Performance evaluation of three features for model-based single channel speech separation problem

M. H. Radfar[†,‡]*, R. M. Dansereau*[†]*, A. Sayadiyan*[‡]

[†]Department of Systems and Computer Engineering, Carleton University, Ottawa, Canada
[‡]Department of Electrical Engineering, Amirkabir University of Technology, Tehran, Iran
{radfar,rdanse}@sce.carleton.ca & eeas35@cic.aut.ac.ir

10.21437/Interspeech.2006-654

Abstract

This paper addresses the efficiency of three features for the model-based single channel speech separation problem. The separability of three features: log spectrum, modulated lapped transform (MLT) coefficients, and a fusion of pitch and envelop information are evaluated using a VQ-based speech separation technique. At the core of this approach are two trained codebooks of the quantized feature vectors of speakers, whereby the main evaluation for separation is performed. The experiments are conducted in two different scenarios: speaker-dependent and speaker independent. The results show that the log spectrum outperforms the other features for speaker-dependent scenario. However, for the speaker-independent scenario, the best results are obtained from applying the pitch-envelop feature.

Index Terms: single channel speech separation, computational auditory scene analysis (CASA), spectrum, modulated lapped transfron (MLT), and pitch-envelope .

1. Introduction

In the context of speech separation, single channel speech separation is treated as an underdetermined problem when the number of observations is less than the number of sources. In this case, the problem is too ill-conditioned to be solved using common separation techniques. The state-of-the-art techniques are able to deliver an appropriate quality, but only in special cases (e.g., based on *a priori* knowledge of speakers [1–4], or for only voiced segments [5]). The techniques that use *a priori* knowledge of underlying speakers to combat the problem are usually referred to as model-based single channel speech separation techniques [1–4]. In these techniques, first a statistical model is fitted to the feature vectors of each speaker. Then, the two speaker models are combined to model the mixed signal. Finally, in the test phase, the states that best match the mixed signal are decoded based on some criteria (e.g., minimum mean square error, likelihood ratio).

Though many efforts have been performed to introduce new models for this problem, less works have been done to examine the efficiency other features rather than the log spectrum on the performance of single channel speech separation systems. In this paper we consider the problem from this perspective. The separation system is based on vector quantization which, in fact, is a special case of model-based single channel speech separation techniques.

The authors would like to thank the Natural Sciences and Engineering Research Council (NSERC) of Canada and the Iran Ministry of Science and Research who partially funded this research.

In the remaining sections, we start by discussing three different feature spaces in Sec.1, namely the log magnitude of the short-time Fourier transform (STFT), the modulated lapped transform [6] which is extensively used in speech and audio coding, and finally an integration of pitch and envelop. In Sec.2, details are given on how the VQ-based speech separation is performed. The experiments performed are described in Sec.3, and finally discussion and conclusions are given in Sec.4.

2. Feature Extraction

An appropriate feature for separation should have three main characteristics. First, the relationship between the mixture and individual signals in the feature space should be straightforward, accurate, and with as few parameters as possible. Second, the dimension of the feature vector should be as low as possible such that the storage space and searching complexity are minimized. Third, the feature should follow the compactness property, where the signal is modeled with as few codevectors as possible.

The speech features we extract are based on frequency domain features. One approach of dealing with frequency information is to use the *mixture-maximum approximation* (MIXMAX) proposed by Nadas *et al.* [7]. Consider that a monaural mixture of two speech signals is given by

$$s^{mix}(t) = s^1(t) + s^2(t) \tag{1}$$

where the superscript of $s^i(t)$, $i = \{1, 2\}$ is used to indicate speaker 1 and speaker 2, respectively, and should not be confused as an exponent. The Fourier transform of (1) gives

$$S^{mix}(j\omega) = S^1(j\omega) + S^2(j\omega). \tag{2}$$

According to [7], the MIXMAX approximation for (2) is

$$\log|S^{mix}(j\omega)| \approx \max\left(\log|S^1(j\omega)|, \log|S^2(j\omega)|\right) \tag{3}$$

where knowledge of $\log|S^{mix}(j\omega)|$ means an approximate knowledge of either $\log|S^1(j\omega)|$ or $\log|S^2(j\omega)|$. Note that (3) is a function of ω, so the MIXMAX approximation is an element-wise approximation.

In the following subsections, we describe the three features extracted for this paper.

2.1. Log Spectrum

Using the idea of the MIXMAX approximation, one feature examined is the log magnitude of the windowed short-time Fourier

transform (STFT). Using $s[n]$ as the notation for the sampled signal $s(t)$, this feature is represented as

$$L_n^{mix}(\omega) = \log|S_n^{mix}(\omega) * W(\omega)| \qquad 0 < \omega < \pi \qquad (4)$$

where $S_n^{mix}(\omega)$ is the STFT of $s[n]$ shifted to sample n and $W(\omega)$ is the Fourier transform of the analysis window. In this paper, the window corresponds to a 30 ms duration Hamming window and the STFT is performed at every 10 ms for an 8 kHz sampling rate. Also, the windowed sequence is zero padded to 512 samples which after the transform leaves 256 components for $L_n^{mix}(\omega)$ since the symmetry is discarded.

2.2. Modulated Lapped Transform

The modulated lapped transform (MLT) was introduced to combat discontinuity artifacts of block transforms (*e.g.*, DCT) in speech and image processing applications [6]. The MLT mitigates blocking artifacts by overlapping adjacent windows of consecutive transform segments, thus dramatically reducing the artifacts.

The analysis equation for the MLT has the form [6]

$$X_{\text{MLT}}[k] = \sum_{n=0}^{2M-1} x[n]p_{n,k} \qquad k = 0, 1, \ldots, M-1 \qquad (5)$$

where M is the number of coefficients and $x[n]$ is the input speech signal. Notice that the number of MLT coefficients is half the length of the analysis window. The transform kernel $p_{n,k}$ in (5) is given by

$$p_{n,k} = h[n]\sqrt{\frac{2}{M}}\cos\left[\frac{(2n+M+1)(2k+1)\pi}{4M}\right] \qquad (6)$$

where $h[n]$ is a lowpass half-band filter that must satisfy certain conditions for the MLT synthesis equation to exhibit perfect reconstruction. One such $h[n]$ is [6]

$$h[n] = \begin{cases} \sin\left[\left(n+\frac{1}{2}\right)\left(\frac{\pi}{2M}\right)\right], & 0 \le n < 2M, \\ 0, & \text{otherwise.} \end{cases} \qquad (7)$$

The cosine factor in (6) modulates the lowpass filter of $h[n]$ to different frequency bands depending on the values of n and k. The MLT synthesis equation is given as

$$x[n] \cong \sum_{k=0}^{M-1}\left(X^{\text{MLT}}[k]p_{n,k} + X_P^{\text{MLT}}[k]p_{n+M,k}\right) \qquad (8)$$

where $X_P^{\text{MLT}}[k]$ is the previous block and the approximately equal turns to equality for the $h[n]$ in (7). It is evident from the $X_P^{\text{MLT}}[k]$ in (8) that the overlap-add method is used for reconstruction.

2.3. Pitch and Envelop

The process of speech production is similar to filtering in the context of signal processing, where an excitation signal, produced by the vocal cords, is filtered out by a semi all-pass filter known as the vocal tract. The excitation signal is either nearly an impulse train during voiced speech or noise during unvoiced speech. This filtering can be formulated in the Fourier domain as follows

$$\hat{S}_n^i(\omega) = [E(\omega) \times H(\omega)] * W(\omega) \qquad (9)$$

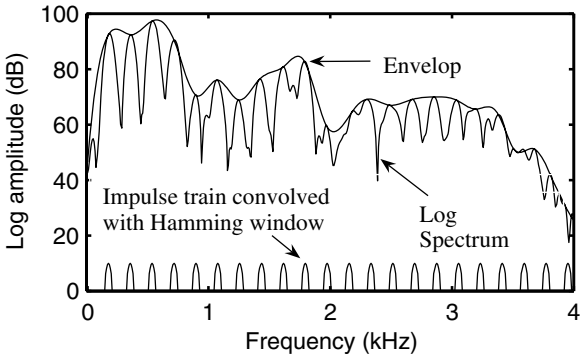

Figure 1: The process of extracting features from log spectrum.

where $\hat{S}_n^i(\omega)$, $E(\omega)$, $H(\omega)$, and $W(\omega)$ are the Fourier transform of the reconstructed speech signal, excitation signal, vocal tract filter, and the applied window, respectively. This filtering is depicted in Fig. 1.

Pitch values are extracted using a multi-pitch tracker, such as with [5]. In order to obtain the envelop of the log spectrum, the method described in [8] is used. During the unvoiced segments no pitch exists, but as shown in [8], we can use an impulse train with fundamental frequency of 100 Hz multiplied by the corresponding envelop. Thus we consider the $E(\omega)$ and $H(\omega)$ as the selected features for the separation process.

3. Separation Models

In this section we describe the three models used for separation. At the heart of these models are codebooks of the quantized feature vectors whereby the main process of separation is performed.

3.1. VQ-based Separation Model for Log Spectra

Figure 2 illustrates the VQ-based log spectra separation system. The main objective of this model is to find two binary masks. These masks are applied to the spectrum of the mixed signal to recover the individual signals.

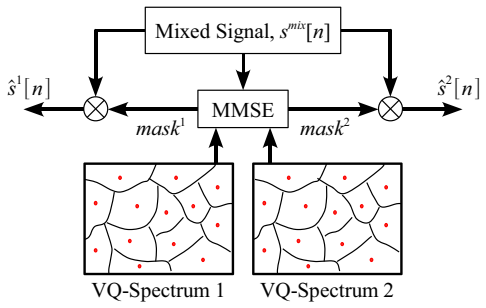

Figure 2: VQ-based separation model for log spectrum.

During the training phase, the two codebooks are trained using the LBG algorithm with *a priori* data from each speaker. In the separation phase, a search is done through the codevectors $\mathbf{c}_j^1$

and $\mathbf{c}_\ell^2$, for $1 \leq j, \ell \leq K$, to find the optimal codevectors, $\mathbf{c}_{opt}^1$ and $\mathbf{c}_{opt}^2$, that when mixed satisfy a minimum distortion criterion compared to the observed mixed signal's feature vector. We opt for the mean square error (MSE) distortion which results in the minimum mean square error (MMSE) between the log spectrum of the mixed signal and those of the individual signals.

The exact relation between the mixed signal and the individual signals requires phase information. However, constructing a model for phase information has been shown to be difficult with the results generally unsatisfactory. In order to obviate this difficulty, we use the MIXMAX approximation to help determine the optimal codevectors $\mathbf{c}_{opt}^1$ and $\mathbf{c}_{opt}^2$. First, let's define an element-wise MIXMAX function for two codevectors as follows

$$\tilde{\mathbf{c}}_{j,\ell}^{mix} = \text{MIXMAX}(\mathbf{c}_j^1, \mathbf{c}_\ell^2) \triangleq \{\max(c_j^1(1), c_\ell^2(1)), \ldots \max(c_j^1(k), c_\ell^2(k)), \ldots, \max(c_j^1(K), c_\ell^2(K))\} \quad (10)$$

where $c_j^i(k)$ is the k^{th} element in the K-dimensional codevector $\mathbf{c}_j^i$. All codevector pairs $\{\mathbf{c}_j^1, \mathbf{c}_\ell^2\}$ are compared to find the MMSE compared to the magnitude of the discrete Fourier transform of the mixed signal $|S^{mix}(k)|$, as follows

$$\{j_{opt}, \ell_{opt}\} = \underset{j,\ell}{\text{argmin}} \sum_{k=1}^{K} \left[|S^{mix}(k)| - \tilde{c}_{j,\ell}^{mix}(k)\right]^2 \quad (11)$$

where $\{j_{opt}, \ell_{opt}\}$ are the codebook indices for the optimal codevector pair $\{\mathbf{c}_{opt}^1, \mathbf{c}_{opt}^2\}$ and the final 2 in (11) is a power.

With the optimal codevector pair $\{\mathbf{c}_{opt}^1, \mathbf{c}_{opt}^2\}$ known, a binary mask is constructed for each speaker, which for speaker 1 is

$$mask^1(k) = \begin{cases} 0, & c_{opt}^1(k) < c_{opt}^2(k) \\ 1, & c_{opt}^1(k) \geq c_{opt}^2(k) \end{cases} \qquad k = 1, 2, \ldots, K. \quad (12)$$

The $mask^2(k)$ for speaker 2 is formed in a similar fashion. Finally, these masks are applied to the corresponding STFT of the mixed speech to separate the speakers as follows

$$\hat{S}_n^i[k] = mask^i(k) \times \left|S_n^{mix}[k]\right| e^{\angle S_n^{mix}[k]} \quad (13)$$

where $S_n^{mix}[k]$ and $\hat{S}_n^i[k]$ are the corresponding K-point STFT of $s^{mix}[n]$ and $\hat{s}^i[n]$, respectively.

3.2. VQ-based Separation for MLT

As mentioned in the previous section, since the phase information is not available the MIXMAX approximation is used to estimate the log spectrum of the mixed signal. One drawback of the MIXMAX approximation is that low energy segments of one speaker are completely masked by the other speaker, thus the separation does not asymptotically approach to a perfect separation system as the codebook size is increased. One solution might be through the use of a real kernel to make the relation between the mixed and individual signals linear in the transform domain. We opt for the MLT that is extensively used in speech and audio coding applications where the mixture is a linear combination as follows.

$$S_{MLT}[k] = S_{MLT}^1[k] + S_{MLT}^2[k] \quad (14)$$

The separation strategy is similar to that of the log spectrum except that the MIXMAX approximation is replaced by a linear operation, *i.e.*, summation, and the individual signals are recovered by taking the inverse MLT of selected codevectors rather than applying the binary mask.

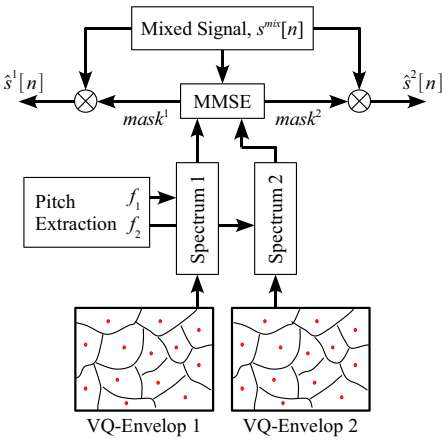

Figure 3: VQ-based separation model for envelop and pitch.

3.3. VQ-based Separation Model for Pitch and Envelop

The main idea behind this model is to integrate the advantages of computational auditory scene analysis (CASA) techniques and model-based approaches. As mentioned in Sec. 1, CASA separation techniques rely on psychoacoustic cues and especially pitch values. The pitch values for the individual speakers need to be extracted from the mixed signal through multi-pitch tracking. Therefore, if we pass the pitch extraction procedure to CASA methods, we just need to decode the envelop of the log spectrum of the underlying signals. Fortunately there are a wide variety of techniques, especially in the context of speech coding, for extracting and quantizing the envelop information. Moreover the envelop feature vector is less redundant, more perceptually important, and speaker independent. Interestingly, the last property makes this system an appropriate candidate for separating unknown speakers. In this model, we first extract the envelop vector from the training data set and construct one codebook for each speaker (or for a general speaker). Then using the computed pitch values and the codebook entries, the log spectrum is generated. In this paper, we assume that we have access to pitch information as *a priori* knowledge to exclude the error caused by multi-pitch extraction. Nevertheless, multi-pitch tracking can be performed on the mixed signal using well-known methods such as in [5]. The remaining procedure is similar to that of the log spectrum model. A schematic of this approach is illustrated in Fig. 3.

4. Experimental Results and Comparison

In order to evaluate the performance of applied features, we conducted the following experiments. We used one hour of speech signals of fifteen speakers. Five speakers among the fifteen speakers are used for the training phase and the remaining speakers are used for the testing phase. The experiments are performed for both speaker dependent and independent cases. For the speaker

dependent case the speakers are used for training and testing are the same, but for the independent case training and testing speakers are different. Throughout all experiments, a Hamming window is used with a duration of 32 ms and a frame rate of 10 ms. Feature extraction (see Sec. 2) was performed on the entire training data set.The test utterances are mixed, with signal-to-signal ratio adjusted to 0 dB for the test phase. For the objective test we used the signal-to-noise ratio (SNR) between the separated and original signals in the time domain. Considering the trade off between the size of the codebook and the complexity of search process, we opt for a codebook of size 1024 codevectors for the selected features. Vector quantization is performed on the training data set using the binary splitting LBG algorithm which performs better than LBG with random initialization.

After constructing the desired codebooks for log spectrum, envelop , and MLT vectors, five utterances from the test data set of each speaker were selected and added digitally in pairs to generate five mixed signals. Finally the mixed signals are fed to the speaker separation algorithms (see Sec. 3). Fig. 4(a) and (b) show the SNR values obtained by applying the five mixed speech files to the separation technique using log spectrum, pitch-envelop model, and MLT for the speaker dependent scenario.

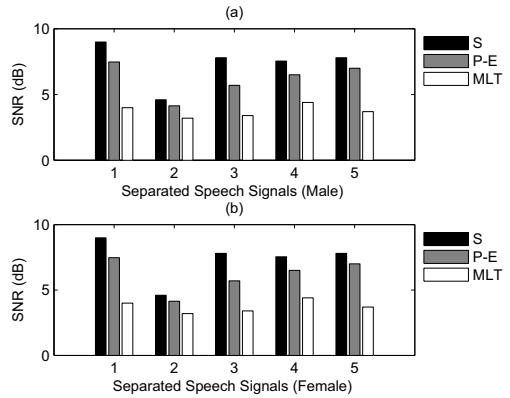

Figure 4: SNR results for separated speech files for the speaker dependent scenario; dark, gray, and white bars show the results obtained by using the log spectrum (S), pitch-envelop (P-E), and MLT (MLT), respectively. Each speech file in the upper panel(a) is separated from its corresponding speech file in the bottom panel.

For speaker independent scenario, the separated results are illustrated in Fig. 5(a) and (b) for log spectrum, pitch-envelop, and MLT. It can be seen from the figures that the performance of the model with the log spectrum , on average, decreased by 4 dB in respect to speaker dependent scenario. For pitch-envelop no difference is observed even for some speech files we can see improvement.

5. Conclusions and Discussion

In this paper we evaluate the performance of three features for the model-based single channel speech separation problem. Although the SNR results obtained by using the log spectrum are better than those using the other two features tested, applying the log spectrum has a main drawback. The log spectrum approach exhibits poor performance when applied for two unknown speakers. In that situation the proposed pitch and envelop feature is an appropriate candidate since the envelop can be trained independent of the speaker and pitch can be extract from the mixture without any *a priori* knowledge of the speakers. Finally, the advantage of MLT is that it obviates the need for a mask. Unfortunately, since the MLT coefficients are less perceptually important than STFT coefficients, a much larger codebook is required for the same SNR performance. Though, if the computational cost of a larger codebook is not a factor, then the separation with the MLT approach would asymptotically approach a perfect separation system.

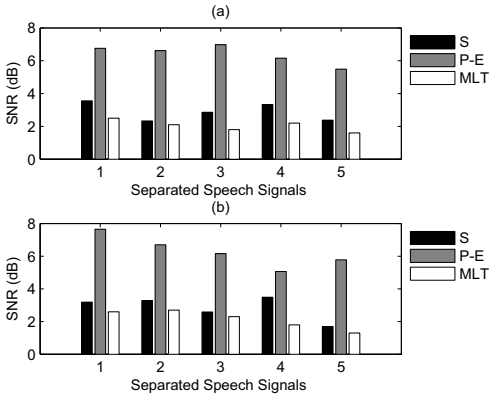

Figure 5: SNR results for separated speech files for the *speaker independent* scenario; dark, gray, and white bars show the results obtained by using the log spectrum (S), pitch-envelop (P-E), and MLT (MLT), respectively. Each speech file in the upper panel(a) is separated from its corresponding speech file in the bottom panel.

6. References

[1] G. J. Jang and T. W. Lee, "A probabilistic approach to single channel source separation," in *Proc. Advances in Neural Inform. Process. Systems*, 2003, pp. 1173–1180.

[2] S. Roweis, "One microphone source separation," in *Proc. Neural Inf. Process. Syst.*, 2000, pp. 793–799.

[3] M. J. Reyes-Gomez, D. Ellis, and N. Jojic, "Multiband audio modeling for single channel acoustic source separation," in *Proc. ICASSP–04*, vol. 5, May 2004, pp. 641–644.

[4] A. M. Reddy and B. Raj, "A minimum mean squared error estimator for single channel speaker separation," in *INTERSPEECH–2004*, Oct. 2004, pp. 2445–2448.

[5] G. Hu and D. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Networks*, vol. 15, no. 5, pp. 1135–1150, Sept. 2004.

[6] H. S. Malvar, *Signal Processing with Lapped Transforms*. Boston, MA: Artech House, 1992.

[7] A. Nadas, D. Nahamoo, and M. A. Picheny, "Speech recognition using noise-adaptive prototypes," *IEEE Trans. Acoust. Speech Sig. Process.*, vol. 37, no. 10, pp. 1495–1503, Oct. 1989.

[8] R. J. McAulay and T. F. Quatieri, "Sinusoidal coding," in *Speech Coding and Synthesis*, W. Kleijn and K. Paliwal, Eds. Elsevier, 1995.