize the relationship of Eqn. (6) using a first order Taylor series expansion at the point z_0, we can write the linearized version of the likelihood

$$p_l(y|x_1,x_2) = p_l(y|z) = N(y; g(z_0) + G(z_0)(z-z_0), \Psi) \qquad (11)$$

where z_0 is the linearization point and $G(z_0)$ is the derivative of g, evaluated at z_0. We can now write a Gaussian approximation to the posterior for a particular speech and noise combination as

$$p_l(x_1,x_2,y|s_1,s_2) = p_l(y|x_1,x_2)p(x_1|s_1)p(x_2|s_2) \qquad (12)$$

It can be shown[5] that the $p(x_1,x_2|y,s_1,s_2)$ is jointly Gaussian with mean

$$\eta_s = \Phi_s \left[\Sigma_s^{-1}\mu_s + G^T\Psi^{-1}(y-g-Gz_0)\right] \qquad (13)$$

and covariance matrix

$$\Phi_s = \left[\Sigma_s^{-1} + G^T\Psi^{-1}G\right]^{-1} \qquad (14)$$

and the posterior mixture likelihood $p(y|s_1,s_2)$ can be shown to be

$$\gamma_s = |\Sigma_s|^{-1/2}|\Psi|^{-1/2}|\Phi_s|^{1/2} \cdot \exp\Big[-\frac{1}{2}(\mu_s^T\Sigma_s^{-1}\mu_s + (y-g+Gz_0)^T\Psi^{-1}(y-g+Gz_0) - \eta_s^T\Phi_s^{-1}\eta_s)\Big].$$

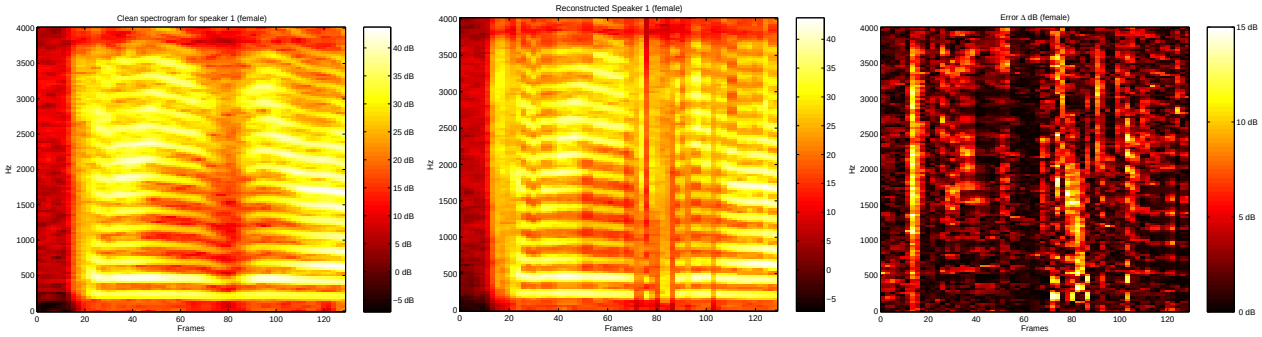
The choice of the linearization point is critical to the accuracy of the approximation. Ideally, we would like to linearize at the mode of the true posterior. In the Algonquin algorithm, we attempt to iteratively move the linearization points towards the mode of the true posterior. In iteration i of the algorithm, the mode of the approximate posterior in iteration $i-1$, μ_{i-1} is used as a linearization point of the likelihood, i.e. $z_i = \mu_{i-1}$. The algorithm converges in 3-4 iterations.

3. EXPERIMENTS

As mentioned above, we use Gaussian mixture models (GMM) to model the speakers. We trained two speaker independent gender dependent models. Each model had 512 mixtures of 128 dimensions. The training set was the clean training set from the Aurora 2 robust speech recognition data set.

Exact inference of a single frame of speech requires the evaluation of every combination of the female and male speaker models. As each models contains 512 mixtures the number of combinations that must be evaluated is 262144. Each combination requires 3-5 iterations in 128 dimensions. Hence, exact inference has complexity $O(m \cdot n \cdot d \cdot i)$ where m is the number of mixtures in speaker model 1, n is the number of mixtures in speaker model 2, d is the number of dimensions (frequency bins) and i is the number of iterations of the algorithm. The computational complexity is therefore considerable.

The test set was constructed from the Aurora 2 test-set. This set contains files with spoken digits, sampled at 8k Hz. Files from test Set A were mixed together at equal signal powers (i.e. 0 dB SNR). Log spectrum feature vectors were computed using an analysis window of 25 ms and a frame shift of 10 ms.



(a) Spectrogram of original signal 1 (female)

(b) Spectrogram for reconstructed speaker 1 (female).

(c) Absolute dB error.

Fig. 3. (a) The spectrogram for the original clean female signal. (b) The restored spectrogram for speaker 1 (female). (c) The error in absolute dB. The scale has been changed to make the errors clearer. Note that the errors do not suggest that the male speaker is substantially present in the spectrogram.

We ran the algorithm on 17 files from this test set[2]. No adaptation due to differences in signal gain was done as utterances in the training and test sets have similar signal levels.

Figure 2(a) shows a spectrogram for a portion of a file from the test set. Figure 2(b) shows the spectrogram for the separated female signal and Figure 2(c) shows the spectrogram for the separated male signal. Notice that the characteristics of the male and female spectrograms are different, where the fundamental frequency of the female speaker is higher, and the harmonics are spaced further apart. Notice also that the harmonics of the male signal are not clearly visible. This may be to due aliasing and may be reduced by lengthening the analysis window.

Figure 3(a) shows the spectrogram for the original female component signal. Compare this to the spectrogram for the recovered signal in Figure 3(b). The absolute dB errors are shown in Figure 3(c). The error plot shows that very little of the male speaker remains in the female signal.

The average average gain in SNR was 6.59 dB for the separated female signal, and 5.51 dB for the separated male signal. The separation of the female signal is better on average than the male signal. Interestingly, the model works best when there is complete overlap. In low energy frames of the female signal the male signal tends to leak into the separated female signal, but not vice versa.

The acoustic quality of the separated signals is impressive given the difficulty of the task. The suppression of the unwanted speaker in the restored signal is substantial, and is often barely audible[3]. The suppression of the unwanted speaker is better than the above numbers suggest, as the algorithm also introduces some distortion.

4. DISCUSSION AND FUTURE WORK

The male-male and female-female cross-talking scenarios require that the two speaker models be the same. As this is a symmetrical problem, the components that generate an observation may be correctly identified, but without temporal dependencies, we cannot associate the components through time. The complexity of the inference problem is not substantially increased introducing time dynamics, however the estimation of speaker models is more involved. We are currently exploring ways to do this.

We are also pursuing approximate inference techniques that promise orders of magnitude reduction in computational complexity without resorting to sub-optimal factorizations or mixing approximations.

In this paper we have proposed a new method for the cross-talker source separation task, that relies on strong high frequency resolution models of speech. We provide a test set based on the Aurora 2 test set and give quantitative results for a portion of this set. The acoustic quality of the results is impressive for this new method.

[2]The scripts to generate the test-set can be found at http://laplace.uscd.edu/∼jhershey or by contacting the authors.

[3]Audio samples can be found at http:// research.microsoft.com/users/traustik/.

5. REFERENCES

[1] Anthony J. Bell and Terrence J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995.

[2] Hagai Attias, "Independent factor analysis with temporally structured sources," in *Advances in Neural Information Processing (NIPS)*, 1999.

[3] J. Hershey and M. Casey, "Audio-visual sound separation via hidden markov models," in *Advances in Neural Information Processing Systems 14*, T. G. Dietterich, S. Becker, and Z. Ghahramani, Eds., Cambridge, MA, 2002, pp. 1173–1180, MIT Press.

[4] S. T. Roweis, "Factorial models and refiltering for speech separation and denoising," *Eurospeech*, 2003.

[5] T. Kristjansson, *Speech Recognition in Adverse Environments: A Probabilistic Approach*, Ph.D. thesis, University of Waterloo, Waterloo, Ontario, Canada, April 2002.

[6] T. Kristjansson and J. Hershey, "High resolution signal reconstruction," in *In Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2003)*, 2003.

[7] Hagai Attias, "New em algorithms for source separation and deconvolution," in *Proc. ICASSP*, 2003.

[8] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," Tech. Rep., Department of Computer Science, University of Sheffield, 1999.