CASA Based Speech Separation for Robust Speech Recognition

Han Runqiang, Zhao Pei, Gao Qin, Zhang Zhiping, Wu Hao , Wu Xihong

Speech and Hearing Research Center
National Laboratory on Machine Perception
Peking University, Beijing, China
{hanrq, zhaopei, gaoqin, zhangzp, wuhao, wxh@cis.pku.edu.cn}

10.21437/Interspeech.2006-20

Abstract

This paper introduces a speech separation system as a front-end processing step for automatic speech recognition (ASR). It employs computational auditory scene analysis (CASA) to separate the target speech from the interference speech. Specifically, the mixed speech is preprocessed based on auditory peripheral model. Then a pitch tracking is conducted and the dominant pitch is used as a main cue to find the target speech. Next, the time frequency (TF) units are merged into many segments. These segments are then combined into streams via CASA initial grouping. A regrouping strategy is employed to refine these streams via amplitude modulate (AM) cues, which are finally organized by the speaker recognition techniques into corresponding speakers. Finally, the output streams are reconstructed to compensate the missing data in the abovementioned processing steps by a cluster based feature reconstruction. Experimental results of ASR show that at low TMR (<-6dB) the proposed method offers significantly higher recognition accuracy.

Index term: speech separation, CASA, speaker recognition, pitch tracking, units grouping, reconstruction

1. Introduction

The performance of an automatic speech recognition (ASR) system usually degrades drastically in a noisy environment, especially when the background noise is speech. Several methods, such as spectral subtraction [1], advanced Front-End for Distributed Speech Recognition [2], have been proposed to implement robust speech recognition system in noisy environment. These methods significantly improve system performances when the noise is stationary. However, in the presence of time-varying noise, the performance improvement of such methods shrinks. Sometime these methods even provide worse performance.

Speech separation, focusing on separating the target speech from the interference speech, is an obvious way to help ASR under speech masking. It is well known that computational auditory scene analysis (CASA) technique, which emulates the character of human auditory system, can extract target speech from complex background. Hence, CASA approach is a promising way to deal with speech processing problem under multi-speaker condition and its effectiveness has been revealed by recent researches [3], [4].

The most commonly used CASA model in robust speech recognition is the one proposed by Barker, Cooke and Ellis [3], where the top-down information search is performed with bottom-up segmentation and grouping. It combines the CASA with speech recognition, and provides a good performance in complex environments. However, the time and space complexity is an issue to be resolved in this circumstance.

Another popular CASA model proposed by Hu and Wang [4] show advantages on dealing with the high-frequency part of speech. In their algorithm, the resolved and unresolved harmonics in low bands and high bands are treated differently. For resolved harmonics, the system generates segments based on temporal continuity and cross-channel correlation, and groups them according to their periodicities. For unresolved harmonics, it generates segments based on common amplitude modulation (AM) in addition to temporal continuity and groups them according to AM rates [11], [4]. Since pitch is the most significant cue, this model is more suitable for continuous voice speech. However, to the unvoiced speech or the speech composed of both voiced and unvoiced components, the separation becomes more difficult. Furthermore, when there is overlap between the target speech and interference speech in time-frequency domain, the target will also be difficult to separate. When the binary masking is adopted, the target spectral is usually removed.

When applying speech separation to speech recognition, the output of CASA model cannot be directly used as input for ASR system, due to the fact that some target bands are removed in CASA processing, especially when there is overlap between the target speech and interference speech in time-frequency domain. To deal with this problem, some methods (e.g., the state-based imputation and marginalization [5]) have been proposed to perform feature reconstruction. These algorithms modify the manner in which state output probabilities are computed within the recognizer. But the feature used here should be log Mel-filter bank energies since they are less effective than the cepstrum feature. Meanwhile, many post-processing techniques cannot be used properly in this situation. So we only use the output of the CASA to identify whether the log Mel-filter energy is reliable. The unreliable log Mel-filter energy components should be estimated. There are many researches on the problem [5], [6], such as feature reconstruction methods, including correlation-based reconstruction and cluster-based reconstruction, and previous research shows that the latter outperforms the former [6].

Considering the analyses mentioned above, a new speech separation system is developed, where Hu & Wang's algorithm is adopted and modified for the CASA model. For speech composed of voiced and unvoiced components, it is divided into several segments according to pitch breaks. The speaker recognition technique is then employed to streams organizing. The final outputs are processed by cluster-based reconstruction.
The rest of this paper is organized as follows. Section 2 presents the overview of the proposed separation system and the six modules of the proposed system are discussed in details. Section

3 shows the experimental results. Then follow the discussion and conclusion in section 4.

2. System Description

The system, illustrated in Figure 1, is composed of two main parts: CASA based speech separation and cluster based speech reconstruction. The entire system can be divided into six modules. The CASA based speech separation part contains the first five modules, i.e. auditory peripheral model, pitch detection and tracking, initial grouping, speaker recognition, and regrouping, while the cluster based speech reconstruction forms the sixth module. The output of the system, i.e. the reconstructed feature is used as the input of the speech recognizer.

Based on the speech signal processing order, six modules regarding to two parts of the system are described as follows in detail.

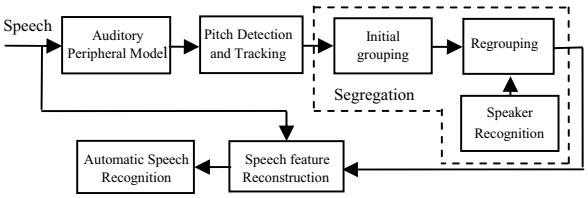

Figure 1 *System scheme*

2.1. CASA based speech separation

Speech signal is imported to the CASA based speech separation procedure and preprocessed via auditory peripheral model as initialization. Pitch detection and tracking modular is then conducted and the detected pitch is used as a major cue to separate speech [8]. The CASA initial grouping module combines the speech segments, which is merged from the time frequency (TF) units, into several speech streams according to pitch information. Then, based on amplitude modulate (AM) cues, the regrouping module is applied to refine those streams, which are finally organized into corresponding speakers by the speaker recognition module. At this stage, the speech signal is separated.

2.1.1. Auditory peripheral model

The gammatone filters, which simulate the cochlear response, are adopted as auditory filters in the system. The response of each filter is further transduced by the Meddis model of inner hair cells. The envelope of the hair cell output is obtained through low-pass filtering.

The system adopts the gammatone filters in the preprocessing includes 128 channels from 80Hz to 5000Hz, and the filter center frequencies are quasi-logarithmically spaced.

2.1.2. Pitch detection and tracking

The task of pitch tracking is to detect the pitch contours that belong to the same speakers. In the speech separation task, two assumptions are set up as follows: (1) in case that the signal-to-noise ratio is high, target speech is dominant, and so is its pitch; (2) the pitch of different speaker differs in frequency, and the pitch contour is coherent. These two assumptions suggest two major features for pitch tracking algorithm: normalized correlogram and frequency dynamic. Let each candidate pitch point be defined as: $p_{i,j} = (a, f)$, where a denotes the normalized correlogram value of the frame index j and frequency index f, and f denotes the frequency. Then the frequency dynamic is calculated by (1)

$$d(f_{i-1}, f_i) = \frac{|f_i - f_{i-1}|}{|f_i + f_{i-1}|} \tag{1}$$

Using dynamic programming, we find the pitch contour with maximum path score

$$S(p) = \sum_{i=1}^{n} a_{i,j} - d(f_i - f_{i-1}), p = \{f_1 \cdots f_n\} \tag{2}$$

After the first pitch contour is obtained, we eliminate the possible harmonic elements of each pitch point, and track the second pitch contour in the same way.

2.1.3. Initial grouping

To build the initial grouping module Hu&Wang's model [4] is adopted and further modified based on our previous analysis in section 1. After decomposition and feature extraction, in this step, an input mixture speech is analyzed by an auditory filter bank in consecutive time frames. This processing of decomposition leads to a two-dimensional time-frequency map (T-F unit). Four kinds of features for each time-frequency unit are utilized:

- Correlogram: autocorrelation of a gammatone filter response in each frame.
- Dominant pitch: the sum of the each channel for one frame.
- Envelop correlogram: autocorrelation of the envelope within each band.
- Cross-Channel correlation: cross correlation between neighbour channels' response.

Based on the pitch temporal continuity and cross-channel correlation of the TF units, they can be merged into segments to capture a perceptually-relevant acoustic component of a single source. Segments are then grouped into an initial foreground stream and a background stream based on domain pitch.

In this processing step, the global pitch is estimated by the previous step, and refined by calculating it in segments labeled to the target group. Pitch is an important cue for low frequency, because harmonics are much easier resolved than higher frequency.

2.1.4. Speaker recognition

This module is the processing for organizing the segments that have no overlap on time, but have the same domain pitch properties for one speaker. The MFCC feature is extracted from the resynthesis [4] speech. The speakers are trained using Gaussian mixture models (GMM) via EM algorithm. Given a speech stream, the speaker recognition technique is used to label the stream by its most likely corresponding speaker on

likelihood. In this way, speech stream could be organized for the same speaker. In the system, MFCC features are used as feature, and 64-component GMM is built for each speaker. The speaker recognition result is the organization cue for regrouping module.

2.1.5. Regrouping

For this processing, the pitch is refined by estimation from the initial foreground stream to obtain higher accuracy. Segments corresponding to unresolved harmonics are generated based on temporal continuity and cross-channel envelope correlation to refine those streams marked in initial grouping step. Amplitude modulation (AM) is then applied to design the high frequency segments' label，since the responses of adjacent channels to unsolved harmonics exhibit very similar AM patterns and their envelops are highly correlated. More detail description can be found in [4]. After this module, the segments are further grouped into the foreground and background stream. With the speaker recognition result, the disconnected stream in time can be grouped to one target.

2.2. Cluster based feature reconstruction

The spectral missing of speech caused by CASA is serious, especially for the unvoiced segments. In our system, the spectral reconstruction technique is employed to retrieve the missing data for the following speech recognition.

The labeled speech stream, the output of the CASA based speech separation part of the system, is used as the input of the Cluster based feature reconstruction module, where the log Mel band energy vector is assumed to be segregated into a number of clusters. Each cluster is assumed to be Gaussian distributed:

$$P(X \mid k) = \frac{\exp\left(-\frac{1}{2}(X-\mu_k)^T \Theta_k^{-1}(X-\mu_k)\right)}{\sqrt{(2\pi)^d |\Theta_k|}} \tag{3}$$

where X represents an arbitrary vector from the k*th* cluster, d represents the dimensionality of X, and μ_k and Θ_k represent the mean vector and covariance matrix of the k*th* cluster, respectively. In a frame, the signal can be divided into K Mel bands. Let $Y(t)$ represent the noisy vector for which the underlying true vector $X(t)$ must be reconstructed. The reliable component vector of $X(t)$, $X_r(t)$, can be approximated by the reliable component vector of $Y(t)$, $Y_r(t)$. The unreliable component vector $X_u(t)$ must be estimated. The overall estimate of the $X_u(t)$ is given by

$$\hat{X}_u(t) = \sum_{k=1}^{K} P(k \mid Y_r(t), X_u(t) \le Y_u(t)) \hat{X}_u^k(t) \tag{4}$$

where $\hat{X}_u^k(t)$ is the estimate for $\hat{X}_u(t)$ of the *kth* cluster's distribution, K is the cluster number. More details can be found in [3].

In our system, the number of clusters K is set to 64. We assume that the output of CASA is part of target, so whether a component of $Y(t)$ is reliable can be judged by estimating the TMR:

$$TMR = 10\log\frac{Mel_k^{CASA(t)}}{Mel_k^{Y(t)} - Mel_k^{CASA(t)}} \tag{5}$$

where $Mel_k^{CASA(t)}$ is the k*th* band energy of the output of the CASA.

3. Experiments and Results

Our system is used to recognize speech from a target speaker in the presence of other speech. 17000 Utterances are provided as training data, which is used to train the models for speaker recognition and models for speech reconstruction. The test data consists of pairs of sentences at a range of TMRs (target-to-masker ratios) from 6 to -9 dB. Only one signal per mixture is provided.

3.1. Recognition test

The recognition test is first conducted to the test speech offered by sponsor without any processing. Another test is conducted on the same data enhanced by advanced Front-End [2], which is based on wiener filtering. The process of recognition for all test data takes about 1 hour on a PC with Pentium IV 1.7GHz CPU.

The test on speech separation of our system is conducted in two manners. One is to recognize the generated speech without reconstruction, the other is to recognize the reconstructed speech. It is important to note that the mixed speech is divided into two utterances in our system. According to the requirement of the evaluation, the target is the one who said "white" at the beginning. Hence, both of the streams separated from the mixture speech are recognized by the recognizer which is offered by sponsor. The stream, which has been recognized to contain the key word "white", is judged to be the target speech. If neither of the sentences begins with "white", the system will select one randomly as the target sentence. The process of the speech separation for all test data takes 5 hours on a 15-node cluster, with Pentium IV 1.7GHz CPU at each node.

3.2. Test Results

The results of above tests are illustrated in the table one to table three respectively.

As to the test on the speech without reconstruction, the recognizer cannot generate any result to most sentences due to the serious mismatch between the processed speech and model of recognizer.

From the results, we find that the conventional speech enhancement has little effect to the speech interference. The system of speech separation cannot deal with the speech recognition when the TMR is higher than -6dB. To the lower TMR, the recognition accuracy is higher than that of unprocessed speech. The results also show that the speech reconstruction is necessary in this application.

The testing data and training data are both distributed by the sponsor [10].

4. Discussion and Conclusion

A speech separation system used for speech recognition is introduced in this paper, aiming to deal with the speech interference problem, where auditory peripheral model, pitch detection and tracking, speaker recognition techniques are

employed. In CASA based speech separation, pitch is the most important cue to be utilized. According to pitch break points, mixed speech is divided into several segments. Within each segment, the time-frequency units from one source are grouped conforming to pitch continuity and consistency. In the segments grouping, speaker recognition is utilized to connect single speaker's segments into one stream.

Speech reconstruction is designed to retrieve the features for speech recognition from the separated stream. The results of experiment show that the reconstruction can overcome the mismatch between the fragmental stream and the recognizer to some extent.

The results of recognition on separated speech are quite lower than that of unprocessed speech. It shows that the method of speech separation will destroy the spectrum of speech especially to the unvoiced segments. On the low TMR speech, the speech separation performs better, showing the effectiveness in this situation.

The system is an attempt on utilizing CASA in ASR, In the future, how to recover the missing data (especially the unvoiced data) is the key problem to be solved. A possible way is to add the unvoiced speech to the separated speech according to phonetics rules or statistical model. In addition, the multipitch tracking is also a challenging work, whose result can offer more accurate cues to speech separation. Another technical issue is the speech reconstruction, which plays an important role in auto recognition to separated speech. In this evaluation the method we followed is not designed specially to the post-processing of CASA. Hence we should find a new way to integrate reconstruction and CASA effectively. Of course, parameters adjustment is also helpful to improve the performance of the system effectively.

5. Acknowledgements

The work was supported in part by NSFC 60435010, 60305030, 60305004, NKBRPC 2004CB318000, a program for NCET, as well as a joint project of Engineering Institute at Peking University.

6. References

[1] Ephraim, Y., "Statistical-model-based speech enhancement systems". Proceedings of the IEEE, 80(10):1526-1555, 1992.

[2] ETSI ES 202 212, "Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Extended advanced front-end feature extraction algorithm; Compression algorithms; Back-end speech reconstruction algorithm," 2003.

[3] Barker, J. P., Cooke, M. P. and Ellis, D. P. W. "Decoding speech in the presence of other sources", Speech Communication 45, 5-25, 2005.

[4] Hu, G. and Wang, D.L., "Monaural speech segregation based on pitch tracking and amplitude modulation". IEEE Transactions on Neural Networks, Vol. 15, 1135-1150, 2004.

[5] Wu, X. H., "Auditory perception mechanism and computational auditory scene analysis", post doctor research report, 1997.

[6] Cooke, M., Green, P., Josifovski, L., Vizinho, A., "Robust automatic speech recognition with missing and uncertain acoustic data". Speech Commun. 34, 267–285. 2001.

[7] Raj, B., Seltzer, ML, Stern, RM, "Reconstruction of Missing Features for Robust Speech Recognition", Speech Communication, Vol. 43, Issue 4, pp.275-296, September 2004.

[8] A. S. Bregman, "Auditory Scene Analysis". Cambridge, MA:MIT Press,1990.

[9] Shao Y. and Wang D.L. "Model-based sequential organization in cochannel speech". IEEE Transactions on Audio, Speech, and Language Processing (formerly IEEE Transactions on Speech and Audio Processing), vol. 14, 289-298, 2006.

[10] Cooke, M. P., Barker, J., Cunningham, S. P. and Shao, X., "An audio-visual corpus for speech perception and automatic speech recognition", submitted to J. Acoust. Soc. Amer. [status: submitted 29 Nov 2005].

[11] R. P. Carlyon and T. M. Shackleton, "Comparing the fundamental frequencies of resolved and unresolved harmonics: evidence for two pitch mechanisms," J. Acoust. Soc. Amer., vol. 95, 3541–3554, 1994.

Table 1. Recognition accuracy on unprocessed test data

TMR	Same Talker	Same Gender	Different Gender	Avg.
-9dB	5.66%	7.26%	7.50%	6.75%
-6dB	9.73%	14.53%	11.50%	11.75%
-3dB	18.10%	20.95%	19.50%	19.42%
0dB	29.64%	32.96%	33.50%	31.92%
3dB	46.15%	44.13%	46.75%	45.75%
6dB	62.44%	64.25%	64.25%	63.58%
clean	——	——	——	98.56%

Table 2. Recognition accuracy on the test data processed by advanced Front-End

TMR	Same Talker	Same Gender	Different Gender	Avg.
-9dB	5.66%	7.54%	7.50%	6.83%
-6dB	9.73%	14.53%	11.50%	11.75%
-3dB	18.10%	20.95%	19.50%	19.42%
0dB	29.64%	32.96%	33.25%	31.83%
3dB	46.15%	43.85%	47.00%	45.75%
6dB	62.90%	64.25%	64.25%	63.75%
clean	——	——	——	98.56%

Table 3. Recognition accuracy on the test data processed by speech separation

TMR	Same Talker	Same Gender	Different Gender	Avg.
-9dB	8.60%	10.89%	11.25%	10.17%
-6dB	9.50%	12.57%	14.00%	11.92%
-3dB	11.54%	16.48%	19.75%	15.75%
0dB	23.30%	27.37%	27.75%	26.00%
3dB	27.15%	29.61%	33.00%	29.83%
6dB	31.45%	37.43%	40.00%	36.08%
clean	64.80%	73.79%	65.13%	67.38%