Model-based Speech Separation with Single-microphone Input

S. W. Lee [1], *Frank K. Soong* [1,2], *and P. C. Ching* [1]

[1] Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong SAR, China
[2] Microsoft Research Asia, Beijing, China
{yswlee, pcching}@ee.cuhk.edu.hk, frankkps@microsoft.com

Abstract

Prior knowledge of familiar auditory patterns is essential for separating sound sources in human auditory processing. Speech recognition modeling is one probabilistic way for capturing these familiar auditory patterns. In this paper we focus on separating speech sources with a single-microphone input only. A model-based algorithm is proposed to generate target speech by estimating its spectral envelope trajectory and filtering irrelevant harmonic structure of the interference. The spectral trajectory is optimally regenerated in the form of line spectrum pair (LSP) parameters. Experiments on separating mixed speech sources are presented. Objective evaluation shows that interference is significantly reduced and the output speech is highly intelligible and sounds fairly clear.[1]

Index Terms: speech separation, speech analysis, speech recognition, speech enhancement

1. Introduction

Sound signals typically reach our ears as a mixture of target signal and competing speech or background noise, overlapped in both time and frequency domain. While human being recovers individual sound components remarkably well in these adverse conditions even with a single ear [1, 2], the performance of most speech processing systems is usually degraded. This observation suggests that modeling how human being separates different sound sources is one possible way to extract the target sound from an input mixture prior to any vulnerable processes [2 – 5]. The exploitation of the human perceptual processing in computational systems is referred as computational auditory scene analysis (CASA). Another popular approach for the separation task is independent component analysis (ICA), which relies on the availability of multiple input mixtures and utilizes the statistical properties between sources [6, 7]. This paper proposes a speech separation algorithm for single-microphone input adopting a CASA-based regeneration approach.

The human perceptual processing is referred as auditory scene analysis (ASA), which performs separation by grouping auditory elements that are likely to come from the same source together. This process can be primitive (bottom-up) or schema-based (top-down) [2, 8]. Primitive grouping means acoustic cues extracted from the input mixture, such as harmonicity, continuity and common fates, are used as regularities to determine which elements should gather together. Schema-based grouping, on the other hand, involves the use of conscious effect or priori knowledge of familiar patterns and the process of searching the best match to the input perception. Typical examples of schemas include semantic knowledge and recognition output (caption is helpful when watching a movie where actors are arguing with each other).

Schema-based grouping is a new direction that recent research work starts to adopt in the design of separation systems [9, 10], compared to primitive grouping. In the following, semantic knowledge in the form of transcription is used for separating the mixed sources in a top-down manner. A set of acoustic models of speech recognizer representing the familiar patterns, is manipulated to match the input mixture and the given transcription. This model-based separation system is applicable in certain occasions. One example is the voice extraction of a singer from recordings with background music.

The proposed separation algorithm regenerates the target speech source by working out the corresponding spectral envelope and harmonic structure. In the first stage, the mixture spectrum is shaped towards the target source with an appropriate signal level by using the model parameters retrieved from a speech recognizer. In the second stage, the harmonic structure of interfering source is further attenuated (this could be replaced by enhancing the target harmonics). Preliminary experiments on continuous open-topic speech were carried out. By inspecting the signal-to-interference ratio and the estimated spectral shape, it is found that irrelevant intrusion from interfering speaker has been removed considerably.

2. Model-based speech separation

To search for or align with any familiar patterns of the input mixture signal, speech recognition or pattern models with reliable feature representation are necessary. LSP is employed to parameterize the input mixture signal and a set of hidden Markov models (HMMs) in phone units is trained. LSP, an alternative linear predictive coding (LPC) spectral representation [11, 12], has been shown to be efficient for quantization, well-suited for interpolation and easy for maintaining filter stability. Before moving to the details of the proposed separation algorithm, some fundamentals of LSP are first reviewed.

$$A_M(z) = 1 + \sum_{m=1}^{M} a_m z^{-m} \tag{1}$$

For a given speech frame, the associated inverse filter *A(z)* is shown in Equation (1). Let a_m, m = 1, 2, …, M, be the LPC coefficients with order M. By setting the $(M+1)^{th}$ reflection coefficients, k_{M+1}, to 1 or –1, the LSP coefficients can be computed. This is identical to setting the boundary condition at the glottis as a complete opening or a complete closure [11]. As a result, the inverse filter becomes,

$$A_{M+1}(z) = A_M(z) \pm z^{-(M+1)} A_M(z^{-1}) \tag{2}$$

[1] This work is partially supported by a grant awarded by the Hong Kong Research Grants Council. The work is also affiliated with the Microsoft-CUHK Joint Laboratory for Human-centric Computing and Interface Technologies.

which are two, symmetric and antoi-symmetric polynomials, *P(z)* and *Q(z)* (for $k_{M+1} = \pm 1$). This leads to an interesting property that the roots of *P(z)* and *Q(z)* are interlaced alternately on the unit circle. Hence, it is sufficient to use merely the angles to represent all the information, which are called the LSP frequencies. Equation (3) shows the relationship between these two parameter sets.

$$A_M(z) = \frac{1}{2}\left[P(z) + Q(z)\right] \tag{3}$$

LSP is an efficient representation of the spectrum of an autoregressive, all pole process. Clustering of LSP frequencies indicate spectral peaks or formants in human vocal tract. Its interpolation property is also well-suited for representing the spectral evolution in speech signals.

2.1. Formulation

The input mixture signal *x(n)* is related to its two constituent source signals $x_1(n)$ and $x_2(n)$ as

$$x(n) = x_1(n) + x_2(n) \tag{4}$$

Fig. 1 depicts the block-diagram of our proposed separation system, where $x_i'(n)$ is the estimate for source i ($i \in [1, 2]$). It is designed based upon the source-filter model. In stage 1, Wiener filtering is used to remove the interference source with the spectral information of the target source obtained from the regeneration block. The resultant spectral envelope behaves properly with peaks at target formant frequencies; however, the excitation sources of both target and interference are still present at the output. In stage 2, a comb filtering block is used to attenuate any harmonic structure associated with the interfering source.

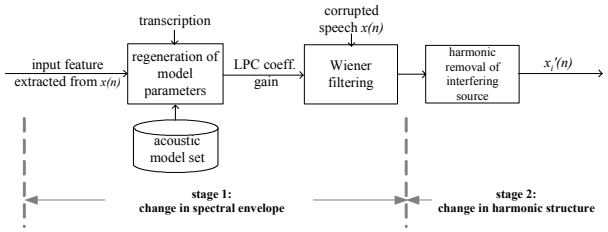

Figure 1: *Block-diagram of the proposed separation system*.

2.2. Regeneration of model parameters

The input mixture signal *x(n)* is first converted to a sequence of feature vectors (referred as feature hereafter). With a given transcription, the regeneration block uses the Viterbi algorithm [13] to perform forced alignment on the input feature so as to generate the optimal time-alignment in state level. By storing a set of commonly found speech patterns (acoustic models), this regeneration block helps to retrieve the expected spectral shape of a speech source. Moreover, the time-alignment defines the boundaries when a particular model state is activated. The mean vectors are extracted from the acoustic model set and replicated according to the time-alignment. By looking at this vector sequence, the expected spectral shape evolution of the target source is illustrated in LSP sense. Fig. 2 shows an example of a time-alignment result and the respective vector sequence.

The model set is HMM-based specifically for Mandarin Chinese, a syllabically paced tonal language. The training data consists of 1,000 sentences in a read speech database recorded by a female speaker (compiled internally by Microsoft Research Asia). The feature vector composes of LSP, gain and fundamental frequency (F0) with both static and dynamic coefficients. They are divided into multiple data streams. The input speech is first analyzed by LPC and the resultant coefficients are converted to LSPs. For LSP and gain, they are modeled by continuous HMMs; while for F0, the multi-space distribution HMM (MSD-HMM) [14] is adopted. F0 sequence is an observation, which bears a continuous value in voiced region, while undefined in unvoiced regions. In the MSD-HMM, a single space with an associated space weight is allocated for every possible dimensionality and discrete symbols. For example, one space is dedicated to continuous F0 value and another space to the unvoiced label.

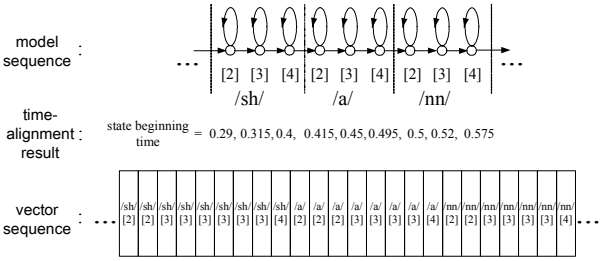

Figure 2: *An example of a time-alignment result and the corresponding vector sequence (The symbols /·/ and [·] denote model name and state number respectively). The longer a state is aligned, the more the mean vector is repeated*.

Segmental tonal modeling is used in the design of modeling unit. In particular, an efficient and tone-carrying phone set called Ph97 [15] is employed. Each Mandarin syllable is constructed by an extended initial (initial and glide) and segmental tonal finals (main and coda). For instance, /han2/ is represented by /h/, /aL/ and /nnH/, where L and H are the low and high value of the 2-scale pitch respectively. These phone units are modeled by 5-state, left-to-right HMMs with single, diagonal covariance Gaussian distribution. Furthermore, context-dependent modeling and stream-dependent state tying are used.

The static component of the vector sequence exhibits the spectral shape evolution of a target source. It is also associated to the dynamic component by linear regression. Simple replication of the mean vector from a model state will produce piecewise constant static components in consecutive frames, but non-zero dynamic components. As a result, the dynamic component, including the delta and delta-delta, is used as a constraint to modify the static component [16]. Let $\mathbf{r_{ik}} = [r_{ik}(1), r_{ik}(2),\ldots, r_{ik}(N)]^T$ be vector taken from kth dimension of the vector sequence, for $k = 1, 2,\ldots, K$, where K and N are the LSP order and number of frames respectively. T denotes the transpose operation. Similarly, we have $\mathbf{\Delta r_{ik}} = [\Delta r_{ik}(1), \Delta r_{ik}(2),\ldots, \Delta r_{ik}(N)]^T$ and $\mathbf{\Delta^2 r_{ik}} = [\Delta^2 r_{ik}(1), \Delta^2 r_{ik}(2),\ldots, \Delta^2 r_{ik}(N)]^T$ as the k^{th} delta and delta-delta sequences respectively. For each LSP dimension, the static parameters are revised by finding the least-squares solution of $\mathbf{r_{ik}'}$ (Equation (6)). $\mathbf{r_{ik}'}$, $\mathbf{W_\Delta}$ and $\mathbf{W_{\Delta^2}}$ are the refined static vector to be found, linear regression coefficient matrices for delta and delta-delta coefficients respectively. Hence, the regenerated LSP trajectory becomes continuous and smoothly varying, showing similar dynamic change as the models described. Before passing to the Wiener filtering, the LSPs are converted back to their corresponding LPC coefficients.

$$\begin{bmatrix} \mathbf{I} \\ \mathbf{W}_{\Delta} \\ \mathbf{W}_{\Delta^2} \end{bmatrix} \mathbf{r_{ik}}' = \begin{bmatrix} \mathbf{r_{ik}} \\ \Delta\mathbf{r_{ik}} \\ \Delta^2\mathbf{r_{ik}} \end{bmatrix} \tag{5}$$

$$\mathbf{r_{ik}}' = \left(\mathbf{I} + \mathbf{W}_{\Delta}{}^{\mathrm{T}}\mathbf{W}_{\Delta} + \mathbf{W}_{\Delta^2}{}^{\mathrm{T}}\mathbf{W}_{\Delta^2}\right)^{-1}\begin{bmatrix} \mathbf{I} & \mathbf{W}_{\Delta}{}^{\mathrm{T}} & \mathbf{W}_{\Delta^2}{}^{\mathrm{T}} \end{bmatrix}\begin{bmatrix} \mathbf{r_{ik}} \\ \Delta\mathbf{r_{ik}} \\ \Delta^2\mathbf{r_{ik}} \end{bmatrix} \tag{6}$$

2.3. Wiener filtering

With the LPC coefficients, the corresponding Wiener filter for the current frame is derived. Wiener filter is a mean-square-error (MSE) optimal linear filter for signals degraded by additive interference. The frequency response is

$$H(\omega) = \frac{P_{x_i x_i}(\omega)}{P_{xx}(\omega)} \tag{7}$$

where $P_{xixi}(\omega)$ and $P_{xx}(\omega)$ are the power spectral densities of the target source $x_i(n)$ and of the mixture input $x(n)$ respectively. In terms of LPC, $H(\omega)$ becomes

$$H(\omega) = G_{x_i}{}^2 B_{x_i}(\omega) \Big/ G_x{}^2 B_x(\omega) \tag{8}$$

where $B_{x_i}(\omega) = \left| 1 \Big/ 1 + \sum_{m=1}^{M} a_{mx_i}\left(e^{j\omega}\right)^{-m} \right|^2$.

Similarly, $B_x(\omega)$ is defined for the mixture input $x(n)$. G_{xi} and G_x are the gains of excitation for $x_i(n)$ and $x(n)$ respectively. The LPC coefficients $\{a_{mxi}\}$ are obtained from the previous regeneration process and $\{a_{mx}\}$ are computed by typical LPC analysis. To properly extract the target source $x_i(n)$ with $H(\omega)$, the excitation gain G_{xi} needs to be determined. Since $\{a_{mxi}\}$ found by the regeneration process represents just spectral shape without gain, speaking softly or loudly with the same transcription will produce identical vector sequences. Therefore, it is necessary to predict the excitation gain G_{xi} beforehand.

2.4. Gain estimation

In this section, the estimation method for determining the excitation gain G_{xi} will be described. Under the assumption of statistical independence of the two zero mean sound sources, the power spectrum of $x(n)$ is equivalent to the sum of power spectra of the two source signals $x_1(n)$ and $x_2(n)$,

$$|X(\omega)|^2 = |X_1(\omega)|^2 + |X_2(\omega)|^2 \text{, then} \tag{9}$$

$$G_x{}^2 B_x(\omega) = G_{x_1}{}^2 B_{x_1}(\omega) + G_{x_2}{}^2 B_{x_2}(\omega) \tag{10}$$

$$\begin{bmatrix} B_{x_1}(1) & B_{x_2}(1) \\ B_{x_1}(2) & B_{x_2}(2) \\ \vdots & \vdots \\ B_{x_1}(U) & B_{x_2}(U) \end{bmatrix} \begin{bmatrix} G_{x_1}{}^2 \\ G_{x_2}{}^2 \end{bmatrix} = \begin{bmatrix} G_x{}^2 B_x(1) \\ G_x{}^2 B_x(2) \\ \vdots \\ G_x{}^2 B_x(U) \end{bmatrix} \tag{11}$$

where U is the number of frequency bins. By finding the least-squares (LS) solution for Equation (11) with the regenerated spectral envelopes of both sources, the excitation gains G_{x1} and G_{x2} can be optimally estimated in the LS sense. Furthermore, U is generally much larger than the number of unknown to be found (i.e. 2 in our case), the resultant gains are often estimated with sufficient precision.

2.5. Harmonic removal of interfering source

A comb filtering in the form of pitch prediction error filter [17] is incorporated in stage 2, which is used to attenuate harmonic structure associated with the interfering source. For frames where there is no harmonic structure shown in the interfering source, an all-pass filter is applied instead. The impulse response of a pitch prediction error filter $h_p(n)$ is,

$$\begin{aligned} h_p(n) = \delta(n) - \beta_1\delta(n-L) \\ - \beta_2\delta(n-(L+1)) - \beta_3\delta(n-(L+2)) \end{aligned} \tag{12}$$

where L is the lag representing the period of the interfering source. A three-tap filter is adopted, which can deal with non-integral samples of pitch periods. By setting the frequency of a sinusoidal wave $d(n)$ to the pitch of the interfering source, $\{\beta_i\}$ are calculated as the least-squares solution of

$$\begin{bmatrix} d(2) & d(1) & d(0) \\ d(3) & d(2) & d(1) \\ \vdots & \vdots & \vdots \\ d(D-1) & d(D-2) & d(D-3) \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} \cong \begin{bmatrix} d(M+1) \\ d(M+2) \\ \vdots \\ d(M+D-2) \end{bmatrix} \tag{13}$$

where D denotes the number of data samples in $d(n)$. Finally, the estimated source $x_i'(n)$ is reconstructed by the overlap-add method. Gain estimation is essential, since the Wiener filtering is on a frame-by-frame basis and appropriate G_{x1} and G_{x2} will maintain suitable power level. Otherwise, estimated speech sources after the overlap-add method would be unnatural with all normalized frames putting together.

3. Experimental results and discussions

The proposed model-based separation system is evaluated on a data set, which contains eight real speech source signals. The speech materials are not domain specific. Speech source signals with similar lengths are artificially mixed together and four mixture signals are obtained. The source signals are scaled to give identical signal power before mixing. Fig. 3 plots the spectral envelope of one such speech source frame, overlaid with the one obtained after regeneration of model parameters from a mixed signal. When comparing the two spectra, similar spectral envelopes were observed.

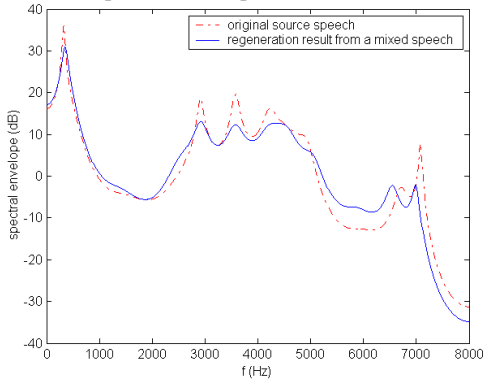

Figure 3: *Plot of spectral envelope versus frequency. The speech model aligned is the tonal final /iL/ .*

The system performance is evaluated in terms of the mean of segmental signal-to-interference ratio (segSIR) suggested by Brown [5]. For a given frame, segSIR is computed as the product of $2/\pi$ and the inverse tangent on the power ratio of signal to interference. As a result, the mean segSIR is a value between 0 (no target signal) and 1 (no intrusion). Besides, the Itakura-Saito distortion (d_{IS}) was also measured, which represents the degree of spectral matching in terms of general spectral shape and overall gain offset. The mean segSIR and d_{IS} values before and after passing the proposed algorithm are tabulated in Table 1 and 2 respectively. After passing the two-stage separation, the mean segSIR in all estimation trials

have been shifted towards to the optimal value '1'. Regarding the Itakura-Saito distortion, the proposed algorithm consistently generate better spectral matching with lower d_{IS}.

mixture	target source	segSIR before	segSIR after
A	1	0.51	0.71
	2	0.46	0.67
B	1	0.45	0.69
	2	0.53	0.73
C	1	0.46	0.70
	2	0.51	0.70
D	1	0.48	0.72
	2	0.51	0.73

Table 1. *Mean segSIR before and after segregation.*

mixture	target source	d_{IS} before	d_{IS} after
A	1	4.71	2.47
	2	59.74	3.47
B	1	57.14	2.75
	2	21.09	2.88
C	1	17.28	2.57
	2	58.90	3.08
D	1	31.17	3.30
	2	11.50	2.40

Table 2. *Mean dIS before and after segregation.*

The fundamental frequencies of source signals are useful not only for removing the harmonics, but also for regeneration of the model parameters. In order to determine the pitch lag L of the interfering signal, the output given by stage 1 is pitch tracked using a method similar to [18]. As the target source is much stronger after Wiener filtering, it is assumed that at most one pitch track exists. The extracted pitch track is consequently associated with the target source and the problem of pitch assignment in most multi-pitch tracking algorithms is avoided. For regeneration of model parameters, the optimal alignment is found by looking at the LSP stream only in our current implementation, since during speech frames with extremely low segmental signal-to-interference ratio, tracking multiple pitches with acoustic information only may not be reliable.

In addition to suppressing interference, maintaining the output speech quality without introducing unnecessary distortions is equally important. In informal listening, most estimated speech sources are highly intelligible and sound clear. A formal subjective test and objective quality assessment are under preparation. The proposed algorithm not only works for speech separation, but also for speech enhancement. With a typical noise estimation scheme, the target speech source can be extracted with a slight change of the gain estimation. Also in the current framework, error-free transcription is used, which implies a high quality recognition result. In case no transcription is available, a robust recognition of mixed speech input would be critical for regenerating speech envelope trajectory.

4. Conclusions

By incorporating speech modeling in signal processing techniques, an algorithm for separating mixed speech sources is proposed. It consists of envelope matching and removal of harmonics of interfering sound source (or enhancing the target harmonics). Speech recognition models are used to estimate the optimal alignment of model states with given transcriptions. With the estimated spectral envelope, the interference is suppressed by Wiener filtering. The pitch of the interfering speech is estimated and corresponding harmonic structure is suppressed by comb filtering. Experimental results show that the proposed algorithm can generate highly intelligible and clear speech estimates, with interfering sources being considerably reduced.

5. References

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *J. Acoust. Soc. Amer.*, vol. 25, pp. 975-979, Sep. 1953.

[2] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*, London: The MIT Press, 1990.

[3] A. S. Bregman, "Psychological data and computational ASA," in *Computational Auditory Scene Analysis*, D. F. Rosenthal and H. G. Okuno, Eds. New Jersey: Lawrence Erlbaum Associates, 1998.

[4] G. J. Brown and M. Cooke, "Computational auditory scene analysis," *Computer Speech and Language*, vol. 8, pp. 297-336, Oct. 1994.

[5] G. J. Brown, "Computational auditory scene analysis: A representation approach," Ph.D. dissertation, University of Sheffield, 1992.

[6] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, New York: John Wiley & Sons, Ltd, 2001.

[7] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*, England: John Wiley & Sons, Ltd, 2002.

[8] G. J. Brown and D. Wang, "Separation of speech by computational auditory scene analysis," in *Speech Enhancement*, J. Benesty, S. Makino and J. Chen, Eds. New York: Springer, 2005.

[9] J. Barker, A. Coy, N. Ma and M. Cooke, "Recent advances in speech fragment decoding techniques," in *ICSLP*, 2006, 85-88.

[10] G. Hu and D. Wang, "An auditory scene analysis approach to monaural speech segregation," in *Topics in Acoustic Echo and Noise Control*, E. Hänsler and G. Schmidt, Eds. New York: Springer, 2006.

[11] F. Itakura, "Line spectrum representation of linear predictor coefficients of speech signals," *J. Acoust. Soc. Amer.*, vol. 57, pp.S35, 1975.

[12] F. K. Soong and B.-H. Juang, "Line spectrum pair (LSP) and speech data compression," in *Proc. ICASSP*, 1984, pp. 37-40.

[13] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

[14] K. Tokuda, T. Masuko, N. Miyazaki and T. Kobayashi, "Multi-space probability distribution HMM," *IEICE Trans. Inf. & Syst.*, vol. E85-D, no.3, pp. 455-464, Mar. 2002.

[15] C. Huang, Y. Shi, J. Zhou, M. Chu, T. Wang and E. Chang, "Segmental tonal modeling for phone set design in Mandarin LVCSR," in *Proc. ICASSP*, 2004, pp. I-901-I-904.

[16] S. W. Lee, F. K. Soong and P. C. Ching, "An iterative trajectory regeneration algorithm for separating mixed speech sources," in *Proc. ICASSP*, 2006, pp. I-157-I-160.

[17] R. P. Ramachandran and P. Kabal, "Pitch prediction filters in speech coding," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 467-478, Apr. 1989.

[18] M. Wu, D. Wang and G. J. Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Trans. Speech Audio Processing*, vol. 11, pp. 229-241, May 2003.