Monaural Speech Separation Based on Computational Auditory Scene Analysis and Objective Quality Assessment of Speech

Peng Li, Yong Guan, Bo Xu, and Wenju Liu

***Abstract*—Monaural speech separation is a very challenging problem in speech signal processing. It has been studied in previous, and many separation systems based on computational auditory scene analysis (CASA) have been proposed in the last two decades. Although the research on CASA has tended to introduce high level knowledge into separation process from primitive data-driven method, the knowledge of speech quality still has not been combined in it. This makes the performance evaluation of CASA mainly focused on the SNR improvement. Actually, whether the result of the separated speech is good does not relevant directly to its SNR. In order to solve this problem, we proposed a new method which combined CASA with objective quality assessment of speech (OQAS). In the grouping process of CASA, we use OQAS as the guide to instruct the CASA system. Through this combination, the performance of the speech separation can be improved not only in SNR, but also in Mean Opinion Score (MOS). Our system is systematically evaluated and compared with previous systems, and it yields substantially better performance, especially for the subjective perceptual quality of separated speech.**

***Index Terms*—Monaural Speech Separation, Computational Auditory Scene Analysis (CASA), Objective Quality Assessment of Speech (OQAS), Grouping, Segmentation**

I. Introduction

IN a natural world, speech signal is frequently accompanied by other sound sources on reaching the auditory systems, yet listeners are capable of holding conversations in a wide range of listening conditions. This is called the well-known 'cocktail party' effect [1]. It is much valuable to make computer have the ability of human to segregate the object source from other interfered sources. An effective system for separating speech from interfered sources would greatly facilitate many applications, including automatic speech recognition (ASR), speaker identification, audio retrieval, digital content management etc. Therefore, the research of speech separation gradually catches the researchers' attentions and it becomes an increasingly popular theme in the field of signal processing.

General methods for signal separation, such as blind source separation using independent component analysis [2] or sensor arrays for spatial filtering [3] require multiple sensors. However, many applications such as telecommunication and audio retrieval need a monaural solution. Since in monaural separation cases only one sensor signal can be used, it is much harder and still-open problem for researchers to solve it.

While monaural speech separation remains a challenge, the human auditory system shows a remarkable capacity for monaural speech segregation, which spurs the researchers to study human auditory perception much deeply. In 1990, Bregman first proposed the concept of *auditory scene analysis* (ASA) [4]. In his book, he argues that the auditory system segregates the acoustic signal into streams, corresponding to different sources, according to ASA principles. His study on ASA offers a new way to deal with the monaural speech separation. It has also inspired considerable work on *computational auditory scene analysis* (CASA). Many CASA system have been built for speech segregation which is adhere to the known principles of ASA [5], [6], [7], [8], [9], [10], [11]. Such systems generally approach speech segregation without making strong assumptions about the acoustic properties of interference, and it can also separate speech with only one or two channel signals. Generally, CASA system follows two main stages: segmentation (analysis) and grouping (synthesis) [4]. In segmentation, the acoustic input is decomposed into sensory segments, each of which should originate from a single source. In grouping, those segments that likely come from the same source are grouped together.

At the beginning of the research on CASA, researchers are concentrated on the primitive data-driven method. This type of CASA system usually extract cues such as pitch, onset and offset, AM rate, etc., from the input data; and use them to separate the target speech from the mixture. In last decade, the focus of CASA is changing from primitive data-driven method to knowledge-based schema-driven method. More and more knowledge in higher level such as acoustic model using in ASR, source character, source location and so on, is introduced into the primitive CASA system to guide the separation [7], [12], [13]. Although the knowledge-based CASA research has

Manuscript received January 31, 2006. This work was supported by the National Grand Fundamental Research 973 Program of China under Grant 2004CB318105.

Peng Li and Bo Xu are with the High-Tech Innovation Center, Institute of Automation, Chinese Academy of Sciences, Beijing, 95 Zhongguancun East Road, P. R. China (e-mail: pengli@hitic.ia.ac.cn; xubo@hitic.ia.ac.cn).

Yong Guan and Wenju Liu are with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, 95 Zhongguancun East Road, P. R. China (e-mail: yguan@nlpr.ia.ac.cn; lwj@nlpr.ia.ac.cn).

achieved great achievement recently with many new types of knowledge being introduced into CASA systems, the knowledge directly relevant to speech perceptual quality has not been included in them yet.

On the other hand, in most CASA systems, the performance evaluation of CASA system is always based on SNR. Although the SNR of speech after separating is improved and CASA system surely reduce the noises, it does not mean that the speech quality in perception improved with it, too. It is not absolutely right that the higher the SNR of a signal is, the better the perceptual quality of that signal is. So, in order to improve both SNR and perceptual quality of the separated speech, we attempt to seek an effective way to combine the subjective perceptual quality of speech with CASA systems.

As we all know, the speech quality is a subjective opinion, based on the user's reaction to the speech signal they actually heard. Subjective methods make use of a listener panel to measure speech quality on an integer scale from 1 to 5, with 1 corresponding to unsatisfactory speech quality and 5 corresponding to excellent speech quality. The average of the listener scores is the subjective Mean Opinion Score, MOS [14]. This has been the most reliable method of speech quality assessment but it is very expensive and time consuming, making it unsuitable for frequent or rapid application. These shortcomings can be overcome by using objective measurement methods, which replace the listener panel with a computational algorithm. Objective methods aim to deliver MOSs that are highly correlated with the MOSs obtained from subjective listening experiments.

Objective quality assessment tests can be classified as *intrusive* or *non-intrusive*. Intrusive measurement depends on some form of distance metric between the input (clean) and output (degraded) speech signals to predict the subjective MOS. Non-intrusive measurement depends only on the degraded speech signal and is a more challenging approach to objective speech quality estimation. Non-intrusive models have been proposed in [15], [16], [17], but only recently has ITU-T released P.563 as its non-intrusive objective quality measurement standard algorithm [18]. In speech separation application, an intrusive approach may not be applicable because the reference speech signal may be unavailable, so the non-intrusive method is recommended. Just for this reason, we select the P.563 algorithm to help our speech quality assessment finally.

Having confirmed the algorithm of the objective quality assessment, what we should do is just to find an appropriate way to integrate it with the segregation process. With this object and considering the characteristics of primitive CASA system, especially the Hu and Wang model [11], we construct the link between the speech quality and CASA processing. On one hand, we use speech quality assessment to evaluate the segments formed by the segmentation of CASA, so that we can select the better segments which were not affected deeply by interference source and use them to track the pitch contour that could be used as the separating cues. On the other hand, in the final grouping stage, we can also use speech quality assessment to evaluate the segments which can not be divided into foreground stream in the former step, judge the accuracy of the former classification and then adjust the corresponding segments back to foreground to enhance the final performance of the separation.

The organization of this paper is as below. Section II first describes the construction of our model by means of analyzing the appropriate selection of OQAS algorithm and CASA system; then, it gives an overview of the new model and explains each components of the model concisely. The key point of the proposed model, the combination of CASA with OQAS, is elaborated in Section III. In Section IV, the proposed system is systematically evaluated and compared with other systems for speech segregation or enhancement. At last, a further discussion is given in Section V.

II. System Construction and Overview

As discussions above, the object of our study is to use OQAS to improve the performance of CASA system. So, the most important problem of our research is what kind of CASA system and OQAS algorithm should be selected, and how to combine the selected CASA system with the selected OQAS algorithm tightly so as to get better separation performance not only in SNR, but also in perceptual quality.

A. Selection of OQAS Algorithm

In Section I, we have discussed that in speech separation application, because of the absence of input reference signal, we have to select non-intrusive method of objective quality assessment of speech. So the ITU-T P.563 algorithm is selected.

The reason why we select the P.563 algorithm is that, it is recommended by ITU as the standard for objective quality assessment in narrow-band telephony applications. The P.563 algorithm is resulted from a collaboration of Psytechnics' NiQA algorithm [19], SwissQual's NiNA [20], and Opticom's P3SQM. Its signal parameterization is divided in three independent functional blocks corresponding to the main classes of distortion; they are: vocal tract analysis, high additional noise, and speech interruptions, muting and time clippings. A total of 51 characteristic signal parameters are calculated. Based on a restricted set of 8 key parameters, a dominant distortion class is selected. The key parameters and the selected distortion class are used for adjusting the speech quality model. Furthermore, for each distortion class, a linear combination of parameters is used to generate an intermediate quality rating that, together with other additional signal features are combined to calculate the (raw) objective quality score.

OQAS algorithms are usually used to test the quality of narrow-band telephone speeches; therefore, the speeches that need to be processed by our model should be telephone speeches with 8 kHz sampling frequency and 16 bit PCM amplitude resolution.

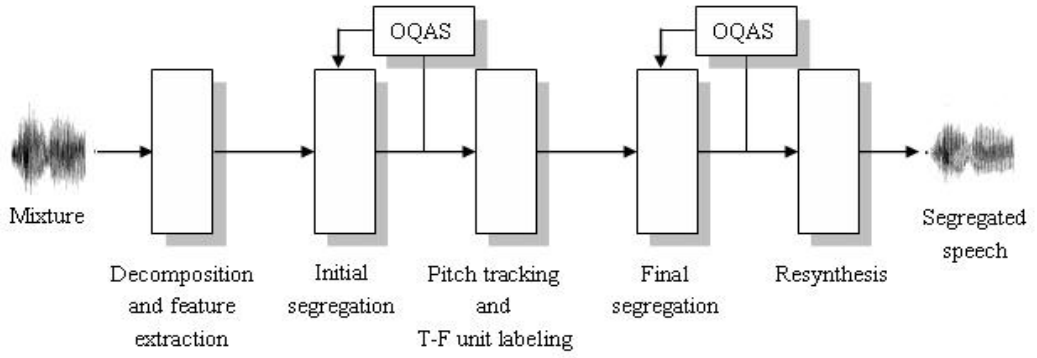

Fig. 1. Schematic diagram of the proposed multistage system

B. Selection of CASA System

The CASA system we employed in proposed model is based on the model of Hu and Wang, which was presented in [11]. Hu and Wang model is a typical primitive CASA model; it uses different segregation methods for resolved and unresolved harmonics. 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. Underlying the segregation process is a pitch contour that is first estimated from speech segregated according to dominant pitch and then adjusted according to psychoacoustic constraints. Based on the above processing, the separation performance of Hu and Wang model is almost the best in primitive CASA system at present, and also better than many other knowledge-based CASA systems in dealing with the voiced speeches.

Another reason why we select Hu and Wang model is that it uses the notion of time-frequency mask. The ideal binary mask is very effective for human speech intelligibility. It is well-defined no matter how many intrusions are in the scene or how many targets need to be segregated and provides an excellent front-end for robust automatic speech recognition. All of these make it easier to combine CASA system with OQAS algorithm.

C. Overview of the Combination System

The main idea of our model is to employ objective quality assessment of speech to guide the separation of distorted speech. The most obvious characteristic is that, in our model, the ITU-T P.563 standard, which is used to evaluate the objective speech quality, is introduced into the separation model. The overall model proposed by us is a multistage system, as shown in Fig.1.

In the first stage, an input mixture is analyzed by an auditory filterbank in consecutive time frames. This processing results in a decomposition of the input into a two-dimensional time–frequency map. Each unit of the map is called a T-F unit, corresponding to a certain filter at a certain time frame. Then the following features are extracted: autocorrelation of a filter response, autocorrelation of the envelope of a filter response, cross-channel correlation, a coarse dominant pitch within each time frame. These features are used in the following stages.

In the initial segregation stage, T-F units are merged into segments [5], [6], [9]. A segment is a larger component of an auditory scene than a T-F unit; it is composed of a spatially contiguous region of T-F units. This segment structure encodes the basic proximity principle in human ASA that applies to both frequency and time dimensions. Segments are then grouped into an initial foreground stream and a background stream based on coarse dominant pitch extracted in the previous stage; the two streams roughly correspond to target speech and intrusion, respectively. Due to the intrusion, the coarse dominant pitch may not be an accurate description of the target pitch. As a result, the foreground stream will usually miss some target speech and include some intrusion. In order to weaken the effect caused by the inaccurate dominant pitch, the objective quality assessment of speech is introduced into the CASA system. It is used as a criterion to judge the segments set in the foreground by the former division. After being processed by this step, the segments in the foreground are more like coming from the target source so we can employ them to track the more accurate dominant pitch from the foreground segments at the later stage.

In the third stage, the pitch of target speech is estimated from the initial foreground stream, and it is used to label units as speech dominant or interference dominant. In the final segregation stage, according to unit labels, segments formed in the initial segregation stage are regrouped into foreground and background stream. This stage corrects some errors of initial grouping due to the inaccuracy of the dominant pitch. In addition, some T-F units are merged into segments that correspond to unresolved harmonics of target speech, and these segments are added to the foreground stream. Here, we also introduced the objective quality assessment of speech. As in the initial segregation stage, we still use objective quality assessment of speech as the criterion to judge whether the relevant segments, which were composed by units that can not be confirmed as the foreground, can be grouped into background or not. Then the foreground stream expands to include neighboring T-F units labeled as speech dominant.

Finally, a speech waveform is resynthesized from the resulting foreground stream using a method described by

Weintraub [10]. Here, the foreground stream works as a binary mask, where 1 indicates T-F units within which target speech dominates and 0, otherwise. The mask is used to retain the acoustic energy from the mixture that corresponds to 1's in the mask and reject the mixture energy corresponding to 0; for more details of this stage, see [5], [9], [10].

Since there are many processing steps are similar to Hu and Wang model, here, we do not explain our model in detail at the corresponding part. To see more detailed depiction, please refer to the Section III to VI of paper [11]. In this article, we would only introduce how to combine the objective speech quality assessment into the Hu and Wang model.

For ease of comparison, let us reiterate the terms that have been introduced so far. A T-F unit is a very local time-frequency region corresponding to a certain filter at a certain time frame. We use to refer to the T-F unit corresponding to filter channel at time frame. A segment is a contiguous time-frequency region that corresponds to a component of a single sound source, and it is a set of connected T-F units. A stream is a group of segments that corresponds to an entire sound source. The target speech, or the target stream, is an utterance we aim to segregate from an acoustic mixture. What constitutes target speech is obviously task-dependent. In this study, target speech refers to an entirely voiced utterance in a sound mixture.

III. COMBINE CASA WITH OQAS

In this section, we will pay attention to state the method which helps to joint the CASA and OQAS in one system. As a whole, the combination of CASA and OQAS is realized by selecting the segments which are more similar to come from the same source in perception sense. Since in the initial segregation stage of CASA, the segments which have been judged and distributed to one stream only using a simple decision are not accurate and do not reflect the level of perception [11], we adopt the ITU-T P.563 algorithm to test the segments that having been distributed by the simply decision of CASA, this effectively improved the perceptual quality of the separated speech.

Concretely, in our model, there are two places where the objective quality assessment of speech is directly linked to the CASA system. One is in the initial segregation stage; the other is in the final segregation stage. The detailed method is cited as below.

A. OQAS in the Initial Segregation Stage

In the initial stage, after the decomposition and extraction processing has been finished, an initial grouping will be executed to give a primitive grouping result.

In Hu and Wang model, this grouping is done by comparing the dominant pitch contour with the periods computed from all the T-F units in a segment. For any segment, if more than half of its units at a certain frame agree with the dominant pitch, the segment would be said to agree with the dominant pitch at this frame. For the segments of target speech, if the dominant pitch at a certain frame is much closed to the true pitch of target speech, all of these segments tend to agree with the dominant pitch at this frame. Hence, segments can be grouped into two streams as follow. First, the longest segment is selected as a seed stream. Since the target speech in this study is all voiced, the longest segment extends through most of the frames of the entire utterance duration. At a certain frame, a segment is said to agree with the longest segment if both segments agree or both disagree with the dominant pitch. If a segment agrees with the longest segment for more than half of their overlapping frames, its T-F units within the duration of the longest segment is grouped into the seed stream. Otherwise, this segment is grouped into the competing stream. The longest segment is also used to determine which stream corresponds to target speech. If it agrees with the dominant pitch for more than half of its frames, it is likely to contain dominant target speech. In this case, the stream containing the longest segment is referred as the foreground stream, while the competing stream as the background stream. Otherwise, the names of the two streams are swapped.

From the description above, we can find that, the initial grouping stage is executed only by a simple decision. Although the grouping result of simple decision can be adjusted by the following stage through iterative estimation and linear interpolation so as to give an acceptable prediction of pitch contour, it yet does not satisfied the requirements of the segregation and would also set some segments which are dominated by the intrusions into the foreground. This will certainly affect the accuracy of the result of pitch tracking.

In order to get more reliable grouping result of foreground and background streams that correspond to the target source and interference sources respectively, a method consists of two steps processing is employed in our model.

In the first step, a simple decision as Hu and Wang used is adopted. Through employing a more conservative plausible pitch range $\theta_P^{'}$ [11], which is set to 0.90 (corresponding to the threshold θ_p, which is 0.95, in Hu and Wang model), the most improbable segments, in which the intrusion is dominant, would be filtrated. After this processing, a coarse classification is acquired, but this grouping must be treated further. This will be executed by the second step.

In the second step, the objective quality assessment is employed to give more accurate predictions in classifying the foreground and background. Considering that the former processing has selected many T-F units with some response energy and sufficiently high-cross-channel correlations to form the segments which are directly relevant to the perceptual quality of speech, and already divided them into foreground and background streams, we decided to use the ideal binary masking method to resynthesize a temporary speech, then adopt the P.563 algorithm to evaluate its quality. Fig. 2 gives a schematic diagram of this combination step. From Fig.2 we can see that, the whole evaluation process can be divided in two steps: First, we select the speech which is resynthesized by preserved all the segments in the foreground while masking all

the T-F units which are not in the foreground as the reference signal and evaluate its quality MOS_r by P.563 algorithm. Then, in order to judge the reliable segments in which the target speech is dominant, a comparison is introduced into the process. Through masking a certain segment while preserving all other segments in the foreground (here, the units not in the foreground are always masked in the resynthesis step), we can get a temporary speech. Using the P.563 algorithm to assess the resynthesized speech's quality, and comparing its MOS with the reference speech's MOS in turn, we can easily confirm the effects to the perceptual quality caused by the masked segment. These means that if there are N segments after primitive grouping in the foreground, well then, we should use P.563 algorithm N+1 times to assess the quality of speeches resyntheisized by masking corresponding segment of the foreground. If MOS_i of a speech, which was synthesized by masking the ith segment in foreground, is higher than the quality of the reference speech, MOS_r, then, it will mean that, if this segment was set to background, the synthesized speech's quality will be improved. So we can adjust this segment to the background, vice versa. The judgment above can be represented as the equation below:

$$Mask(i)=\begin{cases}1, if\ Mos_i - Mos_r > 0\\ 0, if\ Mos_i - Mos_r \le 0\end{cases} i=1,2,\cdots N \qquad (1)$$

Here, *Mask (i)* is the masking value of the ith segment in foreground, 1 represents foreground, while 0 background.

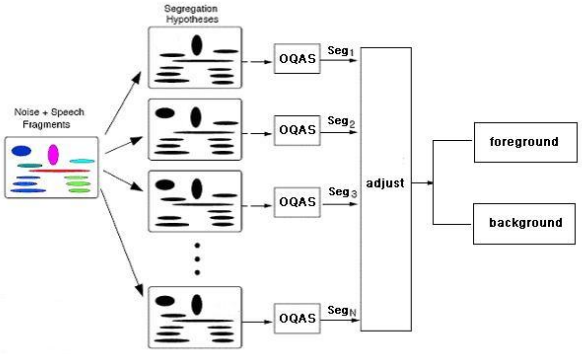

Fig. 2. Sketch map of the combination in initial segregation stage

In real practice, because of the accuracy of the quality assessment algorithm and the complexity of the sources mixture, the adjustment of the segments needs to be more conservative. Here a threshold θ_A is introduced into the judgment, and the number of it is set to be 0.02. Then, Eq.(1) could be modified as:

$$Mask(i)=\begin{cases}1, if\ Mos_i - Mos_r > \theta_A\\ 0, if\ Mos_i - Mos_r \le \theta_A\end{cases} i=1,2,\cdots N \qquad (2)$$

This conservative processing can alleviate the error adjustment caused by the objective quality assessment and yet avoid deleting too many useful segments from the foreground. After this adjustment, the segments which are still kept in the foreground would be more like coming from the target source. Followed by the pitch tracking and unit label steps described in Hu and Wang model, the dominant pitch estimated from the segments staying in the foreground would be more closed to the real pitch of the target speech, and it will also help the further grouping of the foreground and background streams.

B. OQAS in the Final Segregation Stage

In the final segregation stage, since the spectra of target speech and intrusion often overlap and, as a result, some segments generated in the former segmentation still contain units where target dominates as well as those where intrusion dominates. Given unit labels generated in pitch label, a segment in foreground can be further divided into smaller ones so that all the units in a segment have the same label. Then the segments in the foreground are adjusted as follows:

1) Segments with the target label are retained in foreground if they are no shorter than 50 ms;

2) Segments with the intrusion label and no shorter than 50ms are evaluated by OQAS and those caused the speech quality decreased segments are sent to background;

3) Remaining segments are removed from foreground, and they become undecided.

A point need to be emphasized is that, the combination way of OQAS and CASA in Step 2) is similar to that in the initial segregation stage. Through masking every segment no shorter than 50ms with the intrusion label in foreground, and evaluating the corresponding resynthesized speech's perceptual quality, these segments are added into foreground or background to form new streams, respectively. But there is a little difference between them. In this step, the judgment threshold is 0, while in the initial segregation stage, it is 0.02.

After the processing above, background expands iteratively to include undecided segments in its neighborhood. Then, all the remaining undecided segments are added back to foreground.

Finally, individual units that do not belong to either stream are grouped into the foreground stream iteratively if they are labeled as target speech and in the neighborhood of the foreground stream. The result of this is the final segregated stream of target speech. The remaining units are added to the background stream.

The segregated stream from the speech and intrusion mixture contains most of the units where target speech is dominant. In addition, only a little number of units where intrusion is dominant are grouped into foreground incorrectly. Fig. 3 illustrates the segregation result in waveform format for the speech and cocktail party mixture. The clean speech is shown in Fig. 3(a), the mixture in Fig. 3(b), and the segregated speech in Fig. 3(c). To facilitate comparison between these waveforms, an all-one mask is used to synthesize the waveforms in Fig. 3(a) and 3(b). One can easily see that the segregated speech waveform is much more similar to the clean speech than the mixture waveform.

IV. Evaluation and Comparison

In this section, we will elaborate the evaluation of our model in SNR and MOS on a standard corpus and compare its performance with other models.

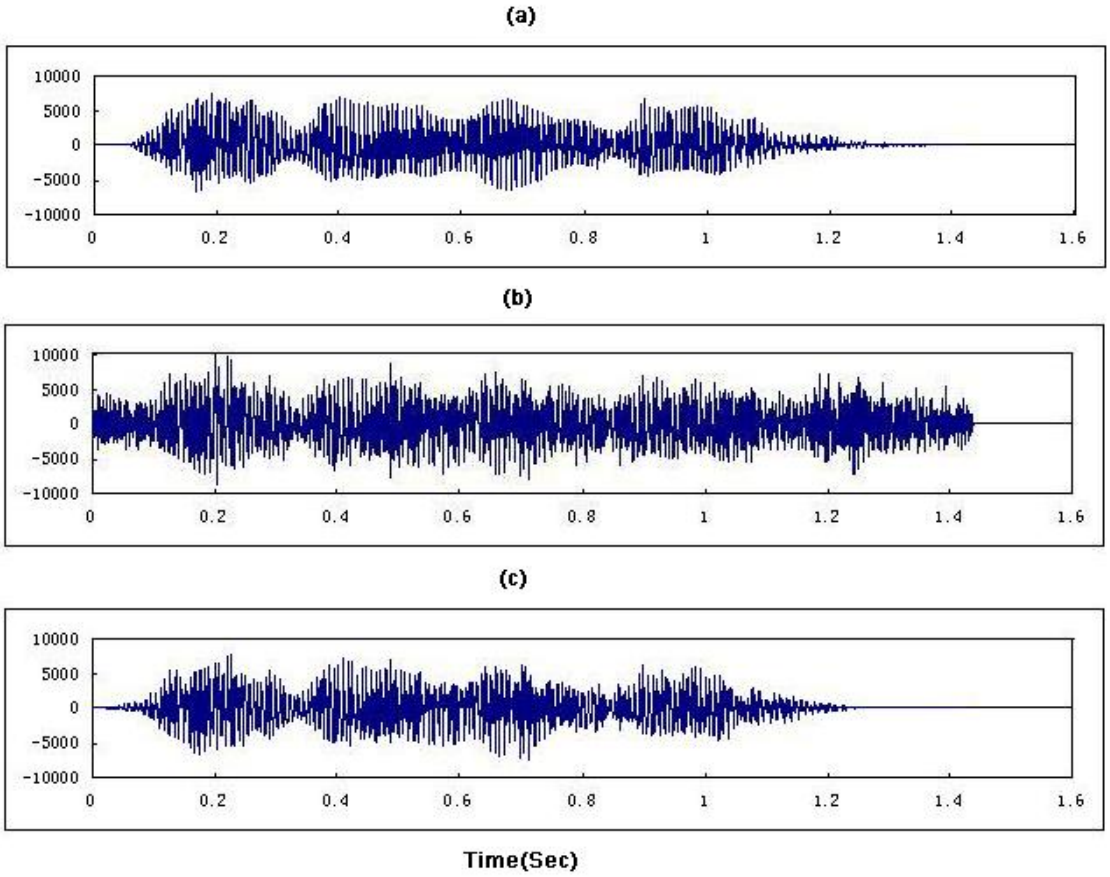

Fig. 3. Waveform results of "Why were you weary?"
(a) Clean Speech. (b) Mixture of the speech and cocktail-party noise. (c) Speech segregated from the mixture

A. Evaluation

Our model is evaluated with a corpus of 100 mixtures composed of 10 voiced utterances mixed with 10 intrusions collected by Cooke [9], which has been used to test CASA systems [5], [6], [7], [9], [21] and, hence, facilitates our comparison. The intrusions have a considerable variety. Specifically, the 10 intrusions are: N0, 1-kHz pure tone; N1, white noise; N2, noise bursts; N3, "cocktail party" noise; N4, rock music; N5, siren; N6, trill telephone; N7, female speech; N8, male speech; and N9, female speech [22]. We use both SNR and MOS as the criterion to quantitatively assess the performance of our new separation system.

The original sampling frequency of the corpus is 16 kHz. Because the OQAS algorithm we employed here is only used to evaluate narrow-band telephone speech, we downsampled the corpus created by Cooke to 8 kHz. Besides this, there are also some requirements need to be satisfied before using the ITU-T P.563 algorithm to test speech quality. These requirements include that the minimum active speech in test speech is 3 second, the maximum signal length is 20 second, the minimum speech activity ratio is 25%, the maximum speech activity ratio is 75%, etc. Since the duration of the speech in Cooke's corpus is mainly about 1.5 second, it is easy to conclude that they could not fulfil the requirements of the P.563 algorithm. In order to successfully apply the P.563 algorithm, we repeat the test speech three times with an interval of 0.5 second between each time. This processing not only makes the test speech fulfil the requirements of OQAS, but also almost have nothing effect on the assessment result. A key point need to be emphasized here is that the P.563 algorithm could not give correct result in some intrusion conditions, such as the music, male speech, female speech, etc. Although it is a limitation we can not cancel, we still adopt the P.563 algorithm in all the test intrusion conditions because of the characteristic that the test speech in our applications is often the signals which have been processed by masking a large part of the intrusion signals.

In order to measure SNR before and after segregation, we use target speech before mixing as signal. To compensate for amplification and distortion effects introduced in the resynthesis process, we use resynthesized target speech with an all-one mask as signal to compute SNR for evaluation cases that involve masks. Table I gives a variety of SNR results, including those of our model and original mixtures. Each value in the table represents the average SNR for one intrusion mixed with 10 target utterances. A further average across all intrusions is shown in the last column of the table. As can be seen in the table, our system improves the SNR for every intrusion, producing a gain of 9.75 dB over the original mixtures. Large SNR improvements are obtained for intrusions whose spectra do not significantly overlap with those of target

TABLE I
SNR RESULT

SNR	N0	N1	N2	N3	N4	N5	N6	N7	N8	N9	Ave
Mixture	-7.380	-8.269	5.474	0.803	0.679	-9.999	-1.609	3.842	9.526	2.749	**-0.418**
SS	7.568	-3.879	6.207	2.651	3.154	-9.609	0.950	4.709	9.968	3.513	**2.523**
Proposed	***11.129***	***3.507***	***14.411***	***5.218***	***6.669***	***12.933***	***14.662***	***9.391***	***11.506***	***3.964***	***9.339***
Hu Wang	10.330	3.346	14.251	5.094	1.095	12.869	15.213	9.040	12.556	5.100	**8.889**
True Pitch	13.044	4.239	14.286	6.147	9.585	12.822	14.855	11.019	13.914	7.226	**10.714**
Ideal Mask	20.001	5.963	18.438	8.122	11.598	17.283	18.992	13.948	17.484	11.176	**14.301**

TABLE II
MOS RESULT OF P.862

MOS	N0	N1	N2	N3	N4	N5	N6	N7	N8	N9	Ave
Mixture	2.239	1.288	1.548	1.654	1.507	0.307	2.056	2.010	2.329	2.263	**1.720**
SS	2.487	1.209	1.046	1.531	1.039	-0.121	2.017	1.386	1.954	1.813	**1.347**
Proposed	***2.673***	***0.699***	***2.247***	***1.087***	***1.162***	***2.255***	***2.372***	***1.761***	***1.886***	***1.306***	***1.745***
Hu Wang	2.630	0.583	2.266	0.922	0.910	2.271	2.459	1.652	2.000	1.673	**1.737**
True Pitch	2.666	0.840	2.241	1.147	1.527	2.277	2.480	1.843	2.139	2.026	**1.919**
Ideal Mask	3.203	1.506	2.820	1.787	2.012	2.691	3.103	2.522	2.824	2.857	**2.533**

TABLE III
MOS RESULT OF P.563

MOS	N0	N1	N2	N3	N4	N5	N6	N7	N8	N9	Ave
Mixture	1.765	1.000	4.686	1.670	1.967	3.357	2.149	3.835	3.769	3.344	**2.754**
SS	3.190	1.000	5.000	2.304	3.298	3.272	2.456	4.321	4.344	3.951	**3.314**
Proposed	***4.214***	***3.720***	***4.269***	***3.838***	***3.783***	***3.748***	***3.976***	***3.628***	***3.805***	***3.605***	***3.858***
Hu Wang	4.178	3.076	4.342	3.582	2.804	3.722	3.783	3.845	3.683	4.194	**3.721**
True Pitch	4.692	3.540	4.228	3.686	4.518	3.909	4.058	4.009	3.898	4.155	**4.069**
Ideal Mask	4.757	4.311	4.586	3.392	3.985	4.235	3.848	4.263	4.355	4.390	**4.210**

utterances (e.g., N0 and N5), whereas improvements are modest for intrusions with significant overlap (e.g., N3, N8, and N9).

Of course, SNR does not indicate the *intelligibility* of the resynthesized speech signal. For example, the model could retrieve a small proportion of the speech energy and totally reject the noise; this would give a very high SNR, but the resynthesized speech would be unintelligible. Accordingly, we complement the SNR metric with a measure of the MOS before and after segregation. This measurement is accomplished by two means. The first way we used is the ITU-T P.862 algorithm, which is an intrusive objective speech quality assessment algorithm. In this way, we treated the target speech before mixing as the reference input and the separated speech as test input. Another way we adopted is the ITU-T P.563 algorithm. In this situation, we straightforwardly treat the synthesis signal as test speech and get its MOS from the output of the ITU-T P.563.

Table II and III give a variety of MOS results acquired by P.862 and P.563 algorithm, respectively, including those of our model and original mixtures. Since our model is based on the P.563 algorithm, it is unfair to use P.563 algorithm to calculate the MOS of test speech, because it's no doubt that the MOS of the speech processed by our model would inevitably be higher than other models. Moreover, the output of P.563 would be incorrect in some intrusion conditions, especially to the mixture in music noises and female speech intrusion. Considering of these shortcomings described above, we only list the MOS calculated by P.563 as an additive reference. The main performance about MOS is referred to the MOS calculated by P.862.

In Table II and III, each value in the tables also represents the average MOS for one intrusion mixed with 10 target utterances. A further average across all intrusions is shown in the last column of the table, too. As can be seen in the Table II, our system improves the MOS for half kinds of intrusions. Large MOS improvements are obtained for intrusions whose spectra do not significantly overlap with those of target utterances (e.g.,

N0 and N5), whereas a bit of decreases are also appeared for intrusions with significant overlap (e.g., N3 and N9). Although the reason why the MOSs for different intrusions do not improved conformably is very complex, we still can say that it is mainly because that in the condition that intrusions has significant overlap with target speech, the separated speech has lost too much information due to the masking processing, which can not be recovered by synthesis.

B. Comparison with Other Models

We have elaborated the evaluation of our model and showed that the proposed model can effectively improve the SNRs and partial MOSs of separated speeches. But whether the model we proposed is better than others? How well on earth does it is? Both of these are need to be proved by comparing with other models.

In order to answer above two questions, we compare the SNRs and MOSs of the original mixtures and the speeches separated by our model together with the results of spectrum subtraction, true pitch model, Hu-Wang model, and ideal binary mask. (To get more definite meaning about these models described above, please refer to [11]). The corresponding SNRs and MOSs of separated speeches offered by these models are listed in Table I, II, and III, respectively. Note that for an original mixture and an output from spectral subtraction, an all-one mask is used against the corresponding ideal binary mask to calculate SNR.

First, we compare spectral subtraction [23], [24], which is a standard method for speech enhancement, with our model. The spectral subtraction method is applied as follows. For each intrusion, we find its duration and obtain its average power spectrum within the duration. This average is used as the estimate of the intrusion. But to N2 intrusion, because it contains a sequence of short noise bursts, spectral subtraction is applied within each burst. Table I - III show a comparison with the spectral subtraction method. With this comparison, we can easily conclude that the spectral subtraction method performs significantly worse than our system in SNR. This is because of its well known deficiency in dealing with nonstationary interference. In perceptual quality aspect, the average MOS of our model is much better than spectral subtraction too. But the MOS on different intrusion condition are not always better than spectral subtraction, especially in complex intrusion. This could also be attributed to the missing of information caused by masking processing.

We also compare our model with Hu and Wang model on SNR and MOS. The Hu and Wang model is a representative CASA system [11] in recent years. Its main processing stages are similar to our model except that it dose not combine the OQAS algorithm. The SNR and MOS results of Hu and Wang model are listed in Table I – III, too. From these tables, we can find that SNRs of the speeches separated by our system are better than Hu and Wang model not only in average level, but also in almost every intrusion. The MOS of our model is also better than Hu and Wang model except in the intrusions of male and female speeches. This is caused by the inaccurate assess of speech quality in the separation stage. We have emphasized that, the P.563 algorithm could not be used in the speech intrusion condition. Although the binary mask reduces many intrusion energy of the test speech, it does not absolutely break the limitations. To conquer this difficulty, we should discover more appropriate OQAS algorithm to eliminate the gap.

Considering that the pitch estimation or pitch-based grouping error in practical situation is inevitable, and it may result in segregation error, to examine more closely the type of error, we employ the use of true pitch information for speech segregation. True pitch is obtained from premixing target speech and further verified manually to ensure high quality. The fifth row of Table I, II, and III give the SNR and MOS results for our system using true pitch instead of estimated pitch. With true pitch, the system performs only slightly better. This suggests that estimated pitch of our system is quite accurate and the performance of our model is very good.

Given the objective of identifying T-F regions that target is dominant, we use an ideal binary mask as the ground truth of target stream. An ideal binary mask is constructed as follows: a T-F unit in the mask is assigned 1 if the target energy in the unit is greater than the intrusion energy and 0 otherwise. With the availability of target and intrusion before mixing, as is the case for our evaluation corpus, ideal binary masks can be readily constructed. We call such a mask “ideal” because it represents our computational objective and it is an *a priori* mask constructed using premixing target and intrusion. The use of ideal masks is supported by the auditory masking phenomenon: within a critical band a weaker signal is masked by a stronger one [25]. The SNR and MOS results from ideal binary masks are shown in Table I - III, and they are uniformly better than all the models mentioned before. Compared with our proposed model, the average SNR improvement for the entire corpus is about 5.0 dB and the average MOS improvement is about 0.8. This gives an indication on how much our model could be further improved in terms of conventional SNR and MOS.

V. Conclusions

Our system segregates voiced speech based on the primitive CASA system combined with OQAS algorithm. The CASA system analyzes temporal information in the input, the temporal fine structure of a resolved harmonic and the temporal envelope of an unresolved harmonic. There is evidence proving that the auditory system uses the temporal patterns of neural spikes to code the input sound [7]. Models based on temporal coding of the input, such as correlogram, have been employed to model auditory perception, especially pitch perception, and have successfully explained many observed perceptual phenomena [26], [27], [28]. The OQAS algorithm is used to classify foreground and background streams. The segments, which are still kept in foreground after OQAS processing, would be used to estimate the dominant pitch. This effectively increases the accuracy of pitch estimation, and surly improves the performance of the separation system.

Like previous CASA systems, our system exploits the

grouping cues of harmonicity and temporal continuity to segregate voiced speech [5], [6], [9], [10]. However, our system is substantially different from previous studies in the following respects.

First, by means of combining CASA with OQAS, we introduced the speech perceptual quality knowledge into separation, and construct a direct link between separated speech and its perceptual quality. It is the first attempt to introduce knowledge about speech quality assessment into CASA systems. From the results of our system, we can conclude that the link between separated speech and its perceptual quality would be valuable to solve speech separation problem.

Second, we find an appropriate representation of speech at middle level. Through decomposition and extraction, speech signal was divided into many T-F units, these units then form many segments. Based on the segments, we resynthesize a temporary speech with one segment masked while others preserved, and send the temporary resynthesised speech to OQAS algorithm to judge its quality. Just because of the representation described above, the combination of CASA and OQAS comes true.

From the comparison with other separation or enhancement system, we have drawn a conclusion that our method is effective in processing the monaural speech separation problem. But, there are also some deficiencies in our research. We must pay more attention to them; analyze the reasons which caused these problems carefully and employ appropriate method to solve them thoroughly.

In our research, the performance of the proposed model depends greatly on the accuracy of an estimated target pitch contour. To get more accurate pitch contour, the segments in foreground which was used to estimate the pitch contour is very important. Since the classification of the foreground and the background in the initial segregation stage is mainly based on the OQAS algorithm, how to improve the accuracy of the OQAS algorithm seems to be the key problem needed to be solved urgently. In additional, because our research on combining CASA with OQAS is primary and not mature, we only find a method to add the whole OQAS algorithm into CASA systems to guide the classification of foreground and background. This combination method does not make fully use of the knowledge of speech quality, therefore, we need to study it further so as to seek the best way in all probable combination methods. This perhaps needs to break OQAS algorithm into smaller units, and only use some of its basic principles to the separation process. Finding an optimal combination method of CASA and OQAS will be our objective in the future.

The proposed system considers the pitch contour of a target source only. However, it is possible to track the pitch contour of intrusion if it has a harmonic structure. With two pitch contours, one could label a T-F unit more accurately by comparing whether its periodicity is more consistent with one or the other. Such a method is expected to lead to better performance for the two-speaker situation [29], e.g., N7, N8, and N9. As indicated in Table I, II, and III, the performance of our system for this kind of intrusions is relatively limited. Our model performs grouping based only on pitch. As a result, it is limited to segregation of only voiced speech. In our view, unvoiced speech poses the biggest challenge for monaural speech segregation. Other grouping cues, such as onset, offset, and timbre, have been demonstrated to be effective for human ASA [4], [30], and may play a role in grouping unvoiced speech. Also, it appears that one must consider acoustic and phonetic characteristics of individual unvoiced consonants. We plan to investigate these issues in future works.

ACKNOWLEDGMENT

This work was supported by the National Grand Fundamental Research 973 Program of China (Grant No. 2004CB318105). The authors wish to thank Guoning Hu and Doh-Suk Kim for fruitful discussions and help; and the two anonymous reviewers for their great help in improving the structure of this article.

REFERENCES

[1] C. Cherry, “Some experiments in the recognition of speech with one and two ears,” *J. Acoust. Soc. Amer*., vol. 25, pp. 975-981, 1953

[2] A. K. Barros, T. Rutkowski, F. Itakura, and N. Ohnishi, “Estimation of speech embedded in a reverberant and noisy environment by independent component analysis and wavelets,” *IEEE Trans. Neural Networks*, vol. 13, pp. 888–893, July 2002.

[3] H. Krim and M. Viberg, “Two decades of array signal processing research: the parametric approach,” *IEEE Signal Processing Mag.*, vol. 13, pp. 67–94, July 1996.

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

[5] G. J. Brown and M. P. Cooke, “Computational auditory scene analysis,” *Comput. Speech Language*, vol. 8, pp. 297–336, 1994.

[6] M. P. Cooke, *Modeling Auditory Processing and Organization*. Cambridge, U.K.: Cambridge Univ. Press, 1993.

[7] D. P.W. Ellis, “Prediction-Driven Computational Auditory Scene Analysis,” Ph.D. dissertation, Dept. Elect. Eng. Comput. Sci., MIT, Cambridge, MA, 1996.

[8] D. F. Rosenthal and H. G. Okuno, *Computational Auditory Scene Analysis*. Mahwah, NJ: Lawrence Erlbaum, 1998.

[9] D. L. Wang and G. J. Brown, “Separation of speech from interfering sounds based on oscillatory correlation,” *IEEE Trans. Neural Networks*, vol. 10, pp. 684–697, May 1999.

[10] M. Weintraub, “A Theory and Computational Model of Auditory Monaural Sound Separation,” Ph.D. dissertation, Dept. Elect. Eng., Stanford Univ., Stanford, CA, 1985.

[11] G. N. Hu and D. L. Wang, “Monaural speech segregation based on pitch tracking and amplitude modulation,” *IEEE Trans. Neural Net*., vol. 15, pp. 1135-1150, 2004.

[12] N. Roman, D. L. Wang and G. J. Brown, “Speech segregation based on sound localization,” J. Acoust. Soc. Am., vol. 114, pp. 2236-2252, 2003.

[13] D. Godsmark, and G. J. Brown, “A blackboard architecture for computational auditory scene analysis, ” Speech Communication, vol.27, pp. 351-366. 1999

[14] ITU-T Rec. P.830, “Subjective performance assessment of telephone-band and wideband digital codecs,” International Telecommunication Union, Geneva, Switzerland, Feb. 1996.

[15] P. Gray, M. P. Hollier, and R. E. Massara, “Nonintrusive speech-quality assessment using vocal-tract models,” in *IEE Proc. - Vision, Image and Signal Processing*, vol. 147, no. 6, pp. 493–501, Dec. 2000.

[16] C. Jin and R. Kubichek, “Vector quantization techniques for output-based objective speech quality,” in *Proc. of the Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 491–494, May 1996.

[17] D. S. Kim, “ANIQUE: An auditory model for single-ended speech quality estimation,” *IEEE Trans. Speech and Audio Processing*, in pressing.

[18] ITU-T P.563, "Single ended method for objective speech quality assessment in narrow-band telephony applications," International Telecommunication Union, Geneva, Switzerland, May 2004.
[19] Psytechnics Limited, "NiQA - Product Description," Tech. Rep., January 2003. [Online]. Available: http://www.psytechnics.com/pages/products/niqa.php
[20] Swiss Qual Inc., "NiNA - SwissQual's non-intrusive algorithm for estimating the subjective quality of live speech," Tech. Rep., June 2001. [Online]. Available: http://www.swissqual.com/HTML/ninapage.htm
[21] L. A. Drake, "Sound source separation via computational auditory scene analysis (CASA)-enhanced beamforming," Ph.D. dissertation, Dept. Elect. Comput. Eng., Northwestern Univ., Evanston, IL, 2001.
[22] D. F. Rosenthal and H. G. Okuno, *Computational Auditory Scene Analysis*. Mahwah, NJ: Lawrence Erlbaum, 1998. R.
[23] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acous. Speech Signal Proc*., vol.27 (2), pp.113-120, 1979.
[24] Martin. "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech and Audio Proc*., vol. 9, pp. 504-512, 2001.
[25] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 4th ed. San Diego, CA: Academic, 1997.
[26] P. Cariani, "Temporal coding of periodicity pitch in the auditory system: an overview," *Neural Plasticity*, vol. 6, pp. 147–172, 1999.
[27] R. Meddis and L. O'Mard, "A unitary model of pitch perception," *J. Acoust. Soc. Amer*., vol. 102, pp. 1811–1820, 1997.
[28] M. Slaney and R. F. Lyon, "On the importance of time—a temporal representation of sound," in *Visual Representations of Speech Signals*, M. P. Cooke, S. Beet, and M. Crawford, Eds. New York:Wiley, 1993, pp.95–116.
[29] M. Wu, D. L. Wang, and G. J. Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Trans. Speech Audio Processing*, vol.11, pp. 229–241, May 2003.
[30] G. Hu and D. L. Wang, "Separation of fricatives and affricates," Proc. ICASSP2005, vol. 1, pp. 1101-1104, 2005.

Peng Li received the B.S. and M.S. degrees in automation from Tianjin University, Tianjin, China, in 2000 and 2003, respectively. He is currently working toward the Ph.D. degree in pattern recognition at the institute of automation, Chinese academy of sciences, Beijing, China.
His research interests include speech segregation, computational auditory scene analysis, speech enhancement, noise reduction, speech recognition, etc.

Yong Guan received the B.S. degrees in automation from Tsinghua University, Beijing, China, in 2002. He is currently working toward the Ph.D. degree in pattern recognition at the institute of automation, Chinese academy of sciences, Beijing, China.
His research interests include speech segregation, computational auditory scene analysis, speaker identification, etc.

Bo Xu received the B.S. degrees in electrical engineering from Zhejiang University, Hangzhou, China, in 1988, and M.S. and Ph. D. degrees in pattern recognition from institute of automation, Chinese academy of sciences, Beijing, China, in 1992 and 1997, respectively. He is currently working as a professor at high-tech innovation center, institute of automation, Chinese academy of sciences, Beijing, China.
His research interests include speech recognition, speech synthesis, speaker recognition, speaker identification, machine translation, computational auditory scene analysis, audio retrieval, digital content management, etc.

Wenju Liu received the B.S., M.S. degrees in mathematics from Beijing University and Beijing University of Post and Telecommunication, and Ph. D. degree in computer applications from Tsinghua University , Beijing, China, in 1983, 1989 and 1993, respectively. He is currently working as an associate professor at the national key laboratory of pattern recognition, institute of automation, Chinese academy of sciences, Beijing, China.
His research interests include speech recognition, speech synthesis, speaker recognition, voice conversion, computational auditory scene analysis, speech enhancement, noise reduction, etc.