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

Ji Ming[†], *Timothy J. Hazen*[‡], *James R. Glass*[‡]

[†]School of Computer Science, Queen's University Belfast, Belfast BT7 1NN, UK
[‡]MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA 02139, USA
j.ming@qub.ac.uk; hazen/jrg@csail.mit.edu

Abstract

This paper considers the recognition of speech given in the form of two mixed sentences, spoken by the same talker or by two different talkers. The database published on the ICSLP'2006 website for Two-Talker Speech Separation Challenge is used in the study. A system that recognizes and reconstructs both sentences from the given mixture is described. The system involves a combination of several different techniques, including a missing-feature approach for improving crosstalk/noise robustness, Wiener filtering for speech restoration, HMM-based speech reconstruction, and speaker-dependent/-independent modeling for speaker/speech recognition. For clean speech recognition, the system obtained a word accuracy rate 96.7%. For the two-talker speech separation challenge task, the system obtained 81.4% at 6 dB TMR (target-to-masker ratio) and 34.1% at -9 dB TMR.

Index Terms: speech recognition, speech separation, speech enhancement, robustness.

1. Introduction

This paper investigates the problem of recognizing speech assuming that the test signal, from a single channel, is a mixed signal consisting of two overlapped speech utterances. The research is conducted on the Speech Separation Challenge database [1], defined as the 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, for example, "place blue at F 2 now", all of an identical syntactical structure: S = <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 utterances are available for training. For testing, pairs of utterances, 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 utterances. 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 utterances, 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. In experiments, no advanced knowledge of the TMR and speaker identities is assumed. The database also provides an additional clean test set, of 600 utterances, for testing clean speech recognition performance assuming prior knowledge of no masker interference.

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. The database further assumes that the target and masker will not speak the same color/letter/number, although the two may share the other non-keywords in the same test case.

This paper describes a system that recognizes the target keywords through the recognition and reconstruction of both the target and masker sentences. The system involves a combination of several different techniques, including a missing-feature approach for improving crosstalk/noise robustness, Wiener filtering for speech restoration, HMM-based speech reconstruction, and speaker-dependent/-independent modeling for speaker/speech recognition. The combination aims to implement a "complete" separation process: taking the mixed speech waveform as input, and producing separated target and masker waveforms as output, along with the word recognition results for both mixing utterances.

2. Proposed System

2.1. An overview

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 an unknown TMR. 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. The system separates the two sentences in five steps, operating in sequence.

In Step 1, the system aims to identify the primary sentence from the mixed signal by treating the secondary sentence as noise. A speaker-dependent (SD) system, consisting of HMMs for the individual speakers, is used in the recognition to exploit both the speaker and the energy ratio information of the primary sentence. Each speaker HMM is a subband union model [3], implementing a missing-feature technique for reducing the crosstalk noise from the secondary sentence. It is assumed that the model for the primary sentence will produce maximum probability due to the matched speaker characteristics, higher energy ratio, and improved noise robustness.

In Step 2, the spectra and waveform of the primary sentence are reconstructed using an algorithm exploiting the most-likely state sequences of the primary sentence. The reconstructed signals are used for waveform output and for Wiener filtering – for restoring the signal of the secondary sentence by removing the signal of the primary sentence from the mixed input. The Wiener filtering

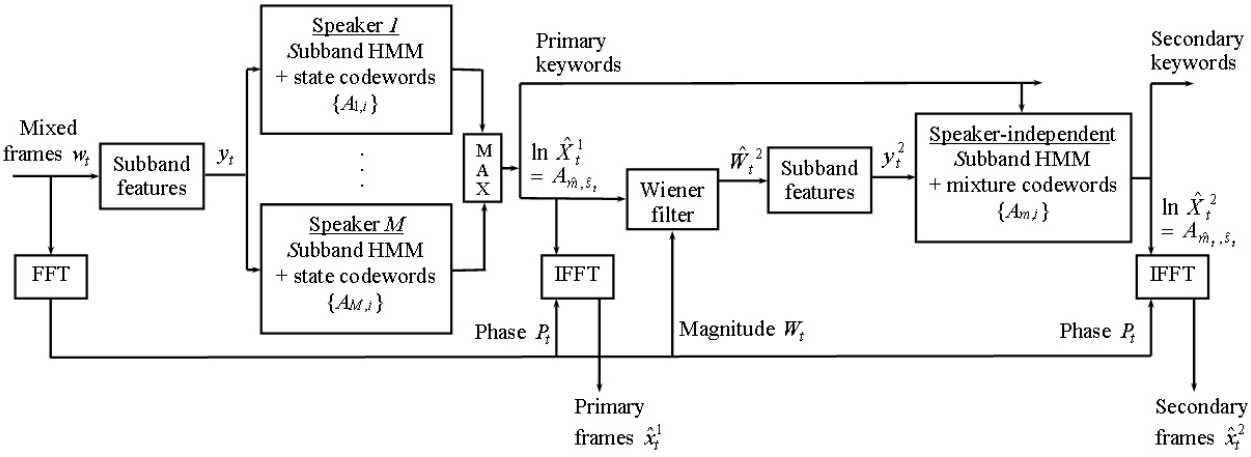

Figure 1: *Schematic diagram of the proposed system for speech separation.*

operation forms Step 3.

In Step 4, the restored signal for the secondary sentence is recognized for words by using a speaker-independent (SI) system, trained using data from all the speakers. The SI system is again a subband union model, for reducing the residual noise in the filtered signal for recognition. The use of an SI system in place of the SD system is found to be important for the filtered signals – for greater robustness to the alteration/mismatch of the speaker characteristics caused by the filtering operation. Step 5 involves the reconstruction of the waveform for the secondary sentence, using an algorithm similar to that for the primary sentence in Step 2. The following describes each component of the system in more detail.

2.2. Subband union model for recognition

The subband union model is used to build both the SD and SI recognition components. The model is a missing-feature method, aiming to focus the recognition on uncorrupted frequency-bands thereby reducing the crosstalk interference/noise on recognition. Let $y = (y(1), y(2), ..., y(B))$ be an input speech frame consisting of B independent subbands $y(b)$ subject to crosstalk/noise corruption (the frame-time subscript is omitted for simplicity). The union model is used to select the clean or usable subbands for recognition. Without assuming prior information on the corruption, the reliable subbands may be defined as the subbands that maximize the probability of the state for y. Denote by $\hat{y}$ the estimate, which is a subset in y, then $\hat{y} = \arg\max_{y' \in y} p(s|y')$, where $p(s|y)$ is the probability of state s given y, defined below:

$$p(s|y) = \frac{p(y|s)p(s)}{\sum_{s'} p(y|s')p(s')} \tag{1}$$

where $p(y|s)$ is the state-conditioned probability of y, $p(s)$ is a state prior, and the summation in the denominator is over all possible states for frame y. For clean-data trained HMMs, the clean data are most likely to produce maximum probabilities for the matched states. Therefore it is likely to find the clean/usable subbands by maximizing the probability of the state for the subbands, as implemented in the above algorithm.

The above model can be incorporated into an HMM by using the maximized state probability, $\max_{y' \in y} p(s|y')$, as the HMM state-emission probability [3]. The maximization for estimating the reliable subbands can be computed efficiently by approximating the probability $p(y'|s)$ in (1), for any subset y', by the probability of the union of all subsets of the same size as y', i.e.,

$$p(y'|s) \propto \sum_{\text{all } y^* \in y, \text{size}(y^*) = \text{size}(y')} p(y^*|s) \tag{2}$$

Note that the union probability (2) is not a function of the identity of subset y' but only a function of the size of y'. This approximation turns the maximization for the identity of the subset of the reliable subbands to the maximization for the size of the subset, which has a much lower complexity. Thus we call the above model the *union* model.

2.3. HMM-based speech reconstruction

An algorithm is developed for reconstructing the spectra $\hat{X}_t^1$, $\hat{X}_t^2$, and waveforms $\hat{x}_t^1$, $\hat{x}_t^2$, of the primary and secondary sentences based on the recognition results from the SD/SI components. 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 utterance 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 by $\hat{s}_t$, $t = 1, 2, ..., T$, 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. Denote the estimate for the short-time log FFT magnitude for the primary sentence as $\ln \hat{X}_t^1 = A_{\hat{m},\hat{s}_t}$. The corresponding waveform estimate, $\hat{x}_t^1$, can be obtained by an inverse FFT of $\hat{X}_t^1$, assuming that the short-time phase can be approximated by the phase of the mixed signal, P_t [5].

The above method, modified slightly, can be applied within the SI 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 characteristics. Denote by $A_{m,i}$ the codeword for mixture component m in state i. The maximization described in Section 2.2, 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 model. The maximized state probability, used as the state-emission probability within the model, is defined as

$$\max_{y' \in y^2} p(s|y') = \sum_m \max_{y' \in y^2} p(s, m|y') \tag{3}$$

where $p(s, m|y)$ is the probability of state s and mixture component m given y, defined similarly to (1) as

$$p(s, m|y) = \frac{p(y|s, m)p(m|s)p(s)}{\sum_{s',m'} p(y|s', m')p(m'|s')p(s')} \tag{4}$$

where $p(y|s, m)$ is the probability of y on state s and mixture component m, $p(m|s)$ 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 component within the state: $\hat{m}_t = \arg\max_{m, y' \in y_t^2} p(\hat{s}_t, m|y')$. Therefore a codeword sequence $A_{\hat{m}_t, \hat{s}_t}$, $t = 1, 2, ..., T$, addressed jointly by the most-likely state sequence $\hat{s}_t$ and mixture-component sequence $\hat{m}_t$, can be retrieved as an estimate for the short-time log FFT magnitudes of the secondary sentence: $\ln \hat{X}_t^2 = A_{\hat{m}_t, \hat{s}_t}$. 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.

2.4. Wiener filtering for speech enhancement

Given the estimate $\hat{X}_t^1$ of 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 model 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:

$$H_t(f) = \frac{P_{\hat{W}_t^2}(f)}{P_{W_t}(f)} \tag{5}$$

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

$$P_{\hat{W}_t^2}(f) = P_{W_t}(f) - \left(g\hat{X}_t^1(f)\right)^2 \tag{6}$$

where $\left(\hat{X}_t^1(f)\right)^2$ 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:

$$g = \arg\min_{g'} \sum_{t=1}^{T} \sum_f \left(W_t(f) - g'\hat{X}_t^1(f)\right)^2 \tag{7}$$

Solving (7) gives

$$g = \frac{\sum_{t=1}^{T} \sum_f W_t(f)\hat{X}_t^1(f)}{\sum_{t=1}^{T} \sum_f \left(\hat{X}_t^1(f)\right)^2} \tag{8}$$

It is assumed that $P_{\hat{W}_t^2}(f) = \alpha P_{W_t}(f)$ if the subtraction in (6) results in a negative value, where α defines the maximum attenuation. An $\alpha = 0.3$ is used in the system.

3. Experimental Results

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 [4], obtaining 26 decorrelated log filter-bank energies (DLFBE). The final frame vector 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 vector for being modeled by the SD/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 and 32 mixtures per state in the SI model. Each mixture component is a Gaussian density with a diagonal covariance matrix. The proposed system is first tested on the clean test set (ssn) for recognizing all three keywords and achieves 96.79% word accuracy rate by the SD model component. The following describes two separation experiments. The first shows the system for recognizing and reconstructing both mixing sentences. The second shows the system for recognizing the target sentence constraining a certain keyword.

3.1. Recognition and reconstruction of two mixing sentences

The experiment considers recognizing both target and masker, with all three keywords, color/letter/number, included in the recognition. In the system, both the SD and SI models are subjected to the syntactical/grammatical constraint S defined in Section 1 for identifying the primary/secondary sentences. The SI model 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. To cope with the condition that there may be only one sentence/speaker in the 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. For each test sentence, the system produces two recognized sentences, one for the target and the other for the masker. There are two possible matches: primary sentence is target/secondary sentence is masker, or vice versa. The closer match, with fewer word errors, is used for scoring. Table 1 shows the word accuracy rates for color/letter/number for the target and masker at different TMRs, calculated based on the algorithm defined in [2].

Fig. 2 shows an example of reconstructing the target and masker signals, produced by the system using the codeword-based algorithm described in Section 2.3. More examples of the reconstructed signals in a WAV format can be found in [6].

Table 1: *Simultaneous recognition of two mixing sentences, showing word accuracy rates (%) for color/letter/number for the target and masker at different target-to-masker ratios (TMRs), for same talker (ST), same gender (SG), different gender (DG) and average.*

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

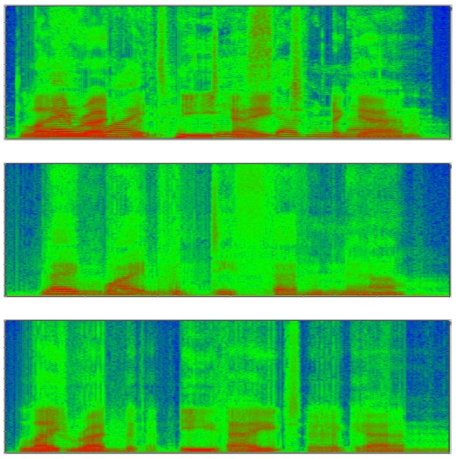

Figure 2: *Separation and reconstruction of sentence t20-lwwd7n-m6-lrwe8a, TMR = 0 dB. From top: mixed signal, reconstructed target, reconstructed masker.*

3.2. Recognizing target

A new experiment is conducted to meet the Challenge requirement – recognizing the letter/number in the target sentence that contains keyword "white". This is slightly different from the above two-sentence recognition task, requiring explicit identification of the target sentence from the mixed signal by correctly recognizing the color keyword. To achieve this, we re-run the recognition of the mixed signal with two recognizer configurations. In the first configuration, the grammar for the SD component forces the word white while the grammar for the SI 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 grammars 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 2 presents the recognition results by the system.

Table 2: *Word accuracy rate (%) for letter/number in recognized target containing word "white".*

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

4. Conclusions

This paper described a system for the recognition and reconstruction of two overlapped sentences, given only the mixed sinal. The system was built upon a combination of different techniques, aiming to exploit simultaneously the speaker, energy-ratio, grammatical constraint, training data and acoustic model information, enhanced by the missing-feature theory for ignoring mismatches. The system was tested on the two-talker database from the Speech Separation Challenge, and was found to perform significantly better than our baseline, 'do-nothing' model. Some of the techniques used in the system were applied earlier to speaker verification [7].

5. References

[1] Cooke, M. and Lee, T.-W., "Speech separation challenge", http://www.interspeech2006.org, 2006.

[2] Ma, N., Lu, Y. and Cooke, M., "Speech separation challenge; baseline recognizer and scoring scripts," http://www.interspeech2006.org, 2006.

[3] Ji, Ming and Smith, F. J., "A posterior union model for improved robust speech recognition in nonstationary noise," ICASSP'2003, pp. 420-423.

[4] Nadeu, C., Hernando, J. and Gorricho, M., "On the decorrelation of the filter-bank energies in speech recognition," Eurospeech'1995, pp. 1381-1384.

[5] Lim, J. S. and Oppenheim, A. V., "Enhancement and bandwidth compression of noisy speech," Proc. of IEEE, Vol. 67, 1979, pp. 1586-1604.

[6] http://www.cs.qub.ac.uk/∼J.Ming/SpeechSeparation.htm

[7] Ji, Ming, Hazen, T. J. and Glass, J. R., "A comparative study of methods for handheld speaker verification in realistic noisy conditions," to appear in IEEE Odyssey 2006.