LATENT DIRICHLET DECOMPOSITION FOR SINGLE CHANNEL SPEAKER SEPARATION

Bhiksha Raj, Paris Smaragdis

Mitsubishi Electric Research Labs
Cambridge MA 02139

Madhusudana V. S. Shashanka

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 modelled 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 learnt 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 separating speakers from mixed 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 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 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. It is based on a statistical model used by [6] that assumes that spectral vectors of speech are the outcomes of a discrete random process that generates frequency bin indices. By this model, 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 modelled as a mixture multinomial distribution. The mixture weights of the component multinomials are modelled to have a prior dirichlet distribution. The component multinomials are assumed to be fixed across frames for any speaker. The component multinomials and the prior dirichlet distributions for each speaker are learned from unmixed signals using iterative procedures.

The spectrum of a mixed signal is modelled as the histogram of repeated draws from a two-level discrete random process. Within each draw, the random process first draws a speaker from the mixture, then a specific multinomial distribution for the speaker, and finally a frequency index from the multinomial. The component multinomial distributions and the dirichlet distribution parameters 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.

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 tranform. the term "frequency" in the subsequent discussion actually refers to the frequencies represented in these spectral vectors.

The latent dirichlet variable model is a generalization of the latent variable model used by Raj et al [6]. It is a generative probabilistic model which is an adaptation of latent dirichlet allocation [7].

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 draw can be described as follows:

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

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

  Generate an observation of θ.

- 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 θ, i.e.

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

- Let β 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 β, where k is the value of z generated in the previous step.

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

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

where θ^s has a prior dirichlet distribution with parameter vector α^s. The superscript s indicates that the terms are specific to the speaker.

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 intial 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.

3. SINGLE CHANNEL SPEAKER SEPARATION

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

3.1. Learning the parameters for speakers

In the learning stage, the mutinomial distributions β^s and the dirichlet parameter vector α^s are learnt 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.

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

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

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

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

The θ values that have been estimated for all time frames are then used to estimate α for the speaker using an iterative procedure, see [8] for details. Figure 1 shows a few examples of typical β_{zf}^s distributions learnt for a female and 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 $\theta_{z,t}^s$ for the t-th analysis frame are estimated by iterations of the following

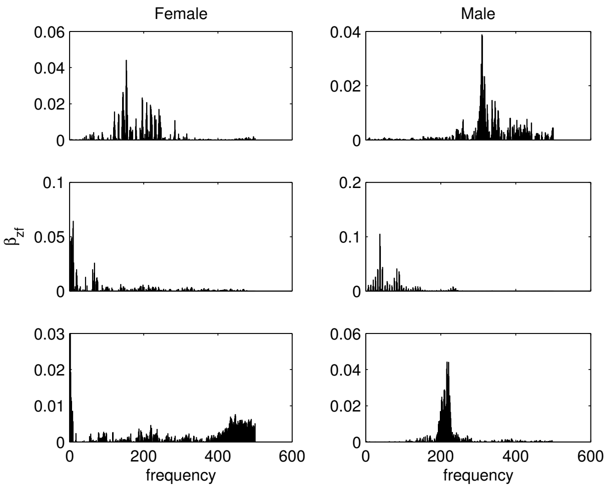

Fig. 1. Typical Component Multinomial Distributions

equations, derived using EM algorithm:

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

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

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

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.

Note: 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. The unknown scaling factor does not appear in equations 7, 8 and 10 since it is factored equally in the numerator and the denominator. However, it is present in equation 11 and we choose its value empirically.

4. EXPERIMENTAL EVALUATION

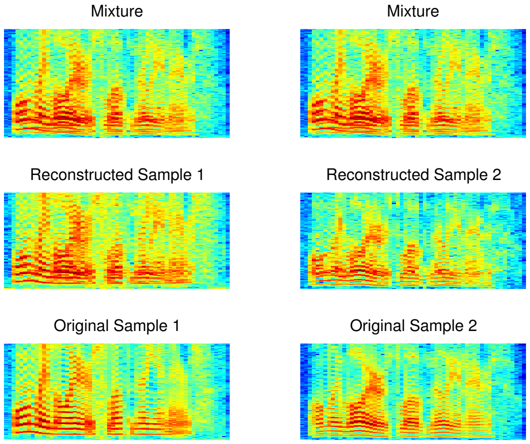

Fig. 2. Example of the output of the separation algorithm

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 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. 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 2 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/.

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.

The proposed algorithm, which is an adaptation of Latent Dirichlet Allocation (LDA, see [7]), is an extension and generalization of the method used by Raj et al [6]. Raj et al used the idea of probabilistic latent semantic indexing (PLSI, see [9]) but it has been shown ([10]) that PLSI is a *maximum a posteriori* estimated LDA model under a uniform dirichlet prior.

6. ACKNOWLEDGEMENTS

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*, 2003, pp. 1009–1012.

[3] A. M. Reddy and Bhiksha Raj, "Soft mask estimation for single channel speaker separation," in *ISCA ITRW on statistical and perceptual audio processing*, 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] Bhiksha Raj and Paris Smaragdis, "Latent variable decomposition of spectrograms for single channel speaker separation," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2005.

[7] David M. Blei, Andrew Y. Ng, and Michael I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, pp. 993–1022, 2003.

[8] P. Minka, Thomas, "Estimating a dirichlet distribution," Tech. Rep., Microsoft Research, 2003.

[9] T. Hoffman, "Unsupervised learning by probabilistic latent semantic indexing," *Machine Learning*, 2001.

[10] Mark Girolami and Ata Kaban, "On an equivalence between plsi and lda," in *SIGIR 2003*, Toronto, Canada, 2003.