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

Latent Dirichlet Decomposition for Single Channel Speaker Separation

Bhiksha Raj, Madhusudana V.S. Shashanka, Paris Smaragdis

TR2006-064   May 2006

Abstract

We present an algorithm for the seaparation 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 modeled as 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 learned from training signals for each speaker. We model the prior distribution of the mixture weights for each speaker as a Dirichlet distribution. 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.

IEEE International Conference on Acoutiscs, Speech and Signal Processing (ICASSP)

LATENT DIRICHLET DECOMPOSITION FOR SINGLE CHANNEL SPEAKER SEPARATION

Bhiksha Raj [1], *Madhusudana V. S. Shashanka* [2], *Paris Smaragdis* [1]

[1] Mitsubishi Electric Research Labs, 201 Broadway, Cambridge MA 02139
[2] Boston University Hearing Research Center, 677 Beacon St, Boston MA 02215

ABSTRACT

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 modeled as 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 learned from training signals for each speaker. We model the prior distribution of the mixture weights for each speaker as a Dirichlet distribution. 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.

1. INTRODUCTION

The problem of monaural speaker separation, *i.e.* the problem of separating concurrent speakers[1] from a mixture of speakers in a monaural recording has historically been approached from the angle of frequency selection. To separate the signal for any speaker, the time-frequency components of the mixed signals that are dominated by the speaker 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 be either 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 learned for the individual speakers from training data. Mixed signals are decomposed into linear combinations of these learned bases. Signals for individual speakers are separated by recombining their bases with appropriate weights. Jang et al [4] derive the bases for speakers through independent component analysis of their signals. Smaragdis [5] derives them through non-negative matrix factorization of their magnitude spectra. Others 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. The latent-variable model for speaker separation, originally proposed by Raj et al [6], assumes that spectral vectors of speech are the outcomes of a discrete random process that generates frequency bin indices. Each analysis window (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 modeled as a mixture multinomial distribution. The mixture weights are assumed to vary from frame to frame while the component multinomials, which are speaker specific, are assumed to be fixed across all frames.

In this framework, the component multinomials may be interpreted as the fundamental modes, or bases that a speaker is able to generate. The spectral magnitude of any analysis window is a (noisy) linear combination of these bases. In the original formulation Raj et al. [6] all linear combinations are assumed equally likely *a priori*, *i.e.* any valid set of mixture weights is as likely as any other set. In this paper we recognize that speakers have biases: they favor some sounds over others. We capture this bias through an *a priori* probability on the mixture weights, that we model by a Dirichlet density. The weights with which component multinomials are combined in any analysis frame are themselves drawn from this density. The parameters of the multinomials and the Dirichlet density are learned from unmixed signals for each speaker using the EM algorithm. The algorithm is thus a supervised one, since the identities of the speakers and the parameters of their distributions must be known.

The spectrum of a mixed signal is modeled as the outcome of repeated draws from a two-level random process. Within each draw, the process first draws a speaker from the mixture, then a specific multinomial for the speaker, and finally a frequency index from the multinomial. To separate the spectrum for each speaker within each analysis frame we obtain maximum *a posteriori* estimates of the mixture weights for each speaker, given the *a priori* probability distribution on the weights and the speaker-specific multinomial component distribution that were learned from training data. 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.

The rest of the paper is organized as follows: In section 2, we briefly describe the latent Dirichlet variable model for magnitude spectra. In section 3, we describe the 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 the results and possible extensions of this work.

2. THE LATENT DIRICHLET VARIABLE MODEL

At the outset it is assumed that all speech signals are converted to sequences of magnitude spectral vectors (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.

[1] The term *speaker* here refers to a person speaking, *i.e.* a talker

The latent Dirichlet variable model is a generative probabilistic model that is an adaptation of latent Dirichlet allocation [7].

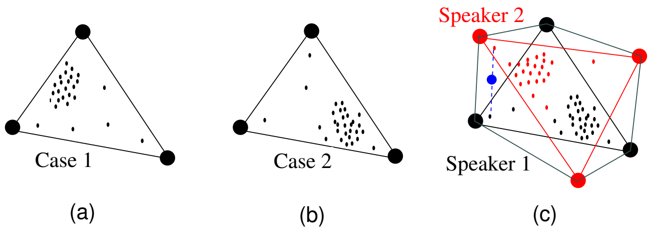

Fig. 1. Illustration of the latent Dirichlet variable model simplex. A Triangle denotes a simplex where each corner represents one of the component multinomials and each point within the simplex represents the mixture multinomial model for the spectrum of one frame of speech. (a) and (b) shows the case of two speakers whose simplexes are similar while the distributions are different. (c) shows the model for a mixed signal where each inner triangle corresponds to the simplex of a different speaker and the outer polygon represents the distribution of the mixed signal.

The model assumes that each spectral vector of a speech signal is the result of several draws from a discrete random process that generates frequency bin indices. The generative process for each spectral vector can be described as follows:

- Let $\boldsymbol{\theta}$ be a K-dimensional Dirichlet random variable that takes values in the $(K-1)$ simplex (a k-vector $\boldsymbol{\theta}$ lies in the $(k-1)$ simplex if $\theta_i \geq 0$, $\sum_{i=1}^{k} \theta_i = 1$) and has the following probability density

$$p(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{i=1}^{K} \alpha_i)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \theta_i^{\alpha_1 - 1} \ldots \theta_K^{\alpha_K - 1} \tag{1}$$

  where the parameter $\boldsymbol{\alpha}$ is a K-vector with components $\alpha_i > 0$ and $\Gamma(.)$ is the Gamma function. Generate an observation of $\boldsymbol{\theta}$.

- Generate several draws from a mixture multinomial whose mixture weights are defined by $\boldsymbol{\theta}$:

  - Let z be a variable that takes values $\{1, 2, \ldots K\}$. Generate a value of z from the probability distribution defined by the vector $\boldsymbol{\theta}$, i.e.

$$p(z = k) = \theta_k \tag{2}$$

  - Let $\boldsymbol{\beta}$ be a $K \times F$ matrix describing frequency probabilities, where F is the number of discrete frequencies in the FFT. The ij-th element of the matrix β_{ij} is the probability of drawing frequency j when the hidden variable z takes the value i, i.e.

$$\beta_{ij} = p(f = j|z = i) \tag{3}$$

    Generate a value of the frequency using the multinomial distribution given by the k-th row of $\boldsymbol{\beta}$, where k is the value of z generated in the previous step.

Thus, the overall mixture multinomial distribution model for a given frame of speech can be written as

$$p(f) = \sum_{k=1}^{K} \theta_k^s \beta_{kf}^s \tag{4}$$

where $\boldsymbol{\theta}^s$ has a prior Dirichlet distribution with parameter vector $\boldsymbol{\alpha}^s$. The superscript s indicates that the terms are specific to the speaker.

Equation 4 represents a multinomial distribution whose parameters lie entirely within a simplex, the corners of which lie at component multinomials that form the rows of $\boldsymbol{\beta}$. This is illustrated by figure 1(a): each corner of the triangle represents one of the component multinomials, and each point within the simplex represents the mixture multinomial model for the spectrum of one frame of speech. Both the simplex and the distribution of points within it are specific to a speaker. In particular, the distribution of points in the simplex can distinguish speakers even when their simplexes are very similar: talkers tend to have a bias towards uttering certain kinds of sounds and this shows up in the scatter of points in the simplex. This is illustrated by figure 1(b) which shows the simplex for a different speaker . This simplex is identical to the one in figure 1(a), except for the scatter of points within it. The latent-variable model of Raj et al. [6] ignores the distribution of points within the simplex and cannot distinguish between two figures. The Dirichlet variable model proposed in this paper, on the other hand, models the distribution of points within the simplex by the Dirichlet distribution over $\boldsymbol{\theta}$ and is thus able to distinguish between the two cases.

The latent Dirichlet variable model for the spectrum of a *mixed* speech signal has an additional level in the hierarchy. A fraction of the spectral content in each frequency is derived from each speaker. Hence, an initial latent variable s first selects a speaker and then a frequency is selected according the generative model for that particular speaker. The overall distribution for the spectral vector is given by

$$p(f) = \sum_{s} p(s) \sum_{k=1}^{K} \theta_k^s \beta_{kf}^s \tag{5}$$

where $p(s)$ is the *a priori* probability of the s-th speaker.

Figure 1(c) illustrates the model for the spectrum of a mixed speech signal. Each triangle represents the simplex for a different speaker (shown by distinct colors). The outer simplex shows the distribution for the mixed signal. A mixed spectrum is represented by a point within this outer simplex, e.g. the blue dot in the figure. This is a linear combination of two points, one lying within the simplex for each speaker (illustrated by the dotted line and the dots the end of the line in the figure). The goal of the separation is to identify the ends of the line, given that the line is of unit length. This is aided greatly by knowing the *a priori* distribution of points within the simplexes of the speakers.

3. SINGLE CHANNEL SPEAKER SEPARATION

The algorithm comprises a learning stage where the component multinomial distributions for speakers are learned, and a separation stage where the learned parameters are used to separate speech.

3.1. Learning the parameters for speakers

In the learning stage, the multinomial distributions $\boldsymbol{\beta}^s$ and the Dirichlet parameter vector $\boldsymbol{\alpha}^s$ are learned for each speaker from a set of training recordings for the speaker. Let $O_{f,t}$ represent the value of the f-th frequency band in the t-th spectral vector. Let $\theta_{k,t}$ represent the value of θ_k that has been estimated for the t-th spectral vector. Since the spectra are assumed to be histograms in 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. However, the unknown scaling factor does not affect the analysis since it is factored equally into the numerator and denominator terms of equations 7 and 8.

The terms of equation 4 are initialized randomly and re-estimated through iterations of the following equations, which are derived through the expectation maximization algorithm:

$$p_t(z|f) = \frac{\theta_{z,t}\beta^s_{zf}}{\sum_{z'}\theta_{z',t}\beta^s_{z'f}} \tag{6}$$

$$\beta^s_{zf} = \frac{\sum_t p_t(z|f)O_{f,t}}{\sum_t\sum_{f'} p_t(z|f)O_{f',t}} \tag{7}$$

$$\theta_{z,t} = \frac{\sum_f p_t(z|f)O_{f,t}}{\sum_{z'}\sum_f p_t(z'|f)O_{f,t}} \tag{8}$$

The $\boldsymbol{\theta}$ values that have been estimated for all time frames are then used to estimate the Dirichlet parameter vector $\boldsymbol{\alpha}$ for the speaker. The maximum likelihood estimate of a Dirichlet distribution is not available in closed form. Hence, we use iterative methods (a fixed-point iteration or Newton-Raphson iteration) to obtain an estimate of $\boldsymbol{\alpha}$, see [8] for a detailed description. Figure 2 shows a few examples of typical β^s_{zf} distributions learned for a female and a male speaker.

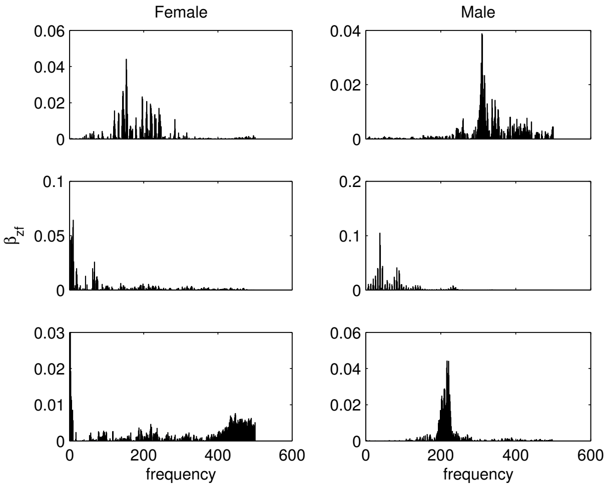

Fig. 2. The three histograms to the left show typical component multinomial distributions obtained for a female speaker. The histograms to the right show typical multinomials for a male speaker.

3.2. Separating speakers from mixed signals

The process of separating the spectra of speakers from a mixed signal has two stages. The parameters $p_t(s)$ and $\theta^s_{z,t}$ for the t-th analysis frame are estimated by iterations of the following equations, derived using EM algorithm:

$$p_t(s,z|f) = \frac{p_t(s)\theta^s_{z,t}\beta^s_{zf}}{\sum_{s'}p_t(s')\sum_{k=1}^{K}\theta^{s'}_{k,t}\beta^{s'}_{kf}} \tag{9}$$

$$p_t(s) = \frac{\sum_{k=1}^{K}\sum_f p_t(s,k|f)O_{f,t}}{\sum_{s'}\sum_{k=1}^{K}\sum_f p_t(s',k|f)O_{f,t}} \tag{10}$$

$$\theta^s_{z,t} = \frac{\sum_f p_t(s,z|f)O_{f,t}C + \alpha^s_z - 1}{\sum_{k=1}^{K}(\sum_f p_t(s,k|f)O_{f,t}C + \alpha^s_k - 1)} \tag{11}$$

Notice the presence of an unknown scaling factor C in equation 11. We empirically find a value of C so that the value of the first term $p_t(s,z|f)O_{f,t}C$ is balanced by the value of $(\alpha^s_z - 1)$ and neither term dominates the answer.

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

$$p_t(f|s) = \sum_{k=1}^{K}\theta^s_{k,t}\beta^s_{kf} \tag{12}$$

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.

$$O_{f,t} = \sum_s O_{f,t}(s) \tag{13}$$

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

$$\hat{O}_{f,t} = E[O_{f,t}(s)] = \frac{p_t(s)p_t(f|s)O_{f,t}}{\sum_{s'}p_t(s')p_t(f|s')} \tag{14}$$

$\hat{O}_{f,t}(s)$ 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 $\hat{O}_{f,t}(s)$ 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-term Fourier transform of the mixed signal is combined with the reconstructed spectrum and an inverse Fourier transform performed to obtain the time-domain signal for the speaker.

4. EXPERIMENTAL EVALUATION

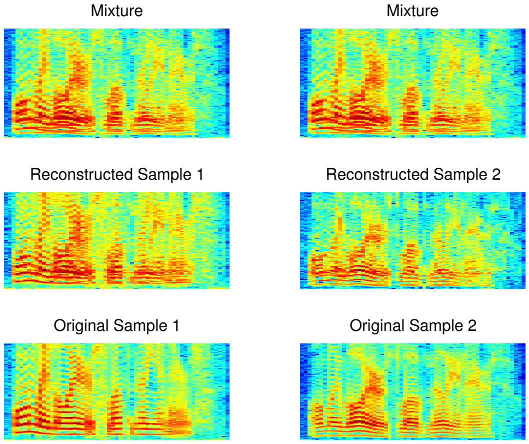

Fig. 3. Example of the output of the separation algorithm. Both speakers uttered the same sequence of words in this example. Spectrograms on the left column correspond to a female speaker while the ones on the right correspond to a male speaker (top row shows the mixture).

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 5 utterances from the TIMIT database comprising approximately 15 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 32 ms overlap between windows. Spectral vectors were modeled by a mixture of 100 multinomial distributions. Thus, a set of 100 multinomial distributions were learned 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. The mixed signals were separated using the method outlined in section 3.2. We empirically chose the value of the unknown scaling factor for equation 11 to be 10000.

Figure 3 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 shown. It can be seen from the figure that considerable separation has been achieved for both speakers. Examples of separated signals can be obtained at http://cns.bu.edu/∼mvss/courses/speechseg/.

It is difficult to obtain unambiguous objective measurements of performance for speaker separation algorithms. The primary problem is that the power in many components of the reconstructed spectra for a speaker is lesser than that in the spectrum of the unmixed signal, leading to negative noise estimates. As a result, several equally unsatisfactory metrics have been proposed. Reyes et al. compute the SNR from only those components where the primary speaker dominates. This however leads to unrealistic SNR estimates. Smaragdis [5] computes a cross correlation between the unmixed and reconstructed signals; however this metric leaves much of the energy in the signal unaccounted for. Raj et al. [6] impose the phase of the unmixed signal on the separated spectrogram and compute the SNR from the complex spectrum. Unfortunately, the SNR estimates are often not meaningful due to the phase distortion introduced by the procedure. Nevertheless, we have attempted to obtain an objective measurement of the separation performance achieved by our algorithm. We report two metrics: the first is the cross covariance between the magnitude spectrum of the original unmixed signal and the separated spectrum for a speaker. The better the separation, the higher this number will be. Unfortunately, the cross-covariance tends to be very high even for the mixed signal, and this metric is not very informative. As an alternative, we have also reported the (equally ineffective) SNR measurements obtained with the SNR estimator of Raj et al.

The table in figure 4 shows the SNR improvement from the mixture to the reconstructed signal and the values of cross covariance between the magnitude spectrum of the original unmixed signal and the spectrum of the mixture/reconstructed signal. The values are for five samples that are available on the website. There is an improvement in all cases for the female speaker with an average of 0.0520 while there is average improvement of 0.0022 in the case of the male speaker. We emphasize again that these numbers are not indicative of the perceptual strength of the separation. Ultimately, the only true method for evaluating separation performance is a subjective test. Subjective tests reveal that the separated signals obtained with our techniques consistently have higher levels of the desired speaker than in the mixture, particularly for the male speaker.

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. In addition, the algorithm has several advantages over most state-of-art techniques. It only requires very small amounts of training data - for the experiments reported in this paper only 15 seconds of training data were used per speaker. Additionally, the computational requirements for separation are minor - the separation of a mixture of 2 speakers can be done in real time on a standard laptop.

Female Speaker			Male Speaker		
Mixed sample	Reconstructed	SNR change	Mixed sample	Reconstructed	SNR change
0.7779	0.8341	2.5 dB	0.7878	0.7991	0.3 dB
0.7367	0.7751	2.8 dB	0.7683	0.7528	-0.2 dB
0.7481	0.8130	2.7 dB	0.7816	0.8018	1.2 dB
0.7374	0.7914	2.8 dB	0.7894	0.7888	1.2 dB
0.7365	0.7832	3.0 dB	0.7498	0.7457	-1.6 dB

Fig. 4. Normalized cross-covariance with the magnitude spectra of unmixed samples and the SNR improvement.

There are many avenues for further improvement in performance. The current model only employs simple Dirichlet densities as priors. Improvements can be gained by having more detailed models such as mixture Dirichlet densities. Temporal dependence between adjacent frames, which are currently being ignored, may be incorporated by using Markovian priors on $\boldsymbol{\theta}$. We expect to address several of these issues in future papers.

We conclude that simplicity of our algorithm, novelty of the approach and the scope for improvements makes it a very interesting method worthy of further research.

6. ACKNOWLEDGMENTS

M. V. S. Shashanka was supported in part by a grant from the Air Force Office of Scientific Research to Barbara Shinn-Cunningham.

7. REFERENCES

[1] J. W. Van der Kouwe, D. Wang, and G. J. Brown, “A comparison of auditory and blind separation techniques for speech segregation,” *IEEE Trans. on Speech and Audio Processing*, 2001.

[2] S. T. Roweis, “Factorial models and re-filtering for speech separation and denoising,” in *EUROSPEECH*, 2003, pp. 1009–1012.

[3] A. M. Reddy and B. Raj, “Soft mask estimation for single channel speaker separation,” in *ISCA SAPA2004*, Jeju, Korea, 2004.

[4] G-J Jang and T-W Lee, “A maximum likelihood approach to single channel source separation,” *Journal of Machine Learning Research*, 2003.

[5] P. Smaragdis, “Non-negative matrix factor deconvolution: Extraction of multiple sound sources from monophonic inputs,” in *Intl. Congress on ICA and Blind Signal Separation*, 2004.

[6] B. Raj and P. Smaragdis, “Latent variable decomposition of spectrograms for single channel speaker separation,” in *IEEE WASPAA*, New Paltz, NY, 2005.

[7] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *Jrnl. of Machine Learning Research*, pp. 993–1022, 2003.

[8] T. P. Minka, “Estimating a dirichlet distribution,” Tech. Rep., Microsoft Research, 2003.