A Minimum Mean Squared Error Estimator for Single Channel Speaker Separation

Aarthi M. Reddy, Bhiksha Raj
Mitsubishi Electric Research Laboratories
Cambridge, MA, 02139
amreddy, bhiksha@merl.com

Abstract

The problem of separating out the signals for multiple speakers from a single mixed recording has received considerable attention in recent times. Most current techniques are based on the principle of masking: in order to separate out the signal for any speaker, frequency components that are not believed to belong to that speaker are suppressed. The signals for the speaker are reconstructed from the partial spectral information that remains. In this paper we present a different kind of technique – one that attempts to estimate all spectral components for the desired speaker. Separated signals are derived from the complete spectral descriptions so obtained. Experiments show that this method results in superior reconstruction to masking based methods.

1. Introduction

Speaker separation refers to the problem of separating the speech signals of individual speakers from mixed recordings, obtained when the speakers speak simultaneously. Conventional solutions to the problem have assumed that the mixed signals have been jointly recorded over at least as many microphone channels as there are speakers. Speaker separation can be performed by one of a variety of methods, e.g. decorrelation algorithms that attempt to minimize the correlation between deconvolved signals [1], deconvolution methods that attempt to estimate the channel [2], and blind source separation techniques that perform independent component analysis (ICA) of the mixed signals [3]. These algorithms make very simple assumptions about the signals, the primary one being that the combined signals are statistically independent of each other.

A more interesting situation arises when the number of recorded signals is fewer than the number of speakers. The problem of separation becomes underconstrained, and explicit statistical knowledge about one or more of the speech signals must be utilized to effect separation, e.g. [4].

An extreme case of the above condition is the problem of single channel speaker separation, where the mixed speech signals have been recorded over only a single microphone. The degree of unconstrainedness becomes severe. Constraints must be enforced externally, such as through models of human perception, assumptions about the acoustic properties of speech, or through detailed statistical models of the speech signal.

Currently the most popular approach to single channel speaker separation combines detailed statistical models with the fact that the energy in the speech signal is unevenly distributed across frequencies, and that speech signals remain intelligible even when several time-frequency components have been excised from them. Roweis [5] models short-time Fourier transform representations of the various speakers by hidden Markov models (HMMs). The parameters of the HMM for any speaker are learnt from training data recorded from the speaker. In addition, Roweis assumes that the log energy in any frequency band of the mixed signal at any time can be attributed to only one of the speakers. This "log-max" assumption is justified by two observations. First, when two or more speakers speak simultaneously, at any time, any given frequency band is usually dominated by a single speaker. Second, in any given frequency band the disparity in the energy levels of the dominant speaker and the other speakers is such that the logarithm of the sum of the energies of the individual speakers can be well approximated by the logarithm of the energy of the dominant speaker. In order to reconstruct the signal for any speaker, Roweis estimates the mask for that speaker, i.e. the identity of the time-frequency locations where the speaker dominates. The entire signal is reconstructed entirely from the masked spectrum for the speaker, i.e. from the spectral components identified by the mask. The results achieved with this method are remarkably good.

Hershey et. al. [6] augment audio recordings with visual features, such as lip and facial movement, in order to enhance the separation. Additionally, they separate the signal into multiple frequency bands, which are then processed independently. As in Roweis’ algorithm, the signals for the individual speakers are reconstructed from masked spectra.

Several other techniques have been proposed along similar lines. All of them employ variants of the same basic themes: separation of the signal into frequency bands, and reconstruction of the final signal from masked spectra. As one may observe, all existing single channel speaker separation algorithms work by suppressing unreliable frequency bands, i.e. frequency bands that cannot reliably be identified with the target speaker, to a floor value and reconstructing the signal from only the reliable spectral components for the speaker.

In this paper, we present a minimum mean squared error estimation (MMSE) single channel speaker separation algorithm that takes a different approach: it attempts to estimate the frequency components for the target speaker, and reconstructs the signal using all spectral components, not merely the reliable ones. As in other single channel speaker separation algorithms, detailed statistical characterizations of the spectra of the individual speakers is required. The distributions of the log spectra of the individual speakers are modeled by mixture Gaussian densities. For simplicity, the mixed recording is assumed to obey Roweis’ log-max approximation. Under these assumptions, an exact minimum mean squared estimator can be obtained for the spectra of the individual speakers. The signals for the individual speakers are reconstructed from the estimated spectra.
The rest of this paper is organized as follows. In Section 2 we discuss the mixing model used to describe the spectral properties of the mixed signal. In Section 3 we describe the statistical model we use. In Section 4 we present the mathematical development of the MMSE estimator. In Section 5 we present experimental analysis of the algorithm. Finally in Section 6 we present our observations and conclusions.

2. The Mixing Model

We consider the specific case where two speakers speak simultaneously. Let $X(t)$ and $Y(t)$ represent the signals generated by the two speakers, who we identify as Speaker $S_X$ and Speaker $S_Y$ respectively. Let $Z(t)$ be the mixed signal recorded by the single microphone. The mixed signal is assumed to be a simple sum of the signals generated by the two speakers.

$$Z(t) = X(t) + Y(t).$$  

(1)

Let $X(\omega)$ represent the power spectrum of $X(t)$, i.e., $X(\omega) = |F[X(t)]|^2$ where $F$ denotes the Fourier transform of the signal, and the $| \cdot |^2$ operation computes a component-wise squared magnitude. Similarly, let $Y(\omega)$ and $Z(\omega)$ represent the power spectra of $Y(t)$ and $Z(t)$ respectively. We assume that $X(t)$ and $Y(t)$ are independent and uncorrelated with each other. Under this assumption:

$$Z(\omega) = X(\omega) + Y(\omega).$$

(2)

Typically, in any frequency band the spectral energy for one of the speakers is much greater than that of the other. This observation can be translated to the following approximation:

$$Z(\omega) \approx \max(X(\omega), Y(\omega)).$$

(3)

Let $z(\omega)$, $x(\omega)$ and $y(\omega)$ be the logarithm of $Z(\omega)$, $X(\omega)$ and $Y(\omega)$ respectively. Equation 3 gives us the following relationship

$$z(\omega) \approx \max(x(\omega), y(\omega)).$$

(4)

We refer to the above approximation as the log-max approximation.

We note that the assumptions in Equations 2 and 4 reflect two separate phenomena that impose conflicting constraints on the manner in which the power spectra of the speech signal is estimated. Equation 2 represents a statistical relationship that is strictly valid only in the long term. It is not guaranteed to hold for power spectra measured from any finite length window. In general, the longer the analysis window over which the power spectrum is estimated, the more valid Equation 2 is. On the other hand, Equation 4 is applicable only to instantaneous power-spectral measurements. As the length of the analysis window increases, the overall spectral characteristics for different speech signals approach each other, and the log-max approximation becomes invalid.

For the work reported in this paper we use an analysis window of 25ms. The window size we use is fairly common; the approximation becomes invalid.

As the length of the analysis window increases, the overall spectral characteristics for power-spectral measurements. As the length of the analysis window over which the analysis window length requirements for both the independence and the log-max assumptions to be valid.

The overall processing of the signals was performed as follows: the signals were segmented into frames of 25ms, where adjacent frames overlapped by 15ms. A 400 point Hanning window was applied to each frame and a 512 point DFT was computed. The log spectrum for each frame was computed by taking the logarithm of the squared magnitude of the DFT for the frame.

3. Model Assumptions

We model the distribution of the log spectral vectors for any speaker by a mixture Gaussian density. Additionally, we assume that given the identity of the Gaussian, the various frequency bands in the log spectral vector are independent of each other. Note that this does not imply that the frequency bands are independent of each other over the entire distribution for the speaker. According to this model, the distribution of the log spectral vectors for speaker $S_X$ can be represented as

$$P(x) = \sum_{k=1}^{K_x} w_k^X \prod_{d=1}^{D} N(x_d; \mu_{x,d}^k, \sigma_{x,d}^k)$$

(5)

where, $K_x$ is the number of Gaussians in the mixture Gaussian density for $S_x$, $w_k^X$ represents the a priori probability for the $k^{th}$ Gaussian in the density, $D$ represents the dimensionality of $x$, i.e. the number of frequency bands in $x$, $x_d$ represents the $d^{th}$ dimension of $x$, $\mu_{x,d}^k$ represents the $d^{th}$ dimension of the mean vector for the $k^{th}$ Gaussian in the mixture Gaussian density for $S_x$, and $\sigma_{x,d}^k$ represents the corresponding variance term. $N(x_d; \mu_{x,d}^k, \sigma_{x,d}^k)$ represents the value of a Gaussian density with mean $\mu_{x,d}^k$ and variance $\sigma_{x,d}^k$ at $x_d$.

The distribution of the log spectral vectors for speaker $S_Y$ is similarly modeled by a mixture Gaussian of the following form:

$$P(y) = \sum_{k=1}^{K_y} w_k^Y \prod_{d=1}^{D} N(y_d; \mu_{y,d}^k, \sigma_{y,d}^k)$$

(6)

The parameters of $P(x)$ and $P(y)$ are learnt from a corpus of training utterances for the two speakers, using the Expectation Maximization (EM) algorithm.

4. Estimation Algorithm

Let $x$ and $y$ represent the log spectral vectors for speakers $S_X$ and $S_Y$ at a given instant of time. Let $x_d$ and $y_d$ represent the $d^{th}$ dimension of $x$ and $y$ respectively. Let $z$ represent the log spectral vector for the corresponding mixed signal, and $z_d$ the $d^{th}$ component of $z$. The relationship between $z_d$, $x_d$ and $y_d$ is given by Equation 4.

Using Bayes’ rule, the a posteriori probability of $x_d$ given $z$ can be expanded as

$$P(x_d|z) = \sum_{k_x=1}^{K_x} \sum_{k_y=1}^{K_y} P(k_x, k_y|z)P(x_d|k_x, k_y, z)$$

(7)

Since $x_d$ and $y_d$ are independent of all other dimensions of $x$ and $y$ respectively, once the Gaussian indices $k_x$ and $k_y$ are known, $z_d$ is also independent of other dimensions of $x$ and $y$, given $k_x$ and $k_y$. Consequently,

$$P(x_d|k_x, k_y, z) = P(x_d|k_x, k_y, z_d)$$

(8)

We introduce the following notation for simplicity:

$$C_x(w|k) = \int_{-\infty}^{\infty} N(x_d; \mu_{x,d}^k, \sigma_{x,d}^k)dx_d$$

(9)

$$P_x(w|k) = N(w; \mu_{x,d}^k, \sigma_{x,d}^k)$$

(10)

$$C_y(w|k) = \int_{-\infty}^{\infty} N(y_d; \mu_{y,d}^k, \sigma_{y,d}^k)dy_d$$

(11)

$$P_y(w|k) = N(w; \mu_{y,d}^k, \sigma_{y,d}^k)$$

(12)
where $w_t$ is a scalar random variable. It can be shown that

$$P(z_t|k_x, k_y) = P_x(z_t|k_x)C_y(z_t|k_y) + C_x(z_t|k_x)P_y(z_t|k_y)$$

and

$$P(x_d|k_x, k_y, z_d) = \begin{cases} P_x(z_d|k_x)P_y(z_d|k_y) & \text{if } x_d \leq z_d \\ P(x_d|k_x, k_y) & \text{otherwise} \end{cases}$$

where $\delta_{x_d}(z_d)$ is a Dirac delta function centered at $z_d$, and captures the probability of event that $y_d < z_d$ and $x_d$ is exactly equal to $z_d$.

The MMSE estimate for $x_d$ is given by

$$\hat{x}_d = E(x_d|z) = \int_{-\infty}^{\infty} x_dP(x_d|z)dx_d$$

Combining Equations 7, 13 and 15, we get

$$\hat{x}_d = \sum_{k_x, k_y} P(x_d|z)$$

$$\left\{ P_y(z_d|k_y)\mu_{k_x, d}C_y(z_d|k_x) - \sigma^2_{k_x, d}P_y(z_d|k_y) \right\}$$

The complete MMSE estimate for the log spectral vector of speaker $S_x$, $\hat{z}$ is obtained by estimating every dimension of $\hat{x}$ by Equation 16.

The overall algorithm for separating the signal for $S_x$ from the mixed signal can be written as follows:

1. For each frame of the signal:
   (a) compute the Fourier transform $Z(\omega)$.
   (b) Compute the log spectral vector $z = log(Z(\omega))$.
   (c) From $z$, derive the MMSE estimate for $\hat{x}$ using Equation 16.

2. Concatenate the re-estimated signals for the various frames using the overlap-add method, to reconstruct the entire signal.

5. Experimental Evaluation

Experiments were conducted to evaluate the proposed MMSE speaker separation algorithm. Two speakers, one male and one female, were used for the experiment. A total of approximately 10 minutes of data were obtained for each speaker. A 256-component mixture Gaussian density was computed for each of the speakers using approximately 8 minutes of the recordings.

Mixed recordings were simulated by digitally adding utterances from the two speakers. The test utterances were not part of the set that was used to compute the mixture Gaussian densities. The signals were added to have an SNR of 0dB; additional scaling of the speakers with respect to each other was not attempted.

Figure 1 shows an example of the signals reconstructed using the MMSE estimator. As a reference, the original signals for the two speakers and the mixed signal are also shown. In addition, as a comparator, signals reconstructed using the MAXVQ technique proposed by Roweis [7] are also shown. The MAXVQ algorithm is similar to the speaker separation algorithm described in [5], except that static Gaussian mixture distributions are used instead of HMMs to model distributions of log spectral vectors.

The spectrograms of the original and the MMSE reconstructed signals shown in Figure 1 are shown in Figure 2.

6. Discussion and Conclusion

We observe from figures 1 and 2 that the proposed MMSE technique is able to separate out the signals for the two speakers quite well. In particular, figure 1 shows that the reconstructed signals for the two speakers can be quite well differentiated from each other.
signals are better than those obtained with MAXVQ. In general, the reconstructed signal for any speaker shows significant suppression of the competing speaker. The MMSE reconstructed signals also have significantly higher SNR than the signals reconstructed by MAXVQ, although the particular spectrograms shown in figure 2 do not demonstrate this clearly.

We note here that the MAXVQ straw man has been chosen simply because it is currently the best algorithm that works under similar statistical constraints: i.e. the joint distribution of all the frequency components in the log spectral vectors for any speaker is modeled by a simple mixture Gaussian. Several researchers have reported significantly improved speaker separation performance with approaches similar to MAXVQ, by modeling different frequency bands of the signal separately, and by using more detailed models such as HMMs, that capture temporal structure. However, the more detailed models used in these techniques can be expected to benefit the proposed MMSE algorithm equally.

Computationally, the proposed MMSE algorithm is far more expensive in its exact form than MAXVQ and related algorithms, because the estimator in Equation 16 iterates over all combinations of Gaussians. The algorithm can be made significantly faster if it is permitted to be inexact, and only the top few combinations of Gaussians are used in the reconstruction. Algorithms such as MAXVQ which employ a “winner-take-all” strategy and perform reconstruction based only on the best combination of Gaussians (or states) for the speakers retain a performance advantage over algorithms such as the MMSE procedure we present, which hedge their bets and consider all combinations of states, in at least one situation: when the mixed signal consists of digitally added recordings of the same speaker. In this situation, algorithms that hedge their bets will naturally find that each log spectral vector from the mixed signal is equally likely to belong to each of the utterances, and reconstruction cannot be effected.

Finally, we note that most single channel speaker separation algorithms assume that the signals for the two speakers have been combined by direct addition. In reality, the signal from each speaker will be convolved with its own specific room response filter. The actual separation algorithm must therefore also account for the unknown linear filter distorting each signal, before matching the signals to the known distributions of the speakers. Masking algorithms such as MAXVQ cannot easily be modified to account for this condition. On the other hand, the MMSE algorithm presented in this paper can be naturally extended to estimate the linear filters in the process of separating the signals. However, the estimation procedure is iterative and has even greater computational complexity than the basic MMSE procedure itself. For this reason, we have not attempted it in this paper.

7. References