MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Latent Variable Decomposition of Spectrograms for Single Channel Speaker Separation

Bhiksha Raj, Paris Smaragdis

TR2005-137 October 2005

Abstract

In this paper we present an algorithm for the separation of multiple speakers from mixed single-channel recordings by latent variable decompoistion of the speech spectrogram. We model each magnitude spectral vector in the short-time Fourier transform of a speech signal as the outcome of a discrete random process that generates frequency bin indices. The distriction of the process is modelled a mixture of multinomial distributions, such that the mixture weights of the component multinomials vary from analysis window to analysis window. The component multinomials are assumed to be speaker specific and are learnt from taining signals for each speaker. The distributions representing magnitude spectral vectors for the mixed signal are decomposed into mixtures of the multinomials for all component speakers. The frequency distribution, i.e. the spectrum for each speaker is reconstructed from this decomposition. Experimental results show that the proposed method is very effective at separating mixed signals.

WASPAA 2005

LATENT VARIABLE DECOMPOSITION OF SPECTROGRAMS FOR SINGLE CHANNEL SPEAKER SEPARATION

Bhiksha Raj, Paris Smaragdis

Mitsubishi Electric Research Laboratories, Cambridge, MA, U.S.A.
bhiksha@merl.com, paris@merl.com

ABSTRACT

In this paper we present an algorithm for the separation of multiple speakers from mixed single-channel recordings by latent variable decomposition of the speech spectrogram. We model each magnitude spectral vector in the short-time Fourier transform of a speech signal as the outcome of a discrete random process that generates frequency bin indices. The distribution of the process is modelled a mixture of multinomial distributions, such that the mixture weights of the component multinomials vary from analysis window to analysis window. The component multinomials are assumed to be speaker specific and are learnt from training signals for each speaker. The distributions representing magnitude spectral vectors for the mixed signal are decomposed into mixtures of the multinomials for all component speakers. The frequency distribution, *i.e.* the spectrum for each speaker is reconstructed from this decomposition. Experimental results show that the proposed method is very effective at separating mixed signals.

1. INTRODUCTION

In this paper we present a new technique for separating the signals for individual speakers from mixed single-channel recordings.

The problem of separating speakers from mixed monaural recordings has historically been approached from the angle of frequency selection. To separate the signal for any speaker, the time-frequency components of the mixed signal that are dominated by the speaker are selected and signals are reconstructed from the resulting incomplete time-frequency representation. The actual selection of time-frequency components for any speaker may be based on perceptual principles e.g. [1], or on statistical models, e.g. [2], and may either be binary or probabilistic, e.g. [3].

In this paper we follow an alternate approach that attempts to construct entire spectra for each of the speakers, rather than partial spectral descriptions. Typically, in this approach, characteristic spectro-temporal structures, or "bases", are learnt for the individual speakers from training data. Mixed signals are decomposed into linear combinations of these bases. Signals for individual speakers are separated by recombining their bases with the appropriate weights. Jang et. al. [4] derive the bases for speakers through independent component analysis (ICA) of their signals. Smaragdis [5] derives them through non-negative matrix factorization (NMF) of their magnitude spectra. Other authors have derived bases through vector quantization, Gaussian mixture models, etc.

The algorithm presented in this paper identifies typical spectral structures for speakers through latent-variable decomposition of their magnitude spectra. It is based on a somewhat unconventional statistical model that assumes that the spectral vectors of speech are the outcomes of a discrete random process that generates frequency bin indices. By this model, each analysis window (which we refer to as a "frame") of the speech signal represents several draws from this process. The magnitude spectrum for the frame represents a scaled histogram of the draws. The distribution of the random process itself is modelled as a mixture multinomial distribution. The mixture weights of the component multinomials in the mixture are assumed to vary from frame to frame; however the component multinomials themselves are assumed to be fixed for any speaker. The component multinomials for each of the speakers are learned from clean (unmixed) signals through an EM algorithm.

The spectrum of a mixed signal is modelled as the histogram of repeated draws from a two-level discrete random process, *i.e.* a process with two latent variables. By this model, within each draw the random process first draws a speaker from the mixture, and then a specific multinomial distribution for the speaker, and finally a frequency index from the multinomial. The component multinomial distributions for each speaker are known *a priori*, having been learnt from training data. The technique is therefore a supervised one, since the actual identities of the speakers in the mixed signal as well as *a priori* knowledge of the component multinomial distributions is required. In order to separate the spectrum for each speaker, maximum likelihood estimates of the mixture weights of all component multinomials and the *a priori* probabilities of the speakers are obtained for each frame. The separated spectrum for the speaker within the frame is finally obtained as the expected value of the number of draws of each frequency index from the mixture multinomial distribution for the speaker. Experiments show that the proposed algorithm can result in a significant degree of separation, as measured by the SNR of signals separated from synthetic mixtures. SNR enhancements of up to 6dB are obtained by the procedure.

The rest of this paper is arranged as follows: In Section 2 we briefly describe the basic latent-variable model for magnitude spectra. In Section 3 we describe EM algorithms for learning multinomial component distributions for speakers and for separation of mixed signals. In Section 4 we present some experimental results. Finally in Section 5 we discuss our results, possible extensions of the model and avenues for future work.

2. The Latent Variable Model

At the outset it is assumed that all speech signals are converted to sequences of magnitude spectral vectors[1] (simply referred to as "spectral" vectors henceforth) through a short-time Fourier transform. The term "frequency" in the subsequent discussion actually refers to the frequencies represented in these spectral vectors.

The latent variable model for the spectral vectors for the signal is best illustrated through the urn-and-ball example of Figure 1a. A stochastic "picker" has a number of urns. Every urn contains a number of balls, each of which is marked with one of *N* frequency

1. It has empirically been determined that the algorithm is more effective when performed with magnitude spectra rather than power spectra.

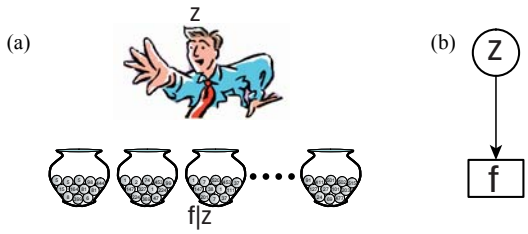

Figure 1. (a) Urn-and-ball illustration of the latent variable model for the magnitude spectrum of a speaker. An "picker" z selects one of several urns at random and draws a ball f from the urn. He repeats this operation several times in every frame. The histogram of the balls drawn represents the spectrum for the frame. (b) Graphical model for the process - latent variable z governs the probability of the observed variable f.

values. Each urn contains a different distribution of balls. The picker randomly selects one of the urns, draws a ball from the urn, notes the frequency on it and returns the ball to the urn. He repeats the process several times, drawing a ball and noting the frequency marked on it each time. He finally plots a histogram of the draws.

The probability distribution of the balls from any urn in this example is a multinomial distribution. The overall distribution of the process is a mixture multinomial distribution. The histogram represents the outcome of a set of draws from this distribution.

The urn-and-ball model above is equivalent to our latent variable model for the spectrum of any frame of speech. The frequencies marked on the balls represent the discrete frequencies in the N-point FFT for the signal. The number of times a particular frequency is drawn by the picker represents the value of the spectrum at that frequency. The mixture multinomial distribution for the entire urn-and-ball process represents a statistical model for the spectrum of the analysis window, while the histogram of the draws represents the actual spectrum obtained.

The latent variable model assumes that the component multinomial distributions for any speaker remain constant across all analysis frames, while the mixture weights for the components vary from frame to frame. In terms of the urn-and-ball simile, this means that the set of urns remains the same for all frames; however the operator selects the urns according to a different distribution in every frame.

The latent variable model for the spectral vectors can be represented graphically as shown in Figure 1b. A latent variable z governs the generation of a frequency f. The conditional probabilities for f are assumed to be constant for any speaker; however the *a priori* probability of the latent variable z varies from analysis frame to analysis frame. Thus the overall mixture multinomial distribution model for the spectrum of the t^{th} frame is given by

$$P_t(f) = \sum_z P_t(z) P_s(f|z) \tag{1}$$

where $P_t(z)$ represents the *a priori* probability of z in the t^{th} frame and $P_s(f|z)$ represents the multinomial distribution of f given the latent variable z. f takes the values of the discrete frequencies of the FFT for the frame, while z takes on as many values as there are component multinomials in the distribution. The subscript s in $P_s(f|z)$ indicates that these terms are specific to the speaker. In urn-and-ball terms, this is equivalent to saying that each speaker is represented by a separate set of urns.

The latent variable model for the spectrum of a *mixed* speech signal is shown in Figure 2. A fraction of the spectral content in each frequency is derived from each speaker. Accordingly, the urn-and-ball equivalent for the mixed signal with two speakers is as shown in Figure 2a. At each draw a new entity, who we term a "caller" randomly calls out one of two pickers who in turn draws a ball from one of his urns. Each picker represents the process that generates the spectrum for one of the speakers and draws balls exclusively from the urns for that speaker. The probability with which the caller selects any of the pickers changes from frame to frame, as does the probability with which the pickers select the urns. Figure 2b shows a graphical representation for the model. An initial latent variable s representing a speaker selects a second latent variable z, which in turn determines the probability of the frequency selected. The constraint here is that z takes on a different set of values for each speaker. The overall distribution underlying the spectral vector for the t^{th} analysis frame is given by

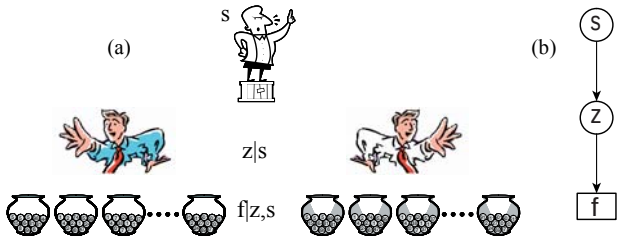

Figure 2. (a) Urn-and-ball illustration of the latent variable model for the spectrum of a mixed signal. A "caller" s randomly calls one of two pickers, who in turn randomly selects an urn and draws a ball from it. Pickers can draw balls from their own urns only and may not draw balls from the other picker's urns. The histogram of the drawn balls represents the spectrum of the mixed signal. (b) A graphical model for the process - an first latent variable s directs a second level latent variable z which in turn determines the probability of the observed variable f.

$$P_t(f) = \sum_s P_t(s) \sum_{z \in \{z_s\}} P_t(z|s) P_s(f|z) \tag{2}$$

where $P_t(s)$ is the *a priori* probability of the s^{th} speaker and $\{z_s\}$ represents the set of values that z can take for that speaker.

3. Single-Channel Speaker Separation

The speaker algorithm comprises a learning stage where the component multinomial distributions for speakers are learnt, and an operational stage where the learnt parameters are used to separate speech. We describe these in the following subsections.

3.1. Learning the parameters for speakers

In the learning stage of the algorithm the latent-variable-conditioned multinomial distributions $P_s(f|z)$ are learnt for each speaker from a set of training recordings for the speaker. Let $N_{t,f}$ represent the value of the f^{th} frequency band in the t^{th} training spectral vector for the speaker. Since the spectra are assumed to be histograms by the model, every spectral component must be an integer. To account for this we assume that the observed spectrum is in fact a scaled version of the histogram. Fortunately, the unknown scaling factor does not affect the analysis since it factored equally into the numerator and denominator terms of all equations.

The various components of the mixture multinomial distribution of Equation 1 are initialized randomly and reestimated through iterations of the following equations, which are derived through

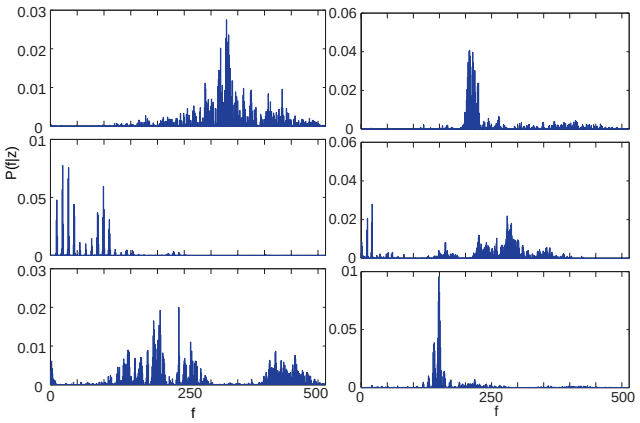

Figure 3. The three histograms to the left show typical component multinomial distributions obtained for a female speaker. The histograms to the right show typical multinomials for a male speaker.

the expectation maximization algorithm:

$$P_t(z|f) = \frac{P_t(z)P_s(f|z)}{\sum_{z'} P_t(z')P_s(f|z')} \tag{3}$$

$$P_s(f|z) = \frac{\sum_t P_t(z|f)N_{t,f}}{\sum_t \sum_{f'} P_t(z|f')N_{t,f'}} \tag{4}$$

$$P_t(z) = \frac{\sum_f P_t(z|f)N_{t,f}}{\sum_{z'} \sum_f P_t(z'|f)N_{t,f}} \tag{5}$$

Only the $P_s(f|z)$ values are used in reconstruction and the rest of the terms are discarded. Figure 3 shows a few examples of typical $P_s(f|z)$ distributions learnt for a female and a male speaker.

3.2. Separating out speakers from mixed signals

The process of separating out the power spectra of speakers from a mixed signal has two stages. In the first, the mixture multinomial distribution of each of the speakers is estimated in each analysis frame. This implies the estimation of all parameters of Equation 2 except the $P_s(f|s)$ terms which are obtained from the training data. The various $P_t(s)$ and $P_t(z|s)$ parameters for the t^{th} analysis frame are estimated by iterations of the following equations, derived through the EM algorithm:

$$P_t(s, z|f) = \frac{P_t(s)P_t(z|s)P_s(f|z)}{\sum_{s'} P_t(s') \sum_{z' \in \{z_{s'}\}} P_t(z'|s')P_{s'}(f|z')} \tag{6}$$

$$P_t(s) = \frac{\sum_{z \in \{z_s\}} \sum_f P_t(s, z|f)N_{t,f}}{\sum_{s'} \sum_{z \in \{z_{s'}\}} \sum_f P_t(s', z|f)N_{t,f}} \tag{7}$$

$$P_t(z|s) = \frac{\sum_f P_t(s, z|f)N_{t,f}}{\sum_{z' \in \{z_s\}} \sum_f P_t(s, z'|f)N_{t,f}} \tag{8}$$

Once all terms have been estimated, the mixture multinomial distribution for the s^{th} speaker in the t^{th} analysis frame is obtained as

$$P_t(f|s) = \sum_{z \in \{z_s\}} P_t(z|s)P_s(f|z) \tag{9}$$

According to the model, the total number of draws of any frequency is the sum of the draws from the distributions for the individual speakers, i.e.

$$N_{t,f} = \sum_s N_{t,f}(s) \tag{10}$$

where $N_{t,f}(s)$ is the number of draws of f from the s^{th} speaker. The expected value of $N_{t,f}(s)$, given the total count $N_{t,f}$ is hence given by

$$\hat{N}_{t,f}(s) = E[N_{t,f}(s)] = \frac{P_t(s)P_t(f|s)N_{t,f}}{\sum_{s'} P_t(s')P_t(f|s')} \tag{11}$$

$\hat{N}_{t,f}(s)$ is the estimated value of the f^{th} component of the spectrum of the s^{th} speaker in the t^{th} frame. The set of $\hat{N}_{t,f}(s)$ values for all values of f and t are composed into a complete sequence of spectral vectors for the speaker. The phase of the short-time Fourier transform of the mixed signal is combined with the reconstructed magnitude spectrum and an inverse Fourier transform performed to obtain the time-domain signal for the speaker.

4. Experimental Evaluation

Experiments were conducted to evaluate the speaker separation performance of the proposed algorithm on synthetic mixtures of signals from a male speaker and a female speaker. A set of 8 utterances comprising approximately 30 seconds of speech was used as training data for each speaker. All signals were normalized to 0 mean and unit variance to ensure uniformity of signal level. Signals were analyzed in 64 ms windows with 32ms overlap between windows. The training data thus comprised approximately 1000 magnitude spectral vectors for each speaker. Spectral vectors were modelled by a mixture of 100 multinomial distributions. Thus, a set of 100 multinomial distributions were learnt from the training data for each speaker.

Mixed signals were obtained by digitally adding test signals for both speakers. The length of the mixed signal was set to the shorter of the two signals. The component signals were all normalized to 0 mean and unit variance prior to addition, resulting in mixed signals with 0dB SNR for each speaker. A total of 20 such mixed recordings were obtained in this manner. The mixed signals were separated using the method outlined in Section 3.2.

Figure 4 shows an example of spectrograms of separated signals obtained for the speakers. The spectrograms of the original signals, the mixed signal and both separated signals are all shown. It can be seen from the figure that considerable separation has been

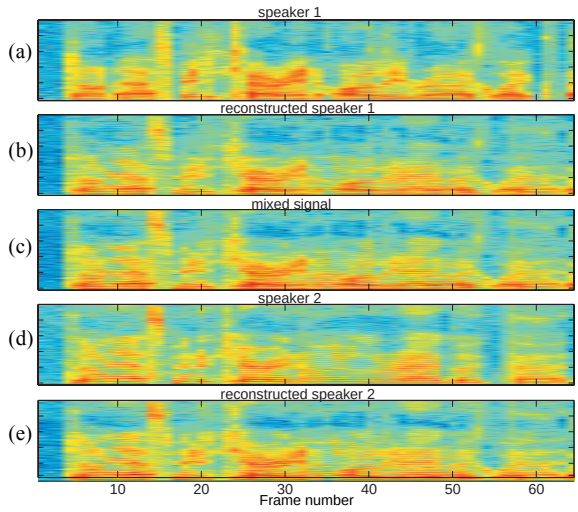

Figure 4. Example of the output of the separation algorithm. Both speakers uttered the same sequence of words in this example. (a) Original unmixed signal for the male speaker, (b) Signal separated for the male speaker from the mixed signal, (c) The mixed signal, (d) Original unmixed signal for the female speaker, (e) Separated signal for the female speaker.

achieved for both speakers. We measured the improvement in SNR for the two speakers. It is difficult to establish a clear definition of the SNR for this problem: since the power in any frequency component of the reconstructed signal is often lower than the power in the corresponding original unmixed signal for one of the two speakers, direct subtraction of the former from the latter to determine noise power can result in negative estimates for noise. We therefore incorporate the phase of the Fourier spectrum of the mixed signal into both the original unmixed signal and the reconstructed signal to normalize their phases and estimate the SNR as:

$$SNR(s) = 10\log 10\left(\frac{\sum_{f,t} N_{t,f}(s)}{\sum_{t,f}\left|\sqrt{N_{t,f}(s)}e^{i\Phi_{t,f}(s)} - \sqrt{\hat{N}_{t,f}(s)}e^{i\Phi_{t,f}(s)}\right|}\right) \quad (12)$$

where $\Phi_{t,f}(s)$ is the phase of the f^{th} frequency band of the fourier spectrum of the t^{th} analysis frame of the mixed signal. Although Equation 12 avoids the problem of negative noise estimates, it is still not perfect and while it is safe to use it to judge the relative level of corruption in two signals, it is not a perfect measure of the absolute degree of corruption in a single signal.

Using the description of SNR given above, the separation method presented in this paper was observed to result in an average SNR improvement of 5.30 dB over the mixed signal for the female speaker and of 5.36dB for the male speaker, averaged over 20 mixed recordings. Figure 5 shows a plot of the SNR improvements for all 20 recordings. Examples of separated signals can be obtained at http://www.cs.cmu.edu/~bhiksha/audio/

5. Observations and Conclusions

The proposed speaker separation algorithm is observed to be able to extract separated signals with significantly reduced levels of the competing speaker. Of particular interest are that the algorithm requires very small amounts of training data and is also computationally significantly less expensive than most other separation

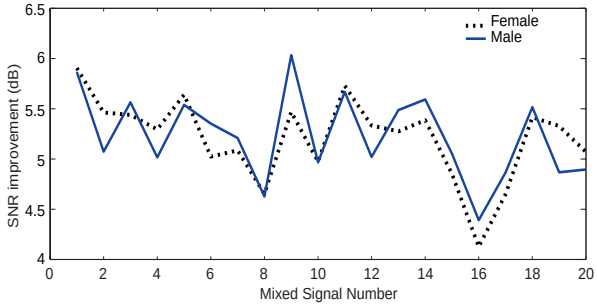

Figure 5. SNR improvement achieved for the two speakers in each of 20 mixed signals from a male and a female speaker.

algorithms that produce similar results. In our experiments the results obtained with our algorithm were perceptually superior to those obtained with MaxVQ [2], although the latter was trained on 30 times as much training data per speaker, and took significantly longer to run (SNR comparisons are difficult to make, however, since MaxVQ only derives partial spectral representations).

Algebraically, the proposed algorithm is very similar to the NMF-based speaker separation method of Smaragdis [5]. The mathematical details, however, differ significantly - a closer simile may be in fact be made to the PLSA algorithm proposed by Hoffman [6]. More importantly, the proposed approach naturally enables clean solutions to various extensions of the basic separation problem. For instance, *a priori* guesses about the relative levels of the two speakers can be incorporated through *a priori* probabilities for $P(s)$. Other work not presented in this paper shows that short-time linear filtering effects are easily incorporated by including an alternative conditioning link from the speaker to the frequencies in Figure 2b. Temporal dependence between adjacent frames may be incorporated through Markovian priors on $P(z|s)$. All of these extensions are easily solved through the EM or belief propagation algorithms. In addition, the formulation also enables the enforcement of sparseness constraints through various well-known methods such as minimum-entropy training or through model-order-estimation criteria such as MDL or BIC. We expect to address several of these issues in future papers.

REFERENCES

1. Van der Kouwe, J. W., Wang, D., Brown G. J. (2001). "A Comparison of Auditory and Blind Separation Techniques for Speech Segregation", IEEE Trans. on Speech and Audio Processing, Vol. 9, No. 3, Mar 2001.

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

3. Reddy, A.M., Raj, B. (2004). "Soft mask estimation for single channel speaker separation", ISCA ITRW on statistical and perceptual audio processing (SAPA2004), Jeju, Korea, 2004.

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

5. Smaragdis, P. (2004). "Non-negative Matrix Factor Deconvolution; Extraction of Multiple Sound Sources from Monophonic Inputs", International Congress on Independent Component Analysis and Blind Signal Separation, September 2004.

6. Hoffman, T. (2001). "Unsupervised learning by probabilistic latent semanting analysis", Machine Learning, vol. 42, pp. 177–196, 2001.