RECOGNIZING SPEECH FROM SIMULTANEOUS SPEAKERS

Bhiksha Raj[1], *Rita Singh*[2], *Paris Smaragdis*[1]

1. Mitsubishi Electric Research Labs, Cambridge, MA, USA
2. Haikya Corp., Watertown, MA, USA

bhiksha@merl.com, rsingh@haikya.com, paris@merl.com

Abstract

In this paper we present and evaluate factored methods for recognition of simultaneous speech from multiple speakers in single-channel recordings. Factored methods decompose the problem of jointly recognizing the speech from each of the speakers by separately recognizing the speech from each speaker. In order to achieve this, the signal components of the target speaker in each case must be enhanced in some manner. We do this in two ways: using an NMF-based speaker separation algorithm that generates separated spectra for each speaker, and a mask estimation method that generates spectral masks for each speaker that must be used in conjunction with a missing-feature method that can recognize speech from partial spectral data. Experiments on synthetic mixtures of signals from the Wall Street Journal corpus show that both approaches can greatly improve the recognition of the individual signals in the mixture.

1. Introduction

In this paper we address the problem of recognizing speech from multiple simultaneous speakers from monaural recordings. This is a difficult problem even for human beings - although we are well able to selectively listen to one of many speakers when hearing the sounds binaurally, our performance is much worse when we hear with only one ear. Needless to say, the problem gets immensely more difficult for automatic speech recognition systems.

Although current scientific literature contains several reports on the *separation* of the individual signals from monaural recordings of concurrent speakers, there is surprisingly little on the *recognition* of such data. However the statistical framework required for a solution is readily available. In most current recognizers, the distribution of the speech signals (or, rather, the sequences of parameter vectors derived from speech signals) is modelled by an HMM. By this model, assuming independence between the signals for the two speakers, the distribution of the mixed signal can be represented by a large *factorial* HMM that includes one state for every combination of states in the HMMs for the individual signals. Specifically, if the HMMs for the two speakers have N and M states respectively, the factorial HMM for the mixed signal has $N \times M$ states. The state output distribution the $(i,j)^{\text{th}}$ state of the factorial HMM is obtained from the state output densities of the i^{th} state of the HMM for the first speaker and the j^{th} state for the second speaker, and the function that relates the parameters for the mixed signal and those for the signals for the individual speakers.

Varga and Moore [1] present a recognition algorithm for decoding such HMMs, that simultaneously retrieves the best state sequences through both the component HMMs. In effect, the algorithm simultaneously recognizes the utterances by both speakers. However, the Varga and Moore algorithm, although theoretically precise, requires the decoding of large factorial HMMs, an extremely difficult proposition for all but relatively small tasks. For instance, in [2] Deoras and Hasegawa-Johnson report applying the algorithm to the recognition of multiple speakers, but restrict themselves to a digits task where both speakers have uttered digit sequences. No results are reported on the application of this technique to larger recognition tasks.

In this paper we follow a simpler approach: we factor the problem of recognizing multiple concurrent speakers into multiple independent recognition procedures, one for each speaker. In each case, the signal components for the target speaker are enhanced in the mixed signal in some manner.

A variety of single-channel speaker separation solutions have been proposed in the literature that can be used to enhance the target speaker. These methods can largely be categorized as spectral-decomposition-based methods and mask-based methods. Spectral-decomposition-based methods learn typical spectral structures, or "bases", for individual speakers from training data. Mixed signals are decomposed into linear combinations of these bases. The signals for individual speakers are obtained by recombining their bases with the appropriate weights. Jang et. al. [3] derive the bases for speakers through independent component analysis (ICA) of their signals. Smaragdis [4] derives them through non-negative matrix factorization (NMF) of their spectra. Other authors have derived bases through vector quantization, Gaussian mixture modelling, etc. The characteristic of the spectral-decomposition-based approach is that it attempts to derive entire spectra for each of the speakers.

Mask based methods, on the other hand, are based on the notion that in a mixed speech signal, any given frequency band is dominated by only one of the speakers at any time. By this model, any speaker can be effectively separated from a mixture by identifying the time-frequency components of the mixed signal in which they dominate and reconstructing a signal from these components alone. The problem then simply becomes one of estimating the spectral *masks* that identify the time-frequency components within which any speaker dominates. In [5] Roweis presents the *max-VQ* algorithm that models the distribution of the log spectra of individual speakers as Gaussian mixtures. The time-frequency components to be associated with each speaker are identified through an efficient branch-and-bound algorithm that identifies the most likely combination of Gaussians for each spectral vector. Other authors attempt to segregate time-frequency components by speaker using perceptual principles (e.g. [6]), or through the use of automated clustering techniques, e.g. [7].

In this paper we have evaluated one spectral-decomposition-based method: the NMF-based separation algorithm of Smaragdis [4], and one mask-based method: the max-VQ algorithm by Roweis [5]. For the NMF-based method,

recognition was performed with features derived from the spectra reconstructed by the algorithm for each speaker. For the mask-based method, on the other hand, we have employed the missing-feature approach proposed by Cooke et. al. [8] that aims to perform recognition with partial spectral information such as might be specified by the spectral masks obtained from max-VQ, for recognition. The specific missing-feature method employed in this paper is the cluster-based imputation method of Raj et. al. [9], although other missing-feature methods are also applicable. Our recognition results, reported on synthetic mixtures of signals from the Wall Street Journal corpus, indicate that both spectral decomposition and mask-based approaches can be used to significantly enhance recognition of individual speakers, at least at relatively low levels of interference from competing speakers.

Before we proceed, we note here that in the rest of this paper we have assumed that a mixed signal comprises speech from two speakers. However, much of the discussion can be extended simply to more complex mixtures, although the recognition results obtained with such mixtures may be expected to worsen with increasing speakers.

The rest of this paper is arranged as follows: In Section 2 we briefly outline Smaragdis' NMF-based speaker separation algorithm. In Section 3 we outline Roweis' max-VQ algorithm for generating spectral masks. In Section 4 we briefly describe the missing feature method employed in conjunction with the mask estimation method of Section 3. In Section 5 we outline the entire recognition process. In Section 6 we describe our experimental results, and finally in Section 7 we present our observations and plans for future work.

2. NMF-based speaker separation

Matrix factorisation algorithms attempt to decompose a real $M \times N$ matrix $\mathbf{V}$ as the product of an $M \times R$ matrix $\mathbf{W}$ and an $R \times N$ matrix $\mathbf{H}$ as:

$$\mathbf{V} \approx \mathbf{W}.\mathbf{H} \tag{1}$$

In such decomposition, the columns of $\mathbf{W}$ may be interpreted as a set of basis vectors and the columns of $\mathbf{H}$ as the coordinates of the column vectors in $\mathbf{V}$ in terms of these bases.

Conventional factorisation techniques such as principal component analysis and independent component analysis permit the entries of both $\mathbf{W}$ and $\mathbf{H}$ to be both negative an positive. However, for strictly non-negative data, such as data sets comprising only power spectral vectors of a signal, the resulting bases and their weights bear no intuitive meaning. In [10], Lee and Seung present a non-negative factorisation technique that ensures that the entries of $\mathbf{W}$ and $\mathbf{H}$ are strictly non-negative. Briefly, the NMF algorithm initialises the non-negative matrices $\mathbf{W}$ and $\mathbf{H}$ and iteratively updates them through repeated application of the updates:

$$\mathbf{H} = \mathbf{H} \otimes \frac{\mathbf{W}^\top.[\frac{\mathbf{V}}{\mathbf{W}.\mathbf{H}}]}{\mathbf{W}^\top.\mathbf{1}} \tag{2}$$

$$\mathbf{W} = \mathbf{W} \otimes \frac{[\frac{\mathbf{V}}{\mathbf{W}.\mathbf{H}}].\mathbf{H}^\top}{\mathbf{W}^\top.\mathbf{1}} \tag{3}$$

where $\otimes$ represents a Hadamard (component-wise) product and all matrix divisions are also per-component.

The bases derived from NMF decomposition of images and text have empirically been observed to represent perceptually meaningful parts of faces, characters, etc. In [4] Smaragdis presents a single-channel speaker separation algorithm that is based on NMF decomposition of spectra.

In this method, the sequences of power spectral vectors derived from windowed short-time Fourier analysis of the signals for each speaker are treated as spectral matrices. In a training step, basis vectors are derived from spectral matrices obtained from training data for each speaker. Let $\mathbf{X}_{tr}$ represent an $M \times T_X$ spectral matrix comprising the sequence of power spectral vectors derived from training data for a speaker S_X. M is the length of the power spectral vectors for the signal (i.e. the no. of unique FFT points for any analysis window). Let $\mathbf{Y}_{tr}$ be an $M \times T_Y$ spectral matrix for the training data for speaker S_Y. $\mathbf{X}_{tr}$ is decomposed into the product of an $M \times R_X$ matrix $\mathbf{W}_X$ and a $R_X \times T_X$ matrix $\mathbf{H}_X$, by iterations of Equations 2 and 3. $\mathbf{Y}_{tr}$ is similarly decomposed into the product of an $M \times R_Y$ matrix $\mathbf{W}_Y$ and an $R_Y \times T_Y$ matrix $\mathbf{H}_Y$. R_X and R_Y represent the number of basis vectors for each of the speakers and must be specified externally to the algorithm.

Given a new mixed recording from both speakers, the bases computed for each of them are used to separate their signals. Let $\mathbf{Z}$ represent an $M \times T_Z$ spectral matrix obtained from the mixed signal. An extended $M \times (R_X + R_Y)$ basis matrix $\mathbf{W} = [\mathbf{W}_X \mathbf{W}_Y]$ is created by concatenating the basis matrices for the two speakers. $\mathbf{Z}$ is decomposed into the product of $\mathbf{W}$ and an $(R_X + R_Y) \times T_Z$ matrix $\mathbf{H}_Z$ through iterations of Equation 2. The separated power spectral matrices for the individual speakers are reconstructed as

$$\hat{\mathbf{X}} = \mathbf{W}_X.\mathbf{U}_X.\mathbf{H}_Z \tag{4}$$

$$\hat{\mathbf{X}} = \mathbf{W}_Y.\mathbf{U}_Y.\mathbf{H}_Z \tag{5}$$

where $\mathbf{U}_X$ is an $R_X \times (R_X + R_Y)$ matrix such that leading R_X diagonal elements are 1 and the rest of the terms are 0, and $\mathbf{U}_Y$ is and $R_Y \times (R_X + R_Y)$ matrix such that the trailing (rightmost) R_Y diagonal elements are one and the rest of the elements are 0.

Equation 4 essentially reconstructs the power spectrum for each of the speakers by recombining their bases with their respective weights from the $\mathbf{H}_Z$ matrix. The signals for the individual speakers are then reconstructed by combining $\hat{\mathbf{X}}$ and $\hat{\mathbf{Y}}$ with the phase of the short-time Fourier transform of the mixed signal and performing an inverse short-time Fourier transform.

3. Estimating Spectral Masks with Max-VQ

Let $Z_t(\omega)$ represent the logarithm of the power spectrum of the mixed speech signal within the t^{th} analysis window. Let $X_t(\omega)$ and $Y_t(\omega)$ represent the log spectra (i.e. the log of the power spectra) respectively of the component signals from the two speakers, within the same analysis window. The max-VQ algorithm makes the assumption that:

$$Z_t(\omega) = \max(X_t(\omega), Y_t(\omega)) \tag{6}$$

The goal of algorithm is to determine, for every analysis window, a binary spectrographic mask $S_t(\omega)$ such that $S_t(\omega) = 1\ if\ X_t(\omega) > Y_t(\omega),\ 0\ otherwise$. The mask $S_t(\omega)$ identifies which of the two speakers is actually represented in $Z_t(\omega)$.

In order to estimate $S_t(\omega)$, it is assumed that the log-spectral vectors for each speaker are drawn from speaker-specific codebooks of vectors. The log spectral vector for the observed mixed signal is the element-wise maximum of the codewords for the individual speakers, plus some zero-mean additive estimation noise. Let $\{\mathbb{X}\}$ represent the set of codewords

in the codebook for the first speaker, and $\mathcal{X}_k$ the k^{th} codeword in the codebook. Similarly, let $\{\mathbb{Y}\}$ represent the codebook for the second speaker and $\mathcal{Y}_j$ the j^{th} codeword in the codebook. We have dropped the ω in this notation, for brevity. $\{\mathbb{X}\}$ and $\{\mathbb{Y}\}$ are learnt from training data for the two speakers. Let Σ represent the variance of the estimation noise. The probability that a log spectral vector Z_t for the mixed signal was generated from the codewords $\mathcal{X}_k$ and $\mathcal{Y}_j$ for the two speakers is given by

$$P(Z_t, \mathcal{X}_k, \mathcal{Y}_j) = \pi_{k,j}\mathcal{N}(Z_t; \max(\mathcal{X}_k, \mathcal{Y}_j), \Sigma) \quad (7)$$

where $\mathcal{N}(Z; \mu, \Sigma)$ represents the value of a Gaussian density with mean μ and variance Σ at Z. $\max(.)$ in Equation 7 is an element-wise maximum, and $\pi_{k,j}$ is the *a priori* probability of the pair of codewords $\mathcal{X}_k$ and $\mathcal{Y}_j$. Max-VQ identifies the most likely pair of codewords to have generated Z as:

$$\hat{\mathcal{X}}_t, \hat{\mathcal{Y}}_t = \text{argmax}_{\mathcal{X}\in\{\mathbb{X}\},\mathcal{Y}\in\{\mathbb{Y}\}} P(Z_t, \mathcal{X}, \mathcal{Y}) \quad (8)$$

Roweis presents a highly efficient branch-and-bound algorithm to solve Equation 8. The spectrographic mask for the vector Z_t is obtained as $S_t = step(\hat{\mathcal{X}}_t - \hat{\mathcal{Y}}_t)$, where $step(.)$ takes the value 1 if its argument is positive and 0 otherwise.

4. Missing feature methods: Cluster-based Reconstruction of Spectra

Consider a log-spectral vector Z derived from one analysis frame of the mixed signal obtained from speakers S_X and S_Y. Let X and Y represent the log spectral vectors for the signals from S_X and S_Y that sum up to compose the mixed signal within the window. Let S represent the spectral mask that identifies the components of Z that belong to X. By definition, then

$$X[i] \begin{cases} = Z[i] & \text{if} \quad S[i] = 1 \\ \leq Z[i] & \text{else} \end{cases} \quad (9)$$

where $X[i]$, $Z[i]$ and $S[i]$ represent the i^{th} component of X, Z and S respectively. The inequality for $S \neq 1$ results directly from Equation 6.

Cluster-based reconstruction attempts to reconstruct the components of X for which only the bound $X[i] \leq Z[i]$ is known. The distribution of the X is modelled by a mixture of Gaussians with diagonal covariance matrices:

$$P(X) = \sum_k c_k \prod_i \mathcal{N}(X[i]; \mu_{k,i}, \sigma_{k,i}) \quad (10)$$

where c_k is the mixture weight of the k^{th} Gaussian in the mixture, and $\mu_{k,i}$ and $\sigma_{k,i}$ are the mean and variance respectively of the i^{th} component of X, for the k^{th} Gaussian in the mixture. All c_k, $\mu_{k,i}$ and $\sigma_{k,i}$ values are trained from corpus of training speech from S_X.

$P(k|X, S)$, the *a posteriori* probability of the k^{th}, given the vector X and its mask S is given by

$$P(k, X, S) = \prod_{i:S[i]=1} \mathcal{N}(X[i]; \mu_{k,i}, \sigma_{k,i}). \prod_{j:S[j]\neq 1} \int_{-\infty}^{Z[j]} \mathcal{N}(x; \mu_{k,j}, \sigma_{k,j}) dx \quad (11)$$

$$P(k|X, S) = \frac{P(k, X, S)}{\sum_j P(j, X, S)} \quad (12)$$

The unknown components of X are computed as

$$X[i] \mid_{S[i]\neq 1} \simeq \sum_k P(k|X, S) \min(Z[i], \mu_k, i) \quad (13)$$

The outcome of the reconstruction process is a complete spectral vector X, where some of the components are derived directly from $Z[i]$ and the rest are estimated by Equation 13.

5. Factored recognition of multiple concurrent speakers

Factored recognition of the multiple speakers in a mixed signal is performed using one of the following procedures:

- The signals for the individual speakers are obtained by the NMF-based speaker separation procedure described in Section 2. Cepstral features are derived from these signals and recognition is performed with them.
- Spectral masks are derived for mel log spectral vectors of the mixed signal by the max-VQ algorithm. These spectral masks are used to reconstruct complete log spectral vectors for each of the speakers, by the procedure outlined in Section 5. Cepstral vectors derived from the reconstructed vectors are used for recognition.

6. Recognition Experiments

Recognition experiments were conducted on synthetic mixtures of signals from two male and two female speakers, selected from the speaker dependent portion of the Wall Street Journal corpus distributed by LDC. A set of 400 utterances (approximately 50 minutes) were used as test data for each speaker. A separate half hour of data from each speaker was set aside as training data. The test utterances were digitally added to simulate mixed single channel recordings at speaker-to-speaker energy ratios (SSR) of -10, -5, 0, 5 and 10 dB. Note that a mixed signal for two speakers S_X and S_Y that has a SSR of 10dB for S_X has an SSR of -10dB for S_Y. In all mixtures, the length of the mixed signal was set to that of the longer of the two component signals. Separate mixtures were created for the combination of two male, a male and a female, and two female speakers.

For the NMF-based method, signals were analyzed using 64 ms windows (corresponding to an FFT size of 1024). 100 NMF bases were trained for each speaker (a number empirically determined to be optimal for such training set sizes). The signals for the individual speakers were separated from each of the mixtures and 13-dimensional mel-cepstral vectors derived from them for recognition. For the mask-based method, 40-dimensional log spectral vectors were computed for each 25 ms segment of speech. Adjacent segments overlapped by 15ms. 1024 component Gaussian mixture densities were trained from the training data for each speaker, to be used both by max-VQ and cluster-based reconstruction. For each mixed signal spectral masks were obtained using max-VQ and used to perform cluster-based reconstruction of complete mel-log-spectral vectors for each speaker, from which 13-dimensional cepstral vectors were derived for recognition.

The CMU Sphinx-III continuous density speech recognition system, trained using the speaker-independent component of the training set in the WSJ0 corpus, was used for all experiments. The feature set used included cepstra, difference and double-difference cepstra. Cepstral mean normalization was also performed. The models were further adapted to the training data for each of the four speakers by supervised maximum-likelihood linear regression. In all experiments recognition of any speaker was performed using the specific models adapted to them. The baseline recognition errors on the unmixed signals for the four speakers, identified as "male1", "male2", "female1" and "female2", were 11.5%, 7.8%, 4.6% and 8%, respectively.

Tables 1, 2 and 3 show the recognition errors obtained for each of the three mixtures (male-male, male-female, and female-female), for both NMF-based and mask-based recogni-

Table 1: *Recognition error(%) on a mixture of two male speakers. "None" refers to recognition of unprocessed mixed signals.*

method	spkr	-10dB	-5dB	0dB	5dB	10dB
None	male1	113.3	111.7	103.9	86.7	66.4
	male2	115.2	115.9	109.4	92.5	72.5
NMF	male1	112.6	109.8	102.4	86.3	67.5
	male2	118.9	116.0	107.4	90.8	69.3
Max-VQ	male1	94.3	93.4	86.9	81.3	70.4
	male2	97.1	96.4	88.7	56.8	35.3

Table 2: *Error(%) on a mixture a male and a female speaker.*

method	spkr	-10dB	-5dB	0dB	5dB	10dB
None	male	115.5	116.8	111.1	91.9	71.4
	female	120.8	119.8	110.2	93.3	72.9
NMF	male	114.9	109.3	95.8	76.8	58.6
	female	121.8	115.6	100.7	80.4	61.9
Max-VQ	male	98.7	99.7	95.3	81.2	65.4
	female	92.4	88.9	75.4	48.6	25.8

Table 3: *Error(%) on a mixture of two female speakers.*

method	spkr	-10dB	-5dB	0dB	5dB	10dB
None	female1	120.8	120.0	108.5	84.0	57.2
	female2	114.1	117.3	112.6	95.2	67.5
NMF	female1	119.5	117.0	106.5	85.0	61.9
	female2	100.2	115.6	109.6	95.1	74.7
Max-VQ	female1	95.8	90.7	81.3	51.2	25.9
	female2	95.0	99.6	92.4	89.6	88.2

tion. Baseline recognition obtained with the unprocessed signals is also shown. Note that the reported error includes insertion errors. Error rates greater than 100 imply that in addition to substitution errors the recognizer has also inserted a large number of spurious words. The recognition error rates reported in the tables are extremely high, exceeding 100% in many cases as a result. An alternate performance metric that might have been reported is the recognition *accuracy* (recall), which measures the percentage of uttered words that were correctly recognized. This number was relatively high, lying between 30% and 90% in all cases. However, since insertion errors are an important phenomenon in the recognition of speech-over-speech data, we have preferred to report the error rate over the recall.

7. Observations and Conclusion

It is clear from the tables that recognition of speech-over-speech data is extremely difficult, even at the modest SSR of 10dB. Encouragingly however, recognition error can be significantly improved by the methods applied in this paper. Though the error remains poor in most cases, in others the improvement is significant, reducing by over 50% at 10dB for female1.

However, improvements are not obtained in all cases - they are much greater for some speakers than others. For instance, the greatest improvements have been obtained for female1, both in the male-female and female-female combinations. Greater improvements have been obtained for male2 than for male1. The techniques appear to be ineffective for female2. Also, NMF based separation is effective only for the male-female combination, failing to register any improvement for same-gender mixtures. This is possibly because NMF bases for people of the same gender tend to be very similar.

Several issues remain to be investigated. The curious speaker-dependent phenomenon needs investigation: although the separation methods simultaneously separate the signals (or masks) for both speakers, greater improvement is obtained for one speaker than the other in any mixture. Also, it remains to be determined if superior recognition may be obtained by modifying the manner in which signals have been analyzed, e.g. by the inclusion of perceptual weighting schemes for NMF or by modifying the number of MEL filters etc. for the mask-based methods. Finally, the choice of missing feature methods is an issue: we have only tested one method. The marginalisation based method proposed by Cooke performs optimal classification, and may result in superior recognition performance. Further, the masks used in this paper are binary: frequency components are uniquely associated with a speaker. In [11] a *soft-mask* technique is proposed that associates each frequency component with every speaker with a probability. The use of this such masks with the *soft* marginalisation approach of Morris et. al. [12] may be expected to result in even greater improvements. All of this remains future work to be reported in a future paper.

8. References

[1] Varga A. P., and Moore, R., "Hidden Markov Model Decomposition of Speech and Noise", Proc. IEEE Conf. on Acoustics Speech and Sig. Proc. (ICASSP), 1990.

[2] Deoras, A. M. and Hasegawa-Johnson, M., "A Factorial HMM Approach to Simultaneous Recognitoin of Isolated Digits Spoken by Multiple Talkers on One Audio Channel", Proc. ICASSP, 2004.

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

[4] Smaragdis, P., "Convolutive Speech Bases and their Application to Supervised Speech Separation", Submitted to the IEEE Trans. on Speech and Audio Processing, 2005.

[5] Roweis, S., "Factorial Models and Refiltering for Speech Separation and Denoising", Proc. Eurospeech 2003.

[6] Wang, D. L. and Brown, G. J., "Separation of speech from interfering sounds based on oscillatory correlation", IEEE Trans. on Neural Networks, Vol. 10(3), 1999, pp. 684-697.

[7] Bach, F. R. and Jordan, M. I., "Blind one-microphone speech separation: A spectral learning approach", Neural Information Processing Systems, Dec. 2004.

[8] Cooke, M., Green, P., Josifovski, L. and Vizinho, A., "Robust automatic speech recognition with missing and unreliable acoustic data", Speech Communication, Vol. 34, 2001, pp. 267-285.

[9] Raj, B., Seltzer, M. L. and Stern, R. M., "Reconstruction of Missing Features for Robust Speech Recognition", Speech Communcation, Vol. 43, 2004, pp. 275-196.

[10] Lee, D., D. and Seung, H. S., "Learning the parts of objects with nonnegative matrix factorisation", Nature Vol. 401, 1999, pp. 788-791.

[11] Reddy, A. and Raj, B., "Soft Mask Estimation for Single Channel Speaker Separation", ISCA ITRW on Statistical and Perceptual Audio (SAPA2004), 2004, Jeju.

[12] Morris, A., Barker, J. and Bourlard, H., "From missing data to maybe useful data: soft data modelling for noise robust ASR", Proc. IoA Workshop on Innovative methods in Speech, 2001.