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 extensively, 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 processes using primitive data-driven methods, the knowledge on speech quality still has not been combined with it. This makes the performance evaluation of CASA mainly focused on the signal-to-noise ratio (SNR) improvement. Actually, the quality of the separated speech is not directly related to its SNR. In order to solve this problem, we propose a new method which combines 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. With 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*—Computational auditory scene analysis (CASA), grouping, monaural speech separation, objective quality assessment of speech (OQAS), segmentation.**

I. Introduction

IN A NATURAL world, a speech signal is frequently accompanied by other sound sources upon reaching auditory systems, yet listeners are capable of holding conversations in a wide range of conditions. This phenomenon is well known as the "cocktail party" effect [1]. It is valuable to make a computer have the ability of a human being to segregate the object source from other interfering sources. An effective separation system can greatly facilitate many applications, including automatic speech recognition (ASR), speaker identification, audio retrieval, digital content management, etc. Therefore, the research on speech separation gradually catches the researchers' attentions, and it has become an increasingly popular topic in the field of signal processing.

Manuscript received January 31, 2006; revised July 18, 2006. This work was supported by the National Grand Fundamental Research 973 Program of China under Grant 2004CB318105. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Oded Ghitza.

P. Li and B. Xu are with the High-Tech Innovation Center, Institute of Automation, Chinese Academy of Sciences, Beijing 100080, China (e-mail: pengli@hitic.ia.ac.cn; xubo@hitic.ia.ac.cn).

Y. Guan and W. Liu are with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100080, China (e-mail: yguan@nlpr.ia.ac.cn; lwj@nlpr.ia.ac.cn).

Digital Object Identifier 10.1109/TASL.2006.883258

There are some general methods which are used to segregate speech, such as blind source separation [2] and spatial filtering [3]. These methods need multiple sensors. However, in many applications, e.g., telecommunication and audio retrieval, there is only one sensor, and a monaural solution is expected. Since in monaural separation cases only one sensor signal could be used, it is much harder and still an open problem for researchers to explore.

Although monaural speech separation is still a challenging problem, the human auditory system shows a remarkable capacity for monaural speech segregation, which spurs the researchers to study human auditory perception further. In 1990, Bregman first proposed the concept of auditory scene analysis (ASA) [4]. In the book, he argues that according to ASA principles the auditory system could segregate acoustic signal into streams, which correspond to different sources. His study on ASA offers a new way to deal with the monaural speech separation. It also inspired considerable work on computational auditory scene analysis (CASA). Many CASA systems have been proposed for speech separation according to the known principles of ASA [5]–[11]. These systems not only are able to realize the objective of speech segregation, but also need no strong assumption on the acoustic properties of interferences. As it is well known, a CASA system generally consists of two main stages: segmentation (analysis) and grouping (synthesis) [4]. At the segmentation stage, the acoustic input is decomposed into sensory segments, each of which should be originated from a single source. At the grouping stage, those segments probably originated 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 kind of CASA system usually extracts cues, such as pitch, onset and offset, AM rate, etc., from the input data and then uses them to separate the target speech from the mixture. In the last decade, the research focusing on CASA changed from the primitive data-driven method to the knowledge-based schema-driven method. More and more knowledge in higher level, such as the acoustic model used in ASR, source characteristics, source location, and some others, is introduced into the primitive CASA systems to guide the separation [7], [12], [13]. Although recent great achievements are gained on the knowledge-based CASA research accompanied with many new types of knowledge being introduced into the CASA systems, the knowledge on speech perceptual quality has not been combined with them yet.

On the other hand, in most CASA systems, the performance evaluation of the CASA system is always based on

the signal-to-noise ratio (SNR). Although the SNR of speech after separation is improved and the noise is surely reduced by the CASA system, it does not mean that the speech quality in perception is also promoted. It is not absolutely right that the higher the SNR of a signal is, the better the perceptual quality of the signal is. Thus, in order to improve both SNR and perceptual quality of the separated speech, we attempt to seek an effective way to combine the perceptual quality measurement with CASA systems.

As it is well known, the speech quality is a subjective opinion, and it is based on the users' reaction to the speech signals they actually hear. Subjective assessment methods make use of a listener panel to measure speech quality on an integer scale from 1 to 5, where 1 corresponds to unsatisfactory speech quality and 5 corresponds to excellent speech quality. The average of the listeners' scores is commonly referred to as mean opinion score (MOS) [14]. It is the most reliable method, but it is very expensive both in time and cost, and as a result, it is unsuitable for frequent or rapid applications. However, these shortcomings can be overcome by using objective measurement methods, which replace the listener panel with a computational algorithm. The objective methods aim to reflect subjective ratings on speech signals in a reliable manner.

Objective quality assessment methods can be classified as intrusive or nonintrusive. Intrusive measurement depends on some form of distance metrics between the reference (clean) and test (degraded) speech signals to predict the subjective MOS. Nonintrusive measurement depends only on the test speech signal and is more challenging to the objective speech quality estimation. Nonintrusive models have been proposed in [15]–[17], but only recently has the ITU-T released P.563 as its nonintrusive objective quality measurement standard algorithm [18]. In speech separation applications, an intrusive approach may be inapplicable because the reference speech signals are unavailable, so the nonintrusive method is recommended. For this reason, the P.563 algorithm is selected to implement our speech quality assessment.

Since the objective quality assessment algorithm has been selected, the rest is just to seek an appropriate way to integrate it with a segregation process. Considering the characteristics of the primitive CASA system, especially the Hu and Wang's model [11], we construct the link between the speech quality and the CASA processing. On one hand, we use the speech quality assessment to evaluate the segments formed by the segmentation of CASA so that we can select better segments which were not affected greatly by interference sources 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 cannot be assigned to the foreground stream in the former step, judge the accuracy of the former classification, and then adjust the corresponding segments back to the foreground to enhance the final segregation performance.

The organization of this paper is as follows. Section II first describes the construction of our model by analyzing the appropriate selection of objective quality assessment of speech (OQAS) algorithm and CASA system; then, an overview of the new model is given, and each component of the model is explained concisely. The key point of the proposed model, the combination of CASA with OQAS, is described in Section III. In Section IV, the proposed system is systematically evaluated and compared with other systems for speech segregation or enhancement. Finally, a further discussion is given in Section V.

II. System Construction and Overview

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

A. Selection of OQAS Algorithm

As discussed in Section I, in speech separation applications, the nonintrusive objective quality assessment method should be selected for the reason of the absence of input reference signals. So, in this paper, the ITU-T P.563 algorithm is selected.

The main reason we select the P.563 algorithm lies in that it is the only algorithm recommended by ITU as the standard for objective quality assessment in narrow-band telephony applications. The P.563 algorithm is derived by combining Psytechnics' NiQA algorithm [19], SwissQual's NiNA [20], and Opticom's P3SQM. Its signal parameterization can be divided into three independent functional blocks that correspond to the main classes of distortion: vocal tract analysis and unnaturalness speech, strong additional noise, and speech interruptions, muting, and time clippings. A total of 51 characteristic signal parameters are calculated. Based on a restricted set of eight key parameters, a dominant distortion class is selected. The key parameters and the selected distortion class are used to adjust the speech quality model. Furthermore, for each distortion class, a linear combination of parameters is used to generate an intermediate quality rating, which is combined to calculate the (raw) objective quality score together with other additional signal features [18].

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

B. Selection of CASA System

The CASA system employed in the proposed model is based on the Hu and Wang's model [11]. This model is a typical primitive CASA model. It segregates resolved and unresolved harmonics in different ways. For resolved harmonics, based on temporal continuity and cross-channel correlation, the system generates segments 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 the AM rates. The key point of the Hu and Wang's model is a pitch contour. To get an accurate pitch contour, a coarse one is first estimated with speech segregated according to the dominant pitch. Then it is revised according to psychoacoustic constraints. This makes the separation performance of the Hu and Wang's model almost the

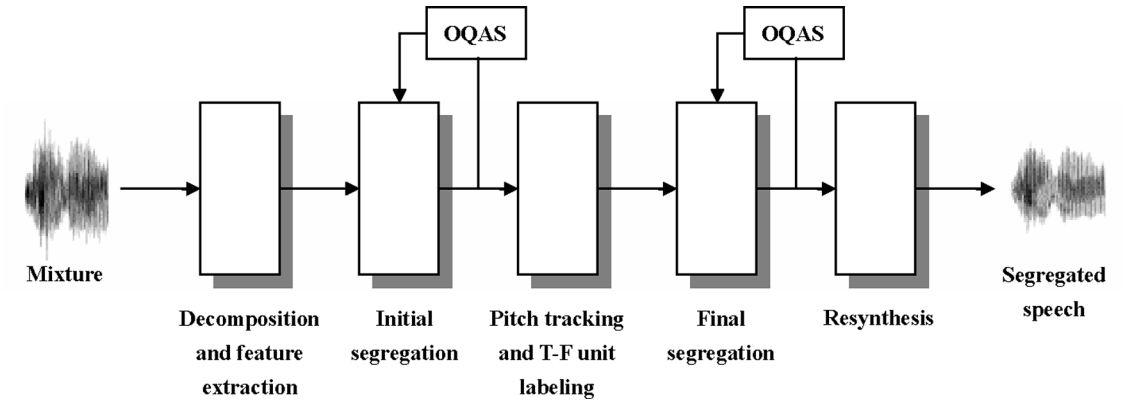


Fig. 1. Schematic diagram of the proposed multistage system.

best in the primitive CASA systems and even better than many other knowledge-based CASA systems in dealing with voiced speeches.

Another reason the Hu and Wang's model is selected lies in that it employs the notion of time-frequency mask. In their model, binary masks are used as output represents in the CASA literature. A separated speech is resynthesized by retaining the units where the target speech is dominant and rejecting the units where the interference is dominant. The idea of binary mask is rooted in ideal binary mask, which is supported by the auditory masking phenomenon: within a critical band a weaker signal is masked by a stronger one [21]. The ideal binary mask is very effective for human speech intelligibility. It is well-defined no matter how many intrusions exist or how many targets need to be segregated and, thus, it provides an excellent front end for robust automatic speech recognition [22]. Of course, it also provides a convenient way to combine the CASA system with the OQAS algorithm.

C. Overview of the Combination System

The objective of the proposed model is to guide the separation of distorted speech with OQAS. That is, the ITU-T P.563 standard, which is used to evaluate the objective speech quality, is introduced into the separation model. The proposed model is a multistage system, as shown in Fig. 1.

In the first stage, an auditory filterbank is used to analyze the input mixture in consecutive time frames. By means of this processing, the input signal is decomposed into a two-dimensional time-frequency map. Each unit of this map is termed as a T-F unit corresponding to a certain filter at a certain time frame. Then, the features, e.g., the autocorrelation of a filter response, the autocorrelation of the envelope of a filter response, the cross-channel correlation, and a coarse dominant pitch within each time frame, are extracted. These features will be used in the following stages.

In the initial segregation stage, some segments are generated. A segment is defined as a larger component of an auditory scene than a T-F unit and it consists of a spatially contiguous region of T-F units [5], [6], [9]. This segment structure encodes the basic proximity principle in human ASA that applies to both the frequency and time domains. Then, based on a coarse dominant pitch extracted in the previous stage, these segments are grouped into two streams: the initial foreground stream and the background stream, corresponding to target speech and intrusion, respectively. Due to the intrusion, the coarse dominant pitch may not be accurate enough, and as a result, the foreground stream may miss some target speech and include some intrusion. In order to weaken the effect caused by the inaccurate dominant pitch, OQAS is introduced into this stage. The OQAS output is used as a rule to judge the assignment accuracy of the segments set in the foreground so as to make a more accurate classification. Then, the segments remained in the foreground are more reliably originated from the target source and they will be used to track a more accurate dominant pitch later.

In the third stage, the pitch of target speech, which is used to label units as speech dominant or interference dominant, is estimated from the initial foreground stream.

In the final segregation stage, segments generated in the initial segregation stage are regrouped into the foreground and the background stream according to the labels of units. Some initial grouping errors caused by the inaccurate dominant pitch are corrected. In addition, segments merged by some T-F units corresponding to unresolved harmonics of target speech, are also composed. These segments are assigned to the foreground stream. Here, the same as in the initial segregation stage, OQAS is also introduced into the CASA system and is used as a rule to judge whether the relevant segments, which are composed of units that cannot be assigned to the foreground, could be grouped into the background. Then, the foreground stream is expanded to include neighboring T-F units labeled as speech dominant.

Finally, using the method proposed by Weintraub [10], we resynthesize a speech waveform from the resulting foreground stream. The resynthesis process can be treated as a binary mask, that is, the T-F units within which the target speech is dominant are labeled as 1, and otherwise labeled as 0. Then, the acoustic energy from the mixture corresponding to 1 in the mask is retained and the mixture energy corresponding to 0 is rejected. More details on the resynthesis are referred to [5], [9], [10].

Since many processing steps of the proposed model are similar to the Hu and Wang's model, we do not touch them much in these parts. More details could be found in [11, Sec. III–VI]. In this paper, we just focus on how to introduce the objective speech quality assessment into the Hu and Wang's model and

describe the difference of the proposed model from the Hu and Wang's model.

For the convenience of comparison, we list all the terms described above and provide their definitions again as follows.

T-F unit	a very local time-frequency region which corresponds to a certain filter at a certain time frame;
segment	a contiguous T-F region corresponding to a component of a single sound source, and it is a set of connected T-F units;
stream	a group of segments which corresponds to an entire sound source;
target speech (target stream)	an utterance that should be separate from an acoustic mixture. (It is no doubt that the constitution of target speech is task-dependent. But in our study, target speech refers to an entirely voiced utterance in a mixture.)

III. Combine CASA With OQAS

This section concentrates on describing in details how to combine CASA with OQAS into one system. As a whole, the introduction of OQAS into the CASA system is realized by selecting the segments which are more probably originated from the same source in perception sense. Since in the initial segregation stage, whether one segment belongs to the foreground stream or the background stream is determined by a simple decision, the assignments of the segments are not accurate enough and are not tightly related to the quality of perception [11]. Therefore, the ITU-T P.563 algorithm is adopted to check the accuracy of the assignments of the segments by CASA so as to improve the final perceptual quality of the separated speech.

Concretely, in our model, there are two parts where OQAS is directly inserted into the CASA system. One is in the initial segregation stage and the other is in the final segregation stage. Details are provided next.

A. OQAS in the Initial Segregation Stage

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

In Hu and Wang's model, the grouping is realized by comparing the periods computed from all the T-F units in a segment with the dominant pitch contour. For any segment, if there are more than half of its units at a certain frame agreeing with the dominant pitch, the segment is assumed to agree with the dominant pitch at this frame. Actually, for a target speech, if the dominant pitch is very close to the true pitch at a certain frame, all the segments in this target speech would tend to be consistent with the dominant pitch at this frame. Therefore, we can group the segments into two streams. First, we could select the longest segment as a seed stream. Since in this study, only voiced target speech is considered, it is reasonable to assume that the longest segment will extend through most of the frames of the entire utterance duration. At a certain frame, if a segment and the longest segment both agree or disagree with the dominant pitch, they are said to be consistent. If more than half of the overlapping frames between a segment and the longest segment are consistent, the T-F units of this segment within the duration of the longest segment will be grouped into the seed stream. Otherwise, this segment will be grouped into the competing stream. In addition, we can also use the longest segment to determine which stream corresponds to target speech. If more than half of the frames of the longest segment are consistent with the dominant pitch, the longest segment is thought to contain dominant target speech. In this condition, the stream containing the longest segment would be regarded as the foreground stream, while the competing stream as the background stream. Otherwise, the names of the two streams should be exchanged.

It is obvious that in the initial grouping stage a simple decision plays an important role. Although the grouping result of the simple decision could be adjusted in the following stage through iterative estimation and linear interpolation so as to give an acceptable prediction of pitch contour, it yet does not satisfy the requirements of the segregation and may also deliver 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 the foreground and background streams corresponding to the target speech and intrusion, respectively, a method consisting of two steps' processing is employed in our model.

In the first step, a simple decision the same as that in the Hu and Wang's model is adopted. Via a more conservative plausible pitch range θ_p, which is set to 0.90 (corresponding to the threshold θ_p, which is 0.95, in the Hu and Wang's model [11]), the most improbable segments, in which the intrusion is dominant, would be filtrated. After this, a coarse grouping is acquired (it will be processed further).

In the second step, the objective quality assessment is employed to give more accurate predictions to classify the foreground and the background. Since many T-F units with some response energy and sufficiently high-cross-channel correlations have been selected to form the segments which are directly relevant to the perceptual quality of speech, and have already been divided into the foreground and the background streams, we decided to use the binary masking method to resynthesize a temporary speech, and then adopt the P.563 algorithm to evaluate its quality.

Fig. 2 gives a schematic diagram of this combination. From Fig. 2, we can see that the whole evaluation process is divided into two stages. In the first stage, we select the speech resynthesized by preserving 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 perceptual quality by the P.563 algorithm. Then, in order to choose the reliable segments in which the target speech is dominant, a comparison is introduced into the process. By 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 processing), a temporary speech is resynthesized. Using the P.563 algorithm to evaluate the resynthesized speech and comparing its quality

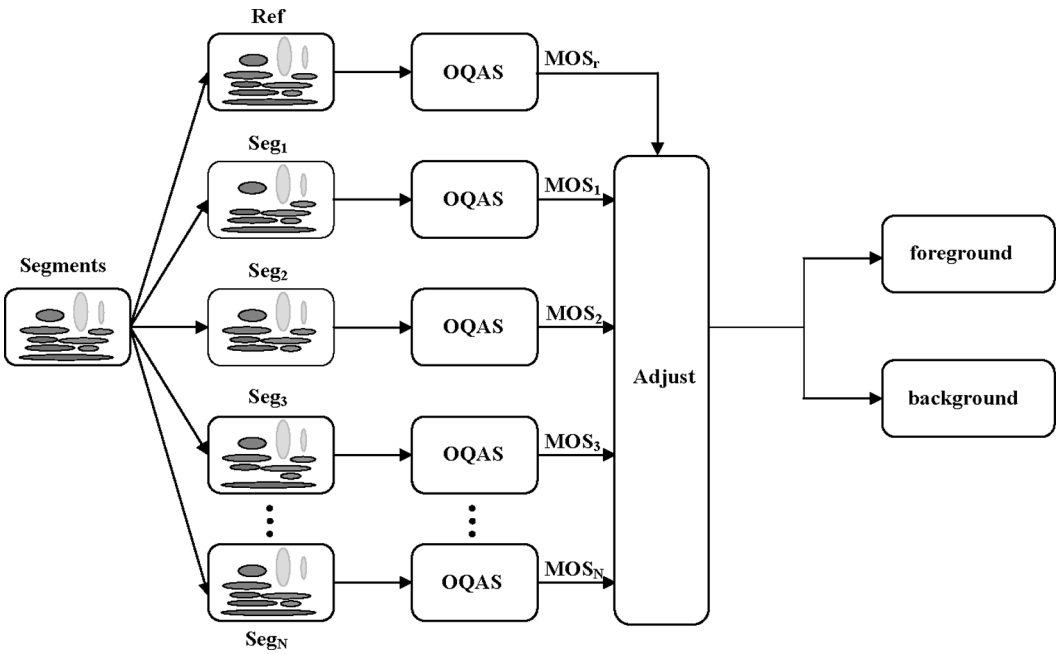


Fig. 2. Sketch map of the combination in initial segregation stage.

with the reference signal mentioned above, we can easily evaluate the effect that the masked segment brings to the speech's quality. More specifically, if there are N segments after primitive grouping in the foreground, then we should use the P.563 algorithm $N+1$ times to assess the quality of speeches resynthesized by masking the corresponding segment in the foreground. If MOS_i, the quality of a speech synthesized by masking the ith segment in the foreground, is higher than the quality of the reference speech MOS_r, it means that if this segment is set to the background, the quality of the resynthesized speech will be improved. That is, this segment can be adjusted to the background, and vice versa. The judgment described above can be performed by the following equation:

$$\mathrm{Mask}(i) = \begin{cases} 1, & \text{if } \mathrm{MOS}_i - \mathrm{MOS}_r \le 0 \\ 0, & \text{if } \mathrm{MOS}_i - \mathrm{MOS}_r > 0 \end{cases} \quad i = 1, 2, \ldots, N \tag{1}$$

where, $\mathrm{Mask}(i)$ is the masking value of the ith segment in the foreground, and 1 represents foreground, while 0 background.

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 it is set to 0.02. Then (1) could be modified as

$$\mathrm{Mask}(i) = \begin{cases} 1, & \text{if } \mathrm{MOS}_i - \mathrm{MOS}_r \le \theta_A \\ 0, & \text{if } \mathrm{MOS}_i - \mathrm{MOS}_r > \theta_A \end{cases} \quad i = 1, 2, \ldots, N. \tag{2}$$

This conservative processing can alleviate the error adjustment caused by the objective quality assessment and avoid deleting too many useful segments from the foreground. After this adjustment, the segments still kept in the foreground are assumed to be more probably originated from the target source. Following the pitch tracking and unit label steps used in the Hu and Wang's model, the dominant pitch estimated from the segments in the foreground is closer to the true 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

Because the spectra of the target speech and the intrusion often overlap, some segments created in the former segmentation would still contain the units where the target dominates as well as the units where the intrusion dominates. Since the process of pitch label has generated a label to each unit, it is possible to further split a segment in the foreground into smaller ones to make all the units in a segment have the same label.

The segments in the foreground can be adjusted as follows.

1) If a segment is labeled as target and is no shorter than 50 ms, it should be kept in the foreground.
2) If a segment is labeled as intrusion and is no shorter than 50 ms, it will be evaluated by OQAS first. Then, if it does not make the speech quality decrease, it will be retained in the foreground. Otherwise, it should be sent into the background.
3) Segments not belonging to the above two types are removed from the foreground. Note that they are not adjusted to the background; they just become undecided.

It is clear that in Step 2), OQAS is introduced into the separation system again. Here, the combination of OQAS and CASA is similar to that in the initial segregation stage. By masking every segment no shorter than 50 ms with the intrusion label in the foreground, and evaluating the corresponding resynthesized speech's perceptual quality, these segments are adjusted to the foreground or the background and form new foreground and background streams, respectively. But the two combinations are

TABLE I
SNR RESULTS.
(**MIXTURE**: THE RESULTS OF THE ORIGINAL DEGRADED SPEECH. **PROPOSED**: THE RESULTS OF THE PROPOSED MODEL. **SS**: THE RESULTS OF THE SPECTRAL SUBTRACTION ENHANCEMENT ALGORITHM. **HU WANG**: THE RESULTS OF HU AND WANG'S MODEL. **TRUE PITCH**: THE RESULTS OF THE PROPOSED MODEL WHERE THE TRUE PITCH IS USED. **IDEAL MASK**: THE RESULTS OF IDEAL BINARY MASKING.)

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**
Proposed	***11.129***	***3.507***	***14.411***	***5.218***	***6.669***	***12.933***	***14.662***	***9.391***	***11.506***	***3.964***	***9.339***
SS	7.568	-3.879	6.207	2.651	3.154	-9.609	0.950	4.709	9.968	3.513	**2.523**
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**

different in some sense. In this stage, the judgment in the combination is made according to (1), while in the initial segregation stage it is performed according to (2).

After the above processing, background stream begins to expand iteratively to involve undecided segments in its neighborhood. When the expansion of the background is over, all the segments remaining in the undecided segments are added back to the foreground.

Finally, those individual units belonging to neither stream are assigned to the foreground stream iteratively if their labels are target and they are located in the neighborhood of the foreground stream. This processing forms the final segregated stream of target speech. Then, the remaining units are grouped into the background. According to the final grouping results, a waveform of segregated speech can be resynthesized.

IV. EVALUATION AND COMPARISON

This section will provide the evaluation of our model in SNR and perceptual quality on a standard corpus and compare its performance with other models.

A. Evaluation

Our model is evaluated with a corpus of 100 mixtures composed of ten voiced utterances mixed with ten intrusions collected by Cooke [9], which has been used to test CASA systems [5]–[7], [9], [23]. The intrusions have a considerable variety. Specifically, the ten 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 [24]. We use both SNR and perceptual quality as the criterion to quantitatively assess the performance of the proposed separation system.

The original sampling frequency of the corpus is 16 kHz. Since the OQAS algorithm utilized in this paper is only used to evaluate narrow-band telephone speech, the corpus is downsampled to 8 kHz. Besides, there are also some requirements which should be satisfied for using the ITU-T P.563 algorithm to test speech quality. These requirements include that the minimum length of active speech in test speech is 3 s, the maximum signal length is 20 s, 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 1.5–2 s long, it is obvious that the test speeches do not satisfy the requirements. In order to apply the P.563 algorithm, we repeat the test speech three times with an interval of 0.5 s. This processing not only makes the test speeches satisfy the requirements of OQAS, but also has no effect on the assessment results. It should be noted that the P.563 algorithm cannot give correct result in some intrusion conditions, such as the music, male speech, female speech, etc. Although it is a limitation difficult to be avoided, the P.563 algorithm is still adopted in all the test intrusion conditions because the test speeches in our application are always the signals 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 the 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 ten target utterances. The average over all intrusions is shown in the last column of the table. As shown in the table, our system improves the SNR for every intrusion and produces a gain of 9.7 dB over the original mixtures. Larger SNR improvements are obtained for the intrusions whose spectra do not significantly overlap with those of target utterances (e.g., N0 and N5), while less improvements are obtained for the intrusions with significant overlap (e.g., N3, N8, and N9).

Of course, SNR does not indicate the perceptual quality of the resynthesized speech signal. For example, the model could retrieve a small proportion of the speech energy and totally reject the noise; it will give a very high SNR, but the resynthesized speech is unintelligible. Accordingly, we append the SNR metric with a measure of speech quality before and after segregation.

As discussed in Section I, there are two ways to estimate the speech quality. One is the objective method and the other is the subjective method. The objective method can estimate speech perceptual quality fast and easily, but the result is only an approximate one and may be different from the real result. On the other hand, although the subjective method can offer more accurate measurement, it is very expensive both in time and cost. Considering the advantage and disadvantage of each method, we adopt both of the two methods to evaluate the performance of our system sufficiently.

To fulfill the objective measurement of speech quality, we employ two OQAS algorithms. One is the ITU-T P.862 standard,

TABLE II
P.862 RESULTS

P.862	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**
Proposed	***2.673***	***0.699***	***2.247***	***1.087***	***1.162***	***2.255***	***2.372***	***1.761***	***1.886***	***1.306***	***1.745***
SS	2.487	1.209	1.046	1.531	1.039	-0.121	2.017	1.386	1.954	1.813	**1.347**
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
P.563 RESULTS

P.563	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**
Proposed	***4.214***	***3.720***	***4.269***	***3.838***	***3.783***	***3.748***	***3.976***	***3.628***	***3.805***	***3.605***	***3.858***
SS	3.190	1.000	5.000	2.304	3.298	3.272	2.456	4.321	4.344	3.951	**3.314**
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**

TABLE IV
MOS RESULTS

MOS	N0	N1	N2	N3	N4	N5	N6	N7	N8	N9	Ave
Mixture	1.59	1.03	1.67	1.28	1.19	1.36	1.33	1.25	1.30	1.26	**1.326**
Proposed	***3.36***	***1.50***	***3.13***	***2.11***	***2.02***	***3.04***	***3.01***	***2.49***	***2.35***	***1.96***	***2.497***
SS	2.13	1.07	1.78	1.31	1.21	1.24	1.36	1.28	1.32	1.32	**1.402**
Hu Wang	3.26	1.41	2.95	1.90	1.41	2.89	2.86	2.34	2.39	2.16	**2.357**
True Pitch	3.62	1.77	3.32	2.47	2.46	3.29	3.21	2.89	2.86	2.69	**2.858**
Ideal Mask	3.85	1.97	3.69	2.92	2.83	3.55	3.42	3.22	3.22	3.16	**3.183**

which is an intrusive objective speech quality assessment algorithm [25]. Since the original speech before mixing is available, it is convenient to apply the ITU-T P.862 algorithm to obtain the intrusive speech quality evaluation result of the separated speech. The other is the ITU-T P.563 algorithm, which is just the OQAS algorithm inserted in our system. With this algorithm, the nonintrusive objective assessment results can be easily gained just given the test speech.

To implement the subjective measurement, a subjective MOS experiment is performed. Ten listeners are ordered to score the test speeches according to their perceptual quality. For each test speech, the average score of the ten listeners is treated as its MOS.

For each speech, we can get three results of perceptual quality. Among these quality results, it is no doubt that the most reliable and accurate one is the MOS obtained by the subjective measurement. However, it does not mean that the objective assessment results are unimportant and could be ignored. Actually, although the objective results may be not accurate enough, they could also reflect the level of performance coarsely and offer some approximate results. Thus, we list the results of P.862, P.563, and MOS in Tables II–IV, respectively, including those of our model and original mixtures.

Since the P.563 algorithm is introduced into our model to find an optimal grouping of the segments, it is not unfair to calculate the perceptual quality of the separated speech with the P.563 algorithm, because it is no doubt that the calculated perceptual quality of the speech processed by our model would inevitably be higher than those processed by other models. Moreover, the output of P.563 is not accurate enough in some intrusion conditions, especially to the mixture in music noises and female speech intrusion. Considering these shortcomings, we just list the speech quality calculated by P.563 as an additive reference.

In Tables II–IV, each value in the tables represents the average speech quality for one intrusion mixed with ten target utterances. The average over all intrusions is shown in the last column of the table, too. As shown in the tables, our system improves speech's perceptual quality for most of the intrusions. Especially in Table IV, the MOSs acquired by means of subjective measurement are improved in all intrusions. The average MOS over all intrusions is 1.326 for the original mixtures and 2.497 for the speeches separated by proposed method. The MOS improvement is about 1.17. It means that the speech quality is improved more than one level in perceptual sense since there are only five levels over the measurement range. More specifically, it is observed that for the original mixtures, the average speech quality is a little higher than the worst quality and unintelligible, while for the proposed method, it is just a little lower than the normal quality and mostly intelligible. Furthermore, we can also find that larger MOS improvements are achieved for the

intrusions whose spectra do not significantly overlap with those of target utterances (e.g., N0 and N5), whereas less MOS improvements are obtained for intrusions with significant overlap (e.g., N3 and N9).

B. Comparison With Other Models

The evaluation of our model has been elaborated and it can be seen that the proposed model can effectively improve the SNRs and most perceptual qualities of the separated speeches. However, is this proposed model better than others? How good on earth is it? In order to answer these two questions, we compare the SNRs and perceptual quality of the original mixtures and the speeches separated by our model together with the results of some other typical separation or enhancement systems.

First, the proposed method is compared with the spectral subtraction algorithm [26], [27], a famous method for speech enhancement. The spectral subtraction method is implemented 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. However, to N2 intrusion, because it contains a sequence of short noise bursts, spectral subtraction is applied within each burst. Tables I–IV show the results of the spectral subtraction method. From the tables, it can be easily concluded that the spectral subtraction method performs significantly worse than our system in SNR. It is because of its well-known deficiency in dealing with nonstationary interference. In the perceptual quality aspect, the separated speeches' average perceptual qualities of our model, calculated by P.563, P.862 and MOS, are all better than spectral subtraction. However, the results of speech quality in different intrusion conditions are not always better than spectral subtraction, especially for the results of P.862 in complex intrusion. Although it is difficult to explain why for the proposed model the results of P.862 for different intrusions are not improved consistently, it can be supposed that the main reason lies in that too much information in the separated speech is lost and cannot be recovered due to the masking processing in the condition that intrusion has significant overlap with target speech.

Our system is also compared with the Hu and Wang's model. Since the main processing stages of their model are similar to our model except the combination with OQAS algorithm, it is valuable to evaluate the effect of the introduction of OQAS on the performance of separation system from this comparison. The SNR and perceptual quality results of Hu and Wang's model are listed in Tables I–IV. From these tables, we can find that SNRs of the speeches separated by our system are better than Hu and Wang's model not only in average level, but also in almost every intrusion. The P.862 and MOS results of our model are also better than the results of Hu and Wang's model except for some intrusions such as male and female speeches. In fact, the exception mentioned above is caused by the inaccurate assessment of speech quality in the separation stage. It has been emphasized that the P.563 algorithm cannot work well in the speech intrusion condition. Although the binary mask reduces many intrusion energy of the mixture speech, it does not absolutely break the limitations. To solve this problem effectively, a more appropriate way should be explored to combine CASA with OQAS more tightly.

Obviously, the error in the pitch estimation or the pitch-based grouping may bring about segregation errors. To evaluate the proposed model more strictly, we employ the 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 Tables I–IV give the SNR and perceptual quality results separately for our system using true pitch instead of estimated pitch. With true pitch, the system performs only slightly better. It implies that estimated pitch of our system is quite accurate and the performance of proposed system is very good.

Given the objective of identifying T-F regions that target is dominant, we use ideal binary mask as the ground truth of target stream. The ideal masks are supported by the auditory masking phenomenon: within a critical band, a weaker signal is masked by a stronger one [21]. It is performed by reserving the T-F regions where the energies of a target sound are stronger than the interference, while discarding the regions where the energies of target sound is weaker than the interference. Ideal binary masking has a number of desirable properties, such as flexibility, well-definedness, etc. [22]. It sets the ceiling performance for all binary masks and is broadly consistent with ASA constraints in terms of audibility and segregation capacity. Since the target and intrusion before mixing is available, ideal binary masks can be easily constructed. The SNR and perceptual quality results of ideal binary masks are shown in Tables I–IV. They are uniformly better than all other results mentioned before. Compared with our proposed model, the average SNR improvement for the entire corpus is about 5.0 dB and the average perceptual quality improvement is about 0.8 in P.862 and 0.7 in MOS. This indicates how much our model could be further improved in terms of conventional SNR and speech perceptual quality.

V. Conclusion

In this paper, a system is proposed to segregate voiced speech based on the primitive CASA system combined with OQAS algorithm. The CASA model in our system analyzes temporal information in the input, the temporal fine structure of a resolved harmonic and the temporal envelope of an unresolved harmonic. It has been proven 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 adopted to model auditory perception, and have explained many observed perceptual phenomena successfully [28]–[30]. The OQAS algorithm is used to classify foreground and background streams. The segments, which are still kept in foreground after OQAS processing, are used to estimate the dominant pitch. This effectively increases the accuracy of pitch estimation and as a result improves the performance of the separation system.

Similar to the 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 the previous studies in the following aspects.

First, by combining CASA with OQAS, we introduced the knowledge on speech perceptual quality into separation and construct a direct link between separated speech and its

perceptual quality. It is the first attempt to introduce the speech quality assessment into CASA systems. From the results of the proposed system, it can be concluded that the link between separated speech and its perceptual quality is valuable to solve speech separation problem.

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

In comparison of the proposed algorithm with other separation or enhancement systems, we can draw a conclusion that our method is effective in processing the monaural speech separation problem.

Of course, there are also some deficiencies in our research. We must pay more attention to them, analyze the reasons, 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, segments in the foreground used to estimate the pitch contour are very important. In Section III, we have emphasized that the classification of the foreground and the background in the initial segregation stage is mainly based on the OQAS algorithm. In fact, although the ITU-T P.563 algorithm employed by our system works very well, it is still a machine estimation and the obtained result is more or less different from the subjective MOS (especially for the segregation task that may introduce new kinds of distortions). Therefore, it seems that it is a key problem to improve the accuracy of the OQAS algorithm. In addition, since our research on combining CASA with OQAS is at a primary stage, only a simple method is found to combine the whole OQAS algorithm with CASA systems and guides the classification of the foreground and the background. This combination does not make full use of the knowledge of speech quality and thus it is necessary to study it further so as to seek a better combination form. This perhaps needs to break OQAS algorithm into local units, or only introduce some of its basic principles into the separation process. It will be our objective in the future to find an optimal combination method of CASA and OQAS.

Our model performs separation based only on pitch. That is to say it can only be used to segregate voiced speech. In fact, unvoiced speech remains a great challenge for monaural speech segregation. To solve the problem, other grouping cues, such as onset, offset, and timbre, which have been demonstrated to be effective for human ASA [4], [31], should be employed in the segregation system. Furthermore, acoustic and phonetic characteristics of individual unvoiced consonants should be considered too. These issues will be investigated in our future work.

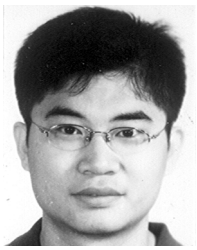
ACKNOWLEDGMENT

The authors would like to thank G. Hu and D.-S. Kim for their fruitful discussions and help. Thanks are also due to two anonymous reviewers for their great help in improving the structure of this paper.

REFERENCES

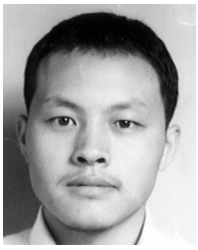
[1] C. Cherry, "Some experiments in the recognition of speech with one and two ears," *J. Acoust. Soc. Amer.*, vol. 25, pp. 975–981, 1953.
[2] A. K. Barros, T. Rutkowski, F. Itakura, and N. Ohnishi, "Estimation of speech embedded in a reverberant and noisy environment by independent component analysis and wavelets," *IEEE Trans. Neural Netw.*, vol. 13, no. 4, pp. 888–893, Jul. 2002.
[3] H. Krim and M. Viberg, "Two decades of array signal processing research: The parametric approach," *IEEE Signal Process. Mag.*, vol. 13, no. 4, pp. 67–94, Jul. 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 Lang.*, 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., Mass. Inst. Technol., Cambridge, 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 Netw.*, vol. 10, no. 3, 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 Netw.*, vol. 15, no. 5, pp. 1135–1150, Sep. 2004.
[12] N. Roman, D. L. Wang, and G. J. Brown, "Speech segregation based on sound localization," *J. Acoust. Soc. Amer.*, vol. 114, pp. 2236–2252, 2003.
[13] D. Godsmark and G. J. Brown, "A blackboard architecture for computational auditory scene analysis," *Speech Commun.*, vol. 27, pp. 351–366, 1999.
[14] "Subjective performance assessment of telephone-band and wideband digital codecs," ITU, Geneva, Switzerland, 1996, ITU-T Rec. P.830.
[15] P. Gray, M. P. Hollier, and R. E. Massara, "Nonintrusive speech-quality assessment using vocal-tract models," *Proc. Inst. Elect. Eng.-Vision, Image Signal Process.*, 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. Int. Conf. Acoust., Speech, Signal Process.*, May 1996, vol. 1, pp. 491–494.
[17] D. S. Kim, "ANIQUE: An auditory model for single-ended speech quality estimation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 13, no. 5, pp. 821–831, Sep. 2005.
[18] "Single-ended method for objective speech quality assessment in narrow-band telephony applications," ITU, Geneva, Switzerland, 2004, ITU-T P.563.
[19] NiQA—product description Psytechnics Limited, 2003 [Online]. Available: http://www.psytechnics.com/pages/products/niqa.php
[20] NiNA—SwissQual's Non-intrusive algorithm for estimating the subjective quality of live speech Swiss Qual Inc., 2001 [Online]. Available: http://www.swissqual.com/HTML/ninapage.htm
[21] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 4th ed. San Diego, CA: Academic, 1997.
[22] D. L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed. Norwell, MA: Kluwer, 2005, pp. 181–197.
[23] 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.
[24] D. F. Rosenthal and H. G. Okuno, *Computational Auditory Scene Analysis*. Mahwah, NJ: Lawrence Erlbaum, 1998.
[25] Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs ITU, Geneva, Switzerland, 2001, ITU-T P.862.
[26] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, Feb. 1979.
[27] Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.
[28] P. Cariani, "Temporal coding of periodicity pitch in the auditory system: An overview," *Neural Plasticity*, vol. 6, pp. 147–172, 1999.
[29] R. Meddis and L. O'Mard, "A unitary model of pitch perception," *J. Acoust. Soc. Amer.*, vol. 102, pp. 1811–1820, 1997.

[30] 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.
[31] G. Hu and D. L. Wang, "Separation of fricatives and affricates," in *Proc. ICASSP*, 2005, vol. 1, pp. 1101–1104.



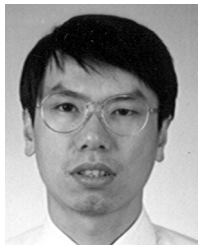
Peng Li received the B.S. and M.S. degrees in automation from Tianjin University, Tianjin, China, in 2000 and 2003, respectively. He is currently working toward the Ph.D. degree in pattern recognition at the Institute of Automation, Chinese Academy of Sciences, Beijing, China.

His research interests include speech segregation, computational auditory scene analysis, speech enhancement, noise reduction, and speech recognition.



Yong Guan received the B.S. degree 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, and speaker identification.



Bo Xu received the B.S. degree from Zhejiang University, Hangzhou, China, in 1988 and the M.S. and Ph.D. degrees from the Speech Recognition Research Center, National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 1992 and 1997, respectively.

His current interests mainly focus on multimedia content management and statistical speech translation. He is the Associate President of CASIA and Deputy Director of the NLPR. He was the Coordinator of International Consortium for Speech Translation Advanced Research (C-Star). He is on the steering committee of the National High-Tech Program in the fields of Chinese information processing and intelligent interfaces.

Wenju Liu received the B.S. degree in mathematics from Beijing University, Beijing, China, in 1983, the M.S. degree in mathematics from and Beijing University of Post and Telecommunications, Beijing, in 1989, and the Ph.D. degree in computer applications from Tsinghua University, Beijing, in 1993.

He is currently an Associate Professor at the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy Of Sciences, Beijing. His research interests include speech recognition, speech synthesis, speaker recognition, voice conversion, computational auditory scene analysis, speech enhancement, and noise reduction.