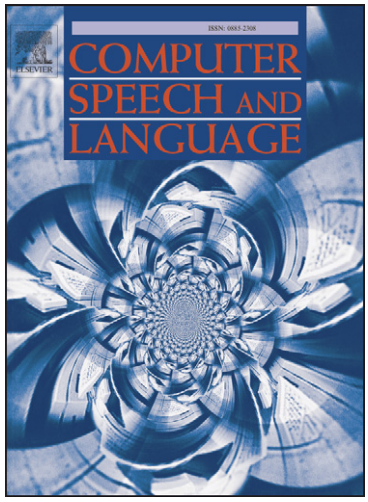
Combining Missing-Feature Theory, Speech Enhancement, and Speaker-Dependent / -Independent Modeling for Speech Separation

Ji Ming, Timothy J. Hazen, James R. Glass



Combining Missing-Feature Theory, Speech Enhancement, and Speaker-Dependent / -Independent Modeling for Speech Separation

Ji Ming, Timothy J. Hazen, James R. Glass

PII: S0885-2308(07)00071-X
DOI: 10.1016/j.csl.2007.12.004
Reference: YCSLA 367

To appear in: *Computer Speech and Language*

Received Date: 29 June 2007
Revised Date: 7 December 2007
Accepted Date: 19 December 2007

Please cite this article as: Ming, J., Hazen, T.J., Glass, J.R., Combining Missing-Feature Theory, Speech Enhancement, and Speaker-Dependent / -Independent Modeling for Speech Separation, *Computer Speech and Language* (2007), doi: 10.1016/j.csl.2007.12.004

Combining Missing-Feature Theory, Speech Enhancement, and Speaker-Dependent / -Independent Modeling for Speech Separation

Ji Ming [a], [1] Timothy J. Hazen [b], James R. Glass [b]

[a] *School of Electronics, Electrical Engineering and Computer Science*
Queen's University Belfast, Belfast BT7 1NN, UK

[b] *Computer Science and Artificial Intelligence Laboratory*
Massachusetts Institute of Technology, Cambridge, MA02139, USA

Abstract

This paper considers the separation and recognition of overlapped speech sentences assuming single-channel observation. A system based on a combination of several different techniques is proposed. The system uses a missing-feature approach for improving crosstalk/noise robustness, a Wiener filter for speech enhancement, hidden Markov models for speech reconstruction, and speaker-dependent/-independent modeling for speaker and speech recognition. We develop the system on the Speech Separation Challenge database, involving a task of separating and recognizing two mixing sentences without assuming advanced knowledge about the identity of the speakers nor about the signal-to-noise ratio. The paper is an extended version of a previous conference paper submitted for the challenge.

Key words: speech separation, speech enhancement, speaker modeling, speech recognition, missing-feature theory, posterior union model

[1] Corresponding author: tel: 028 90971705/90974723; fax: 028 90971702; email: j.ming@qub.ac.uk

1 Introduction

There are currently two major approaches to speech enhancement. One approach assumes the availability of single-channel data (i.e., the speech and noise are available only in a single mixed form), and the other assumes the availability of multi-channel data (i.e., the speech and noise are available in different combination forms, from a set of two or more spatially-distributed transducers). Current single-channel techniques include optimal filtering, for example, spectral subtraction, Wiener filtering, Kalman filtering, or subspace decomposition (see, for example, Boll, 1979; Lim & Oppenheim, 1979; McAulay & Malpass, 1980; Ephraim & Trees, 1995; Gannot, Burshtein & Weinstein, 1998; Jensen & Heusdens, 2007). Other existing single-channel techniques include optimal estimation, for example, minimum mean-square error or maximum *a posteriori* estimators (see, for example, Ephraim & Malah, 1984; Ephraim, 1992; Sameti et al., 1998; Lotter & Vary, 2005; Hendriks & Martin, 2007). Most techniques produce estimates of short-time speech spectra by filtering the noise components, based on knowledge about the statistics (e.g., power spectra, variances, or signal-to-noise ratio) of the noise and speech. When statistics of the noise are not available, they are predicted using previous data without significant speech content (e.g., Martin, 2001; Cohen, 2003). These algorithms work for stationary or slowly-varying noise, but less so for speech-like or heavily nonstationary noise. This is because of the weak predictability of fast-varying noises.

In some applications (e.g., meeting-room or car environments), it is possible to place several microphones to simultaneously record speech and background sounds. Based on the multi-channel data, assuming mutual independence between the sources, it is possible to separate the individual source signals without having to assume prior information. The approach, so called blind source separation, has been studied in speech enhancement as a means of removing

the requirement for prior information about the noise (see, for example, Cichocki & Ehlers, 2007). The multi-channel approach is not the focus of this paper.

In this paper, we study the problem of separating and recognizing overlapped speech sentences assuming single-channel data. In this research, the background noise is crosstalk speech. This problem is challenging not only because the noise is nonstationary, but also because the noise has characteristics of speech signals. It could be more difficult to separate this type of noise from the targeted speech than for other non-speech noises. In the paper, we describe an approach that combines several techniques as a possible solution. We develop the approach on the Speech Separation Challenge database (Cooke & Lee, 2008), involving a task of separating and recognizing two overlapped sentences spoken by the same or two different speakers, assuming the availability of only single-channel data. Our proposed system includes speaker dependent and independent modeling for speaker and speech recognition, missing-feature processing for crosstalk and noise robustness, Wiener filtering for speech enhancement, and hidden Markov models (HMMs) for speech reconstruction.

The remainder of the paper is organized as follows. Section 2 provides an overview of the proposed system for the speech separation challenge, for separating and recognizing two mixing sentences given single-channel data. Section 3 presents the details of the algorithms used to implement the system. Speech separation experiments are described in Section 4, followed by a summary in Section 5.

2 Overview of Proposed System

Fig. 1 illustrates the structure of the proposed system. The input speech waveform is divided into short-time frames, denoted by w_t. Each w_t is a mixed signal

of target and masker, of unknown speaker identities and an unknown target-to-masker ratio. For convenience, we note the sentence with a higher energy ratio as the primary sentence, and the sentence with a lower energy ratio as the secondary sentence. Here, the energy ratio between the two sentences can be defined in the same way as conventional global signal-to-noise ratio, by treating one sentence as speech and the other as noise. Our system separates the two sentences in five steps, operating in sequence.

In Step 1, the system aims to identify the primary sentence by exploiting its higher energy ratio and hence potentially higher recognition accuracy. In the recognition, the lower energy secondary sentence is treated as noise. A speaker-dependent (SD) recognition system is used to select and recognize the primary sentence from the mixed signal. The SD system contains a set of acoustic HMMs for each speaker, with each HMM modeling an appropriate lexical unit for the speaker. For the speech separation challenge task, whole-word HMMs are used. Each HMM is a subband union model (Ming, Lin & Smith, 2006), which uses a missing-feature approach to improve the robustness to the crosstalk noise. It is assumed that the HMM sequence for the primary sentence, matching both the word sequence and speaker characteristics, is likely to produce maximum probability due to the higher energy ratio (and hence smaller distortion) of the primary sentence, and due to the speaker discrimination and improved noise robustness of the HMMs. Thus, the SD recognizer producing the highest probability HMM sequence is selected, with the HMM sequence defining a most-likely primary sentence. This approach is also applied to the situations in which the two mixed sentences have similar energy ratios. As observed in our experiments, when the global signal-to-noise ratio between the two sentences is 0 dB, the system chooses between the two sentences randomly, depending on which sentence produces a higher global probability.

In Step 2, the primary sentence recognized in Step 1 is reconstructed using

an algorithm exploiting the most-likely state sequence of the sentence. The reconstructed speech process consists of short-time spectral estimates $\hat{X}_t^{<1>}$, and the corresponding waveform estimate $\hat{x}_t^{<1>}$. Here, we use the superscript $< 1 >$ to note the variables associated with the primary sentence. Likewise, later we the superscript $< 2 >$ to note the variables associated with the secondary sentence. The short-time spectral estimates $\hat{X}_t^{<1>}$ will be passed to Step 3 for enhancing the secondary sentence, described below.

In Step 3, a Wiener filter is used to enhance the signal of the secondary sentence by filtering out the primary sentence from the mixed signal. The short-time spectral estimates for the primary sentence, produced in Step 2, are used in the operation. The operation takes the short-time spectra of the mixed signal W_t as input, and generates enhanced short-time spectra $\hat{W_t^{<2>}}$ for the secondary sentence.

In Step 4, speech recognition is performed on the enhanced signal for the secondary sentence. A speaker-independent (SI) system is used for the recognition, which consists of an acoustic-linguistic HMM trained using data from all the speakers. The SI system is again a subband union model, for improving robustness to the residual noise in the enhanced signal. The use of an SI system in place of the SD system is found to be important for the recognition – for greater robustness to the distorted speaker characteristics in the enhanced signal, caused by the Wiener filtering operation.

In Step 5, the secondary sentence recognized in Step 4 is reconstructed, using an algorithm similar to that for reconstructing the primary sentence. Again, the reconstructed speech process consists of short-time spectral estimates $\hat{X}_t^{<2>}$ and the corresponding waveform estimate $\hat{x}_t^{<2>}$.

By this process, the system produces speech recognition results for both the primary and secondary sentences. The system therefore implements a "complete" separation process: taking the mixed speech waveform as input, and

producing separated target and masker waveforms as output, along with the speech recognition results for both mixing sentences. In the following section we describe each component of the system in more detail.

3 More Details of Proposed System

3.1 Subband union model for recognition

The subband posterior union model (Ming, Lin & Smith, 2006) is used to build both the SD and SI recognition components. As shown in Fig. 1, they have input y_t and $y_t^{<2>}$, respectively, both representing a short-time feature vector consisting of subband features. The union model is a missing-feature approach, aiming to focus the recognition on uncorrupted subbands thereby improving the robustness to crosstalk interference and/or noise. Let $y = \{y(1), y(2), ..., y(B)\}$ be a frame feature vector, consisting of B independent subbands $y(b)$, subject to crosstalk and/or noise corruption. The union model is used to select the clean or usable subbands for recognition. Without assuming prior information about the corruption, the reliable subbands may be defined as the subbands that maximize the probability of the state for frame y. Denote by $\hat{y}$ an estimate for the reliable subbands, which is a subset of y, then

$$\hat{y} = \arg\max_{z \subset y} p(s|z) \tag{1}$$

where $p(s|z)$ is the probability of state s given subband feature set z. Using Bayes' Rules $p(s|z)$ can be expressed as

$$p(s|z) = \frac{p(z|s)p(s)}{\sum_{s'} p(z|s')p(s')} \tag{2}$$

where $p(z|s)$ is the state-conditioned probability of z, $p(s)$ is a state prior, and the summation in the denominator is over all possible states for frame z. For clean-data trained HMMs, clean data are most likely to produce maximum probabilities for the correct states. Therefore, it is possible to find the clean or reliable subbands by selecting the subbands that maximize the probability of a potential state, as implemented in Eq. (1).

Searching for the optimal set of reliable subbands to maximize the state probability can be computationally expensive, of a complexity $O(2^B)$, for a system using a large number of subbands B. This problem can be relieved by replacing the probability $p(\hat{y}|s)$, for the sought optimal set $\hat{y}$, with the probability of the union of all subsets in y of the same size as $\hat{y}$. Assuming that $\hat{y}$ contains Q subbands, the union probability can be expressed as (Ming, Jancovic & Smith, 2002)

$$p(\bigcup_{z \subset y_B^Q} z|s) \propto \sum_{z \subset y_B^Q} p(z|s) \tag{3}$$

where y_B^Q is the collection of all subsets of Q subbands chosen from the full B subbands in y, and the proportionality is due to ignoring the joint probabilities between the different subsets. Since Eq.(3) contains marginal probabilities of all possible feature subsets, it contains the marginal probability of the optimal feature subset that can be assumed to dominate the sum because of the best data-model match. Therefore, Eq.(3) can be used in place of $p(\hat{y}|s)$ for maximum-probability based recognition. Note that the union probability is not a function of the identity of $\hat{y}$ but only a function of the size of $\hat{y}$. Therefore, substituting Eq. (3) into Eq. (2) for $p(z|s)$, we reduce the problem of finding the optimal set of reliable subbands to finding the optimal number of reliable subbands, but not the exact set, resulting in a lower complexity $O(B)$. This can be expressed as

$$\hat{Q} = \arg\max_{Q} p(s|Q, y) \tag{4}$$

where, by definition,

$$p(s|Q,y) = \frac{\sum_{z \subset y_B^Q} p(z|s)p(s)}{\sum_{s'} \sum_{z \subset y_B^Q} p(z|s')p(s')} \tag{5}$$

As noted by us (Jancovic, 2002) and independently by Chan & Siu (2005), assuming independence between the subbands, an efficient, recursive algorithm exists for calculating the union probability Eq. (3) for Q from 1 to B. The above model, named posterior union model, can be incorporated into an HMM by replacing the conventional state-emission probability with the state probability optimized for the number of reliable subbands, i.e., $\max_Q p(s|Q,y)$ (Ming, Lin & Smith, 2006). Operating on a frame-by-frame basis, the optimal subband selection offers robustness to nonstationary corruption.

3.2 HMM-based speech reconstruction

An algorithm is developed for reconstructing the short-time spectral sequences $\hat{X}_t^{<1>}$, $\hat{X}_t^{<2>}$, and waveforms $\hat{x}_t^{<1>}$, $\hat{x}_t^{<2>}$, of the primary and secondary sentences based on the recognition results from the SD and SI components (Step 2 and 5). In training the SD/SI subband HMMs, a prototype spectrum – suitable for speech reconstruction – is estimated for each HMM state or mixture component using the training data frames assigned to the state or mixture component. In the system, the average log FFT magnitude, taken over all the training frames within the state or mixture component, is used as the prototype spectrum (codeword). Consider the SD recognition component. Denote by $A_{m,i}$ the codeword for speaker m in state i. Given a mixed test sentence w_t, $t = 1, 2, ..., T$, the subband SD model produces an estimate for the primary speaker/sentence, which can be represented by $\hat{m}$ for the speaker and $\hat{s}_t$, $t = 1, 2, ..., T$, for the most-likely state sequence of the primary sentence spoken by the speaker. The $\hat{m}$ and $\hat{s}_t$ can be used to retrieve a clean codeword sequence $A_{\hat{m},\hat{s}_t}$, $t = 1, 2, ..., T$, for reconstructing the spectra and waveform of

the primary sentence, thereby separating the sentence from the mixed signal. Let $\ln \hat{X}_t^{<1>} = A_{\hat{m},\hat{s}_t}$ represent the estimate of the short-time log FFT magnitude of the primary sentence. The corresponding waveform estimate, $\hat{x}_t^{<1>}$, can be obtained from $\hat{X}_t^{<1>}$ by an inverse FFT, assuming that the short-time phase can be approximated by the phase of the mixed frame, P_t (Lim & Oppenheim, 1979).

The above method, modified slightly, can be applied within the SI recognition component for reconstructing the signal of the secondary sentence based on the SI recognition result. The difference is that in the SI model a codeword is estimated for each mixture component within each state, thereby obtaining a good resolution for reconstructing the speaker individualities. Denote by $A_{k,i}$ the codeword for mixture component k in state i. The maximization described in Section 3.1, for estimating the reliable subbands, can be moved inside the state and applied over the individual mixture components, to obtain a most-likely mixture component for each given frame for reconstruction. Let $y^{<2>}$ denote an input frame consisting of subband features for the SI HMM system. The maximized state probability, used as the state-emission probability within the system, is defined as

$$\max_Q p(s|Q) = \sum_k \max_Q p(s,k|Q,y^{<2>}) \tag{6}$$

where $p(s,k|Q,y^{<2>})$ is the union-based probability of state s and mixture component k given $y^{<2>}$, defined similarly to Eq. (5) as

$$p(s,k|Q,y^{<2>}) = \frac{\sum_{z\subset(y^{<2>})_B^Q} p(z|s,k)p(k|s)p(s)}{\sum_{s',k'}\sum_{z\subset(y^{<2>})_B^Q} p(z|s',k')p(k'|s')p(s')} \tag{7}$$

where $p(z|s,k)$ is the probability of feature set z on state s and mixture component k, $p(k|s)$ is the mixture weight in state s, and $p(s)$ is a prior probability of state s. Given the most-likely state $\hat{s}_t$ for frame $y_t^{<2>}$, the most-likely mixture component can be obtained by choosing the maximum-probability compo-

nent within the state: $\hat{k}_t = \arg\max_{k,Q} p(\hat{s}_t, k|Q, y^{<2>})$. Therefore a codeword sequence $A_{\hat{k}_t,\hat{s}_t}$, $t = 1, 2, ..., T$, addressed jointly by the most-likely state sequence $\hat{s}_t$ and most-likely mixture-component sequence $\hat{k}_t$, can be retrieved as an estimate for the short-time log FFT magnitudes of the secondary sentence: $\ln \hat{X}_t^{<2>} = A_{\hat{k}_t,\hat{s}_t}$. The corresponding waveform estimate $\hat{x}_t^{<2>}$ can be obtained from $\hat{X}_t^{<2>}$ by an inverse FFT, using the short-time phase P_t from the mixed input signal w_t.

3.3 Wiener filtering for speech enhancement

Given the estimate $\hat{X}_t^{<1>}$ for the primary sentence, we can obtain an estimate $\hat{W}_t^{<2>}$ for the secondary sentence by removing $\hat{X}_t^{<1>}$ from the mixed input W_t, assuming all three quantities in the same short-time FFT magnitude format. The enhanced signal $\hat{W}_t^{<2>}$ is then used as the input for the SI component for recognizing the secondary sentence. In the system, a Wiener filter is used for the enhancement: $\hat{W}_t^{<2>}(f) = H_t(f)W_t(f)$. The short-time filter function has a simple form:

$$H_t(f) = \frac{P_{\hat{W}_t^{<2>}}(f)}{P_{W_t}(f)} \tag{8}$$

where $P_{W_t}(f)$ is a smoothed periodogram of the mixed input signal w_t, and $P_{\hat{W}_t^{<2>}}(f)$ is a smoothed periodogram of the secondary sentence estimated using the following spectral subtraction

$$P_{\hat{W}_t^{<2>}}(f) = P_{W_t}(f) - [g\hat{X}_t^{<1>}(f)]^2 \tag{9}$$

where $[\hat{X}_t^{<1>}(f)]^2$ is the codeword-based periodogram for the primary sentence treated as noise, and g is a gain factor for matching the gain of the codeword to the gain of the primary sentence in the mixed observation $W_t(f)$. In the system, g is decided on a sentence-by-sentence basis, by minimizing

the sentence-level mean square error between $\hat{X}_t^{<1>}(f)$ and $W_t(f)$ over all periodogram bins and frames:

$$g = \arg\min_{g'} \sum_{t=1}^{T} \sum_{f} [W_t(f) - g' \hat{X}_t^{<1>}(f)]^2 \tag{10}$$

Solving Eq. (10) results in

$$g = \frac{\sum_{t=1}^{T} \sum_f W_t(f) \hat{X}_t^{<1>}(f)}{\sum_{t=1}^{T} \sum_f [\hat{X}_t^{<1>}(f)]^2} \tag{11}$$

It is assumed that $P_{\hat{W}_t^{<2>}}(f) = \alpha P_{W_t}(f)$ if the subtraction in Eq. (9) results in a negative value, where α defines the maximum attenuation. An $\alpha = 0.3$ is used in the system.

4 Experimental Results

4.1 Database and acoustic-linguistic modeling

The above system has been tested on the Speech Separation Challenge database (Cooke & Lee, 2008), containing a two-talker speech recognition task. The database consists of 34 speakers (16 female, 18 male). The sentences by each speaker have a command-like form, all of an identical grammatical structure: <command:4> <color:4> <preposition:4> <letter:25> <digit:10> <adverb:4>, where the number in the brackets indicates the number of choices at each point. Of the six words forming a sentence, the color, letter and number are defined as the keywords for recognition. For each speaker, 500 sentences are available for training. For testing, pairs of sentences, one being treated as "target" and the other being treated as "masker", are mixed at different target-to-masker ratios (TMRs) to form the test sentences. The database provides test data at 7 different TMRs: 6, 3, 0, -3, -6, -9 dB and clean, where

"clean" corresponds to the test data without masker speech. Each test TMR condition contains 600 test sentences, of which, one third are masked by the same talker, one third are masked by talkers of the same gender, and the remaining are masked by talkers of different genders. By definition of the database, of the two mixing sentences forming a test case, one will contain the word "white". This is the target sentence. The recognition task is to identify the letter and number in the target sentence.

In our experiments, we use a 13-subband, 39-stream feature vector to represent each frame. This frame vector is derived from the output of a 27-channel mel-warped filter bank detailed as follows. The speech signal, sampled at 25 kHz, is divided into frames of 20 ms at a frame period of 10 ms. Each frame is analyzed by a 512-point FFT, followed by a 27-channel mel-warped filter bank producing 27 log-scale energies. The 27 log filter-bank energies are then passed to a high-pass filter $H(z) = 1 - z^{-1}$ for decorrelation (Nadeu, Hernando & Gorricho, 1995), resulting in 26 decorrelated log filter-bank energies (DLFBE). The final frame feature vector, i.e., y_t and $y_t^{<2>}$, is formed by grouping the 26 DLFBE uniformly into 13 subbands, with the addition of the first-order and second-order derivatives for each subband, resulting in a 13-subband, 39-stream frame feature vector for modeling by the SD and SI union models for recognition. The 257 short-time FFT magnitudes derived from the FFT are used to form the codewords, associated with the states/mixture components of the SD/SI model, for speech reconstruction.

Each word is modeled by a 14-state left-to-right HMM without state skipping, with one mixture per state in the SD model (trained using speaker-specific data) and 32 mixtures per state in the SI model (trained using all speaker's data). Each mixture component is a Gaussian density with a diagonal covariance matrix. Both the SD and SI recognizers adopt a word bigram language model applied to the Viterbi algorithm for finding the most-likely state sequence given a test signal. The language models reflect the grammatical con-

straint defined above for identifying the primary/secondary sentences; the SI recognizer is additionally subjected to a no-repetition constraint in identifying the secondary sentence, i.e., the keywords that have been recognized for the primary sentence are not assumed to occur again in the secondary sentence. This is indicated in Step 4, Fig. 1, as an additional input containing the disallowed primary keywords into the SI component. To cope with the condition that there may be only one sentence/speaker in the test signal, a silence state, trained using data without speech and allowed to have an unlimited number of self loops, is included in the SI model to absorb the signal from the Wiener filter with the only sentence being removed from the input signal.

In the following we describe two separation experiments. The first shows the system for recognizing the target sentence containing the specified keyword. The second shows the system for recognizing and reconstructing both mixing sentences.

4.2 Recognizing target sentence

Of the two mixing sentences, the target sentence contains keyword "white". The task is to recognize the remaining keywords, letter and number, in this target sentence. We achieve this by simultaneously identifying the target sentence and recognizing the target keywords using the system described in Fig. 1. We run the recognition with two system configurations. In the first configuration, the language model for the SD recognition component forces the word "white" while the language model for the SI recognition component disallows the word "white". This produces two recognized sentences, with respective probability scores $p_{SD(w)}$ (for the primary sentence from the SD component with word "white"), and $p_{SI(no\_w)}$ (for the secondary sentence from the SI component without word white). In the second configuration, the language models for the SD and SI components are swaped, i.e., SD disallowing word "white"

while SI forcing word "white". This produces two new recognized sentences, with respective probability scores $p_{SD(no\_w)}$ (for the primary sentence without word "white"), and $p_{SI(w)}$ (for the secondary sentence with word "white"). Then a decision is made to choose either the first or second configuration result as output dependent on which of the joint probabilities, $p_{SD(w)}p_{SI(no\_w)}$ or $p_{SD(no\_w)}p_{SI(w)}$, is greater. Table 1 presents the recognition results by the system.

4.3 Recognizing and reconstructing both mixing sentences

The following describes further experiments of using the proposed system to recognize and reconstruct *both* mixing sentences from the mixed signal. The task is slightly different from the above task in that we do not aim a specific sentence; instead, we aim to recognize the keywords for each of the mixing sentences. In the experiments, we run the system only once for each mixed test signal, using the more "general" language models described in Section 4.1 without aiming a specific sentence. Also, we consider all three keywords, color/letter/number, in the recognition instead of two keywords in the above experiments. As described in Section 2, the proposed system is capable of producing both speech recognition results and reconstructed speech waveforms simultaneously for both mixing sentences.

For each test signal, the system produces two recognized sentences O_1, O_2. We compare these two recognized sentences with the transcripts of the two input sentences I_1, I_2. There are two possible matches between them: (1) $O_1 - I_1$, $O_2 - I_2$, and (2) $O_1 - I_2$, $O_2 - I_1$. The match producing a higher overall word accuracy rate is used to summarize the final results. Table 2 shows the word accuracy rates for color/letter/number in the target and masker sentences, respectively, produced by the system.

As indicated in Fig. 1, the proposed system uses the union model, Wiener filtering, and speaker-independent modeling for improved separation and recognition performance. To understand the contribution of each of these components, we rerun a set of experiments from using a basic system without these components, to using a refined system with these components added one after another, till the final system. Table 3 shows the improvement on the word accuracy rates for the target and masker sentences, averaged over all the talkers. With reference to Fig. 1, the basic system uses the same speaker-dependent models (in Step 1) for recognizing both the primary and secondary sentences; it uses the full set of subband features for recognition and has no Wiener filter. Therefore, it separates the two sentences by just disallowing repetition of the primary keywords in the secondary sentence. The use of union model for the speaker-dependent models allows the selection of optimal set of subbands for recognition, which reduces the crosstalk noise and offers improved recognition accuracy throughout all noisy conditions. The addition of Wiener filtering improves recognition accuracy for the sentences with low signal-to-noise ratios. For example, it increase the accuracy rate for the target sentences at TMR = -9 dB, from 26.7% to 31.8%, and the accuracy rate for the masker sentences at TMR = 6 dB, from 38.2% to 42.1%. The improvement, however, is smaller for the sentences with high signal-to-noise ratios. This is because the filtering operation tends to alter the characteristics of the speaker while removing the crosstalk noise, thereby causing a mismatch between the speaker-dependent model and the filtered signal for recognition. Replacing the speaker-dependent model with a speaker-independent model may help reduce the mismatch and thereby gain more benefit from the noise reduction. This is evident in Table 3, which shows that the use of a union-based speaker-independent model improves the recognition performance for the filtered signals.

Finally, Fig. 2 shows an example of the reconstructed signals for the target and masker sentences, generated by the codeword-based algorithms described

in Section 3.2. More examples of the reconstructed signals in a WAV format can be found in (Ming, Hazen & Glass, 2007).

5 Summary

This paper described a system for the separation and recognition of two overlapped sentences, given only single-channel data. The system was built upon a combination of several different techniques, aiming to exploit simultaneously the speaker, energy ratio, language-model constraint, training data and acoustic model information, enhanced by the missing-feature theory for ignoring mismatches, to identify and separate the two mixing sentences. The system was tested on the two-talker database from the Speech Separation Challenge, and showed useful improvements. Some of the techniques used in the system were applied earlier to speaker verification (Ming, Hazen & Glass, 2006).

References

[1] Boll SF. Suppression of acoustic noise using spectral subtraction. IEEE Transactions on Acoustics, Speech, and Signal Processing 1979; 27:113-120.

[2] Chan A, Siu M. Efficient computation of the frame-based extended union model and its application in speech recognition against partial temporal corruptions. Computer Speech and Language 2005; 19:301-319.

[3] Cichocki A, Ehlers F, editors. Advances in blind source separation. EURASIP Journal on Advances in Signal Processing 2007; special issue.

[4] Cohen I. Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging. IEEE Transactions on Speech and Audio Processing 2003; 11:466-475.

[5] Cooke M, Lee T-W. The 2006 speech separation challenge. Computer Speech and Language 2008.

[6] Ephraim Y. A Bayesian estimation approach for speech enhancement using hidden Markov models. IEEE Transactions on Signal Processing 1992; 40:725-735.

[7] Ephraim Y, Malah D. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. IEEE Transactions on Acoustics, Speech, and Signal Processing 1984; 32:1109-1121.

[8] Ephraim Y, Trees HLV. A signal subspace approach for speech enhancement. IEEE Transactions on Speech and Audio Processing 1995; 3:251-266.

[9] Gannot S, Burshtein D, Weinstein E. Iterative and sequential Kalman filter-based speech enhancement algorithms. IEEE Transactions on Speech and Audio Processing 1998; 6:373-385.

[10] Hendriks RC, Martin R. MAP estimators for speech enhancement under normal and Rayleigh inverse Gaussian distributions. IEEE Transactions on Audio, Speech, and Language Processing 2007; 15:918-927.

[11] Jancovic P. Combination of multiple feature streams for robust speech recognition. Ph.D. Thesis, Queen's University Belfast, 2002.

[12] Jensen J, Heusdens R. Improved subspace-based single-channel speech enhancement using generalized super-Gaussian priors. IEEE Transactions on Audio, Speech, and Language Processing 2007; 15:862-872.

[13] Lim, JS, Oppenheim AV. Enhancement and bandwidth compression of noisy speech. Proceedings of IEEE 1979; 67:1586-1604.

[14] Lotter T, Vary P. Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model. EURASIP Journal on Applied Signal Processing 2005; 7:1110-1126.

[15] Martin R. Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Transactions on Speech and Audio Processing 2001; 9:504-512.

[16] McAulay RJ, Malpass KL. Speech enhancement using a soft-decision noise suppression filter. IEEE Transactions on Acoustics, Speech, and Signal Processing 1980; 28:137-145.

[17] Ming J, Jancovic P, Smith FJ. Robust speech recognition using probabilistic union models. IEEE Transactions on Speech and Audio Processing 2002; 10:403-414.

[18] Ming J, Hazen TJ, Glass JR. A comparative study of methods for handheld speaker verification in realistic noisy conditions. Proceedings of IEEE Odyssey - The Speaker and Language Recognition Workshop 2006.

[19] Ming J, Lin J, Smith FJ. A posterior union model with applications to robust speech and speaker recognition. EURASIP Journal on Applied Signal Processing 2006; Article ID 75390.

[20] Ming J, Hazen TJ, Glass JR. Examples of speech separation challenge results: http://www.cs.qub.ac.uk/∼J.Ming/SpeechSeparation.htm

[21] Nadeu C, Hernando J, Gorricho M. On the decorrelation of the filter-bank energies in speech recognition. Proceedings of Eurospeech 1995; 1381-1384.

[22] Sameti H, Sheikhzadeh H, Deng L, Brennan RL. HMM-based strategies for enhancement of speech signals embedded in nonstationary noise. IEEE Transactions on Speech and Audio Processing 1998; 6:445-455.

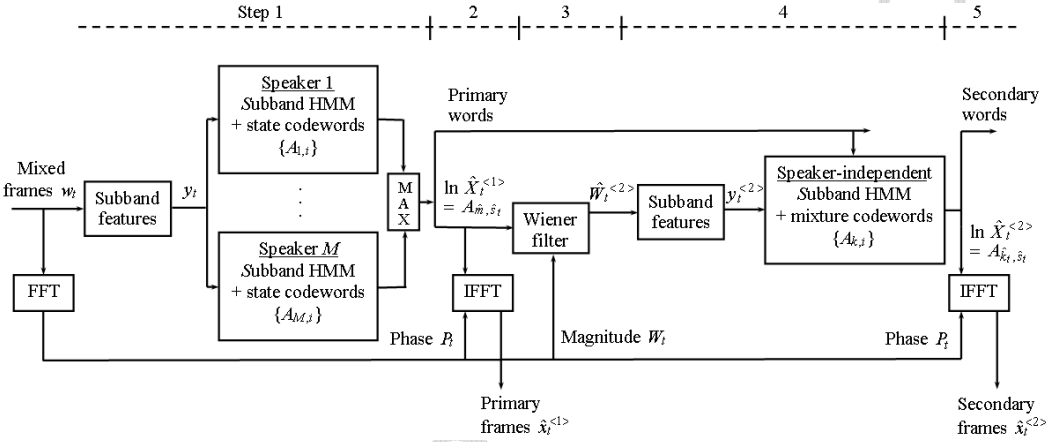


Fig. 1. Schematic diagram of the proposed system for speech separation.

Table 1

Word accuracy rates (%) for recognizing the letter/number keywords in the target sentences containing keyword "white", for different target-to-masker ratios (TMRs), and for two mixing sentences from the same talker (ST), same gender (SG), and different gender (DG).

TMR (dB)	ST	SG	DG	Average
clean				95.17
6	73.08	85.75	86.75	81.42
3	61.54	80.45	82.25	74.08
0	52.49	65.36	72.75	63.08
-3	46.15	56.42	62.75	54.75
-6	38.24	41.89	49.25	43.00
-9	32.81	31.56	38.00	34.17

Table 2

Simultaneous recognition of both mixing sentences, showing the respective word accuracy rates for the target and masker sentences for the color/letter/number keywords.

TMR	Target				Masker			
(dB)	ST	SG	DG	Average	ST	SG	DG	Average
clean				96.94				
6	79.94	89.01	90.00	86.00	48.72	55.87	55.50	53.11
3	70.44	83.79	87.17	80.00	55.35	67.04	69.83	63.67
0	60.03	74.67	80.50	71.22	57.92	74.86	80.83	70.61
-3	54.45	66.67	69.00	62.94	66.67	83.99	88.33	79.06
-6	48.72	54.38	58.50	53.67	75.57	89.76	94.00	85.94
-9	43.59	43.95	48.17	45.22	85.82	93.29	95.67	91.33

Table 3

Contribution of individual techniques, showing improvement on average word accuracy for the target/masker keywords (color, letter, number), from a basic system to the proposed system with the additions of union model, Wiener filtering, and speaker-independent modeling.

TMR (dB)	Basic system	+ Union	+ Wiener filter	+ SI modeling
clean	97.44	96.72	96.67	96.94
6	77.78 / 26.89	84.11 / 38.28	83.89 / 42.17	86.00 / 53.11
3	66.22 / 35.72	76.00 / 48.72	76.39 / 53.44	80.00 / 63.67
0	52.78 / 47.78	63.28 / 62.72	66.06 / 65.22	71.22 / 70.61
-3	38.06 / 65.11	49.28 / 74.78	54.22 / 76.28	62.94 / 79.06
-6	28.89 / 77.89	36.39 / 84.11	40.78 / 84.22	53.67 / 85.94
-9	22.67 / 86.17	26.72 / 90.50	31.83 / 90.39	45.22 / 91.33

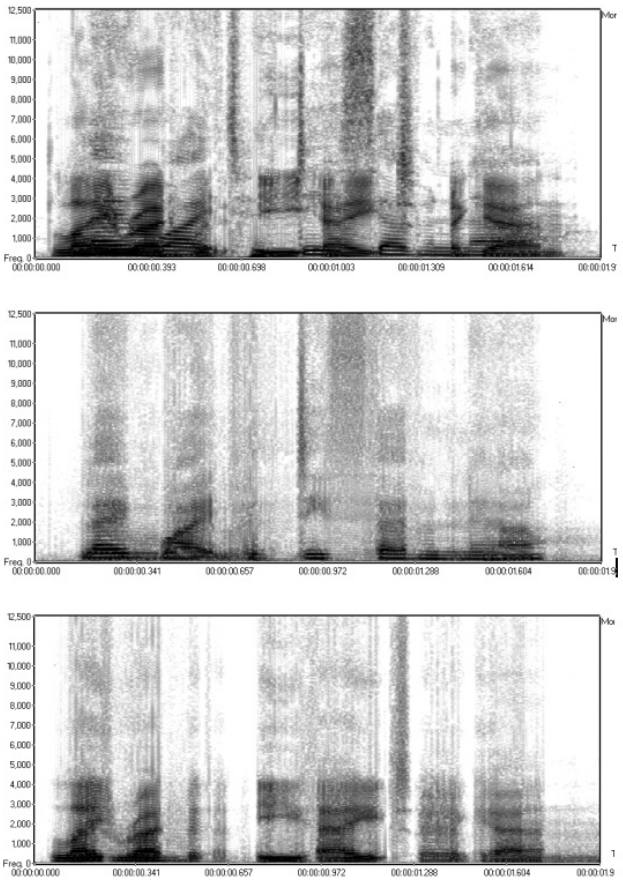


Fig. 2. Separation and reconstruction of sentence t20-lwwd7n-m6-lrwe8a, TMR = 0 dB, DG. From top: mixed signal, reconstructed target sentence, reconstructed masker sentence.