Available online at www.sciencedirect.com
ScienceDirect
Computer Speech and Language xxx (2008) xxx–xxx
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

A computational auditory scene analysis system for speech segregation and robust speech recognition

Yang Shao [a,*], Soundararajan Srinivasan [b,1], Zhaozhang Jin [a], DeLiang Wang [a,c]

[a] Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210, USA
[b] Biomedical Engineering Department, The Ohio State University, Columbus, OH 43210, USA
[c] Center for Cognitive Science, The Ohio State University, Columbus, OH 43210, USA

Received 5 July 2007; received in revised form 17 January 2008; accepted 3 March 2008

Abstract

A conventional automatic speech recognizer does not perform well in the presence of multiple sound sources, while human listeners are able to segregate and recognize a signal of interest through auditory scene analysis. We present a computational auditory scene analysis system for separating and recognizing target speech in the presence of competing speech or noise. We estimate, in two stages, the ideal binary time–frequency (T–F) mask which retains the mixture in a local T–F unit if and only if the target is stronger than the interference within the unit. In the first stage, we use harmonicity to segregate the voiced portions of individual sources in each time frame based on multipitch tracking. Additionally, unvoiced portions are segmented based on an onset/offset analysis. In the second stage, speaker characteristics are used to group the T–F units across time frames. The resulting masks are used in an uncertainty decoding framework for automatic speech recognition. We evaluate our system on a speech separation challenge and show that our system yields substantial improvement over the baseline performance.

Keywords: Speech segregation; Computational Auditory Scene Analysis; Binary time–frequency mask; Robust speech recognition; Uncertainty decoding

1. Introduction

In everyday listening conditions, the acoustic input reaching our ears is often a mixture of multiple concurrent sound sources. While human listeners are able to segregate and recognize a target signal under such conditions, robust automatic speech recognition remains a challenging problem (Huang et al., 2001). Automatic speech recognition (ASR) systems are typically trained on clean speech and face the mismatch problem when

* Corresponding author. Tel.: +1 614 292 7402.
E-mail addresses: shaoy@cse.ohio-state.edu (Y. Shao), srinivasan.36@osu.edu (S. Srinivasan), jinzh@cse.ohio-state.edu (Z. Jin), dwang@cse.ohio-state.edu (D. Wang).
[1] Present address: Research and Technology Center, Robert Bosch LLC, Pittsburgh, PA 15212, USA.

0885-2308/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2008.03.004

tested in the presence of interference. In this paper, we address the problem of recognizing speech from a target speaker in the presence of either another speech source or noise.

To mitigate the effect of interference on recognition, speech mixtures can be preprocessed by speech separation algorithms. Under monaural conditions, systems typically depend on modeling the various sources in the mixture to achieve separation (Ephraim, 1992; Jang and Lee, 2003; Kristjansson et al., 2004; Roweis, 2005; Raj et al., 2005). An alternate approach to employing speech separation prior to recognition involves the joint decoding of the speech mixture based on knowledge of all the sources present in the mixture (Varga and Moore, 1990; Gales and Young, 1996; Deoras and Hasegawa-Johnson, 2004). These model-based systems rely heavily on the use of *a priori* information of sound sources. Such approaches are fundamentally limited in their ability to handle novel interference (Allen, 2005). For example, systems that assume and model the presence of multiple speech sources only, do not lend themselves easily to handling speech in (non-speech) noise conditions.

In contrast to the above model-based systems, we present a primarily feature-based computational auditory scene analysis (CASA) system that makes weak assumptions about the various sound sources in the mixture. It is believed that the human ability to function well in everyday acoustic environments is due to a process termed auditory scene analysis (ASA), which produces a perceptual representation of different sources in an acoustic mixture (Bregman, 1990). In other words, listeners organize the mixture into *streams* that correspond to different sound sources in the mixture. According to Bregman (1990), organization in ASA takes place in two main steps: segmentation and grouping. Segmentation decomposes the auditory scene into groups of contiguous time–frequency (T–F) units or segments, each of which mainly originates from a single sound source (Wang and Brown, 2006). A T–F unit denotes the signal at a particular time and frequency. Grouping involves combining the segments that are likely to arise from the same source together into a single stream (Bregman, 1990). Grouping itself is comprised of simultaneous and sequential organizations. Simultaneous organization involves grouping of segments across frequency, and sequential organization refers to grouping across time.

From an information processing perspective, the notion of an ideal binary T–F mask has been proposed as a major computational goal of CASA by Wang (2005). Such a mask can be constructed from the *a priori* knowledge of target and interference; specifically a value of 1 in the mask indicates that the target is stronger than the interference within the corresponding T–F unit and 0 indicates otherwise. The use of ideal binary masks is motivated by the auditory masking phenomenon in which a weaker signal is masked by a stronger one within a critical band (Moore, 2003). Additionally, previous studies have shown that such masks can provide robust recognition results (Cooke et al., 2001; Roman et al., 2003; Srinivasan et al., 2006; Srinivasan and Wang, 2007). Hence, we propose a CASA system that estimates this mask to facilitate the recognition of target speech in the presence of interference. When multiple sources are of interest, the system can produce ideal binary masks for each source by treating one source as target and the rest as interference.

In this paper, we present a two-stage monaural CASA system that follows the ASA account of auditory organization as shown in Fig. 1. The input to the system is a mixture of target and interference. The input mixture is analyzed by an auditory filterbank in successive time frames. The system then generates segments based on periodicity and a multi-scale onset and offset analysis, producing voiced and unvoiced segments respectively (Hu and Wang, 2006). In the simultaneous grouping stage, the system estimates pitch tracks of individual sources in the mixture and employs periodicity similarity to group voiced segments into simultaneous streams. A simultaneous stream comprises multiple segments that overlap in time. Subsequently, a sequential grouping stage employs speaker characteristics to organize simultaneous streams and unvoiced segments across time into whole streams corresponding to individual speaker utterances. Within this stage, we first sequentially group simultaneous streams corresponding to voiced speech and then unvoiced segments. Finally, the CASA system outputs an estimate of the ideal binary mask corresponding to an underlying speaker in the input mixture. Such a mask is then used in an uncertainty decoding approach to robust speech recognition (Srinivasan and Wang, 2007). This approach reconstructs missing feature components as indicated by the binary masks, and also incorporates reconstruction errors as uncertainties in the speech recognizer. Finally, in the case of multiple speech sources, a target selection process is employed for identifying the target speech.

The rest of the paper is organized as follows. Sections 2–5 provide a detailed presentation of the various components of our proposed system. The system is systematically evaluated on a speech separation challenge

Please cite this article in press as: Shao, Y. et al., A computational auditory scene analysis system for speech ..., Computer Speech and Language (2008), doi:10.1016/j.csl.2008.03.004

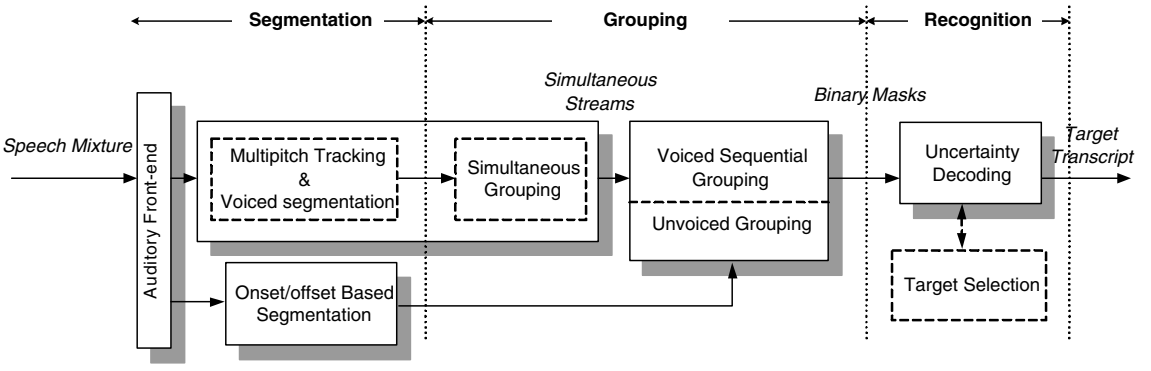

Fig. 1. Schematic diagram of the proposed two-stage CASA system and its application to ASR. The auditory front-end decomposes input signal into a T–F representation called a cochleagram (see Section 2). In the segmentation stage, our system generates both voiced segments and unvoiced segments. Then in the grouping stage, a simultaneous grouping process uses periodicity similarity to group voiced components and produces simultaneous streams. Subsequently, a sequential grouping algorithm organizes these simultaneous streams and unvoiced segments across time. The resulting binary T–F masks are used by an uncertainty decoder and a target selection mechanism to recognize the target utterance.

(SSC) task that involves the recognition of a target speech utterance in the presence of either a competing speaker or speech-shaped noise (Cooke and Lee, 2006) The evaluation results are presented in Section 6. Finally, conclusions and future work are given in Section 7.

2. Auditory based front-end

Our system first models auditory filtering by decomposing an input signal into the time–frequency domain using a bank of Gammatone filters. Gammatone filters are derived from psychophysical observations of the auditory periphery and this filterbank is a standard model of cochlear filtering (Patterson et al., 1992) The impulse response of a Gammatone filter centered at frequency f is:

$$g(f,t) = \begin{cases} t^{a-1}\,\mathrm{e}^{-2\pi bt}\cos(2\pi ft), & t \geqslant 0. \\ 0, & \text{else}, \end{cases} \tag{1}$$

t refers to time; $a = 4$ is the order of the filter; b is the rectangular bandwidth which increases with the center frequency f. We use a bank of 128 filters whose center frequency f ranges from 50 Hz to 8000 Hz. These center frequencies are equally distributed on the ERB scale (Moore, 2003) and the filters with higher center frequencies respond to wider frequency ranges.

Since the filter output retains original sampling frequency, we down-sample the 128 channel outputs to 100 Hz along the time dimension, which includes low-pass filtering and decimation. This yields the corresponding frame rate of 100 Hz, which is used in many short-term speech feature extraction algorithms (Huang et al., 2001). The magnitudes of the down-sampled outputs are then loudness-compressed by a cubic root operation. The resulting responses compose a matrix, representing a T–F decomposition of the input. We call this T–F representation a cochleagram (Wang and Brown, 2006), analogous to the widely used spectrogram. Note that unlike the linear frequency resolution of a spectrogram, a cochleagram provides a much higher frequency resolution at low frequencies than at high frequencies. We base our subsequent processing on this T–F representation. Fig. 2 shows a cochleagram of a two-talker mixture. The signals from these two talkers are mixed at 0 dB signal-to-noise ratio (SNR).

We call a time frame of the above cochleagram a Gammatone feature (GF). Hence, a GF vector comprises 128 frequency components. Note that the dimension of a GF vector is much larger than that of feature vectors used in a typical speech recognition system. Additionally, because of overlap among neighboring filter channels, the Gammatone features are largely correlated with each other. Here, we apply a discrete cosine transform (DCT) (Oppenheim et al., 1999) to a GF in order to reduce its dimensionality and de-correlate its components. The resulting coefficients are called Gammatone frequency cepstral coefficients (GFCC) (Shao

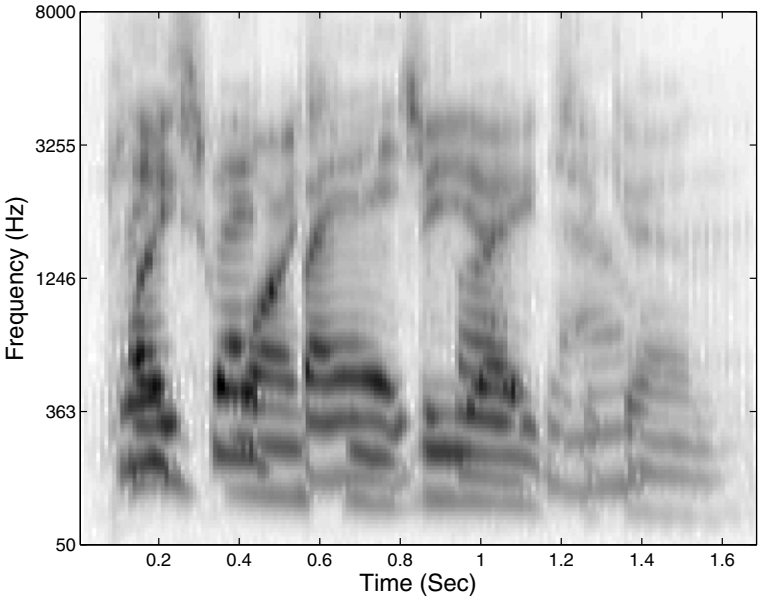

Fig. 2. Cochleagram of a two-talker mixture mixed at 0 dB signal-to-noise ratio. Darker color indicates stronger energy within the corresponding time–frequency unit.

et al., 2007). Rigorously speaking, the newly derived features are not cepstral coefficients because a cepstral analysis requires a log operation between the first and the second frequency analysis for the purpose of deconvolution (Oppenheim et al., 1999). Here we regard these features as cepstral coefficients because of the functional similarities between the above transformation and that of a typical cepstral analysis.

In the previous study by Shao et al. (2007), the 23 lowest order GFCC coefficients are used as a feature vector due to computational considerations. After performing inverse DCT of GFCCs, we find that by including up to 30 coefficients almost all the GF information is captured while the GFCCs above the 30th are close to 0 numerically (Shao, 2007). Hence, we use the 30-dimensional GFCCs as a feature vector in this paper. Fig. 3 illustrates a GFCC transformed GF and a cochleagram. The top plot shows a comparison of a GF frame at 1 s in Fig. 2 and the resynthesized GF from its 30 GFCCs. The bottom plot presents the resynthesized cochleagram from Fig. 2 using 30 GFCCs. As can be seen from Fig. 3 the 30 lowest order GFCCs retain the majority information in a 128-dimensional GF. This is due to the "energy compaction" property of the DCT (Oppenheim et al., 1999). Besides the static feature, a dynamic feature that is composed of delta coefficients is calculated to incorporate temporal information (Young et al., 2000). Specifically, a vector of delta coefficients z_D at time frame m is calculated as,

$$z_D(m) = \sum_{w=1}^{W} w(z(m+w) - z(m-w))/2 \sum_{w=1}^{W} w^2. \quad (2)$$

Here, z refers to static GFCCs; w is a neighboring window index; W denotes the half-window length and it is set to 2.

3. Segmentation

With the T–F representation of a cochleagram, our goal here is to determine how the T–F units are segregated into streams so that the units in a stream correspond primarily to one source and the units that belong to different streams correspond to different sources. One could perform this separation by directly grouping individual T–F units. However, a local unit is likely too small for robust global grouping. On the other hand, one could utilize local information to combine neighboring units into segments that allow for the use of more global information such as spectral envelope that is missing from an individual T–F unit. The segment infor-

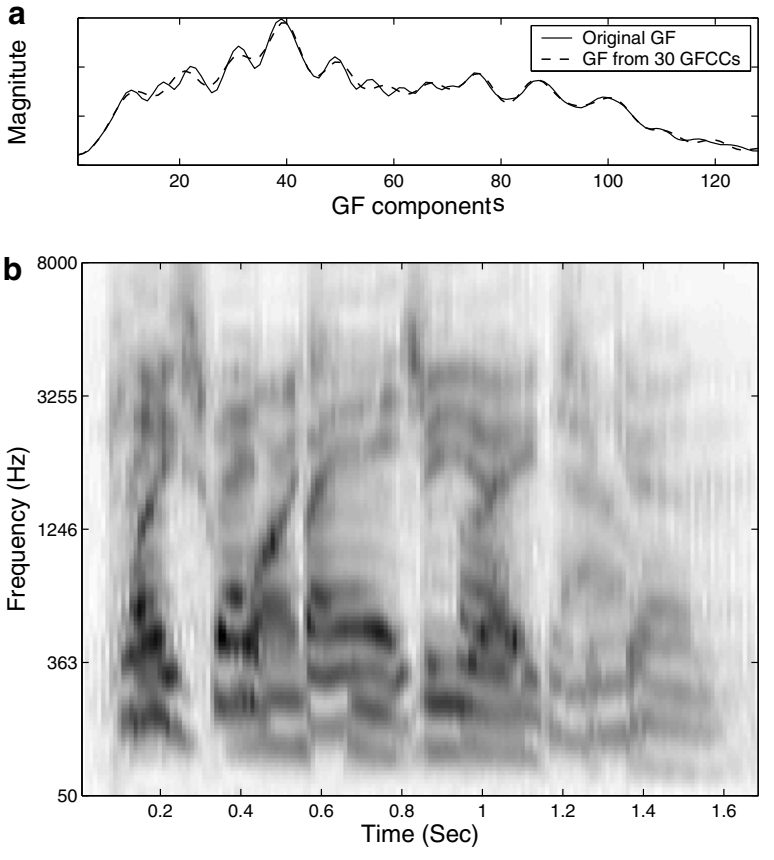

Fig. 3. Illustrations of energy compaction by GFCCs. Plot (a) shows a GF frame at time 1 s in Fig. 2. The original GF is plotted as the solid line and the resynthesized GF by 30 GFCCs is plotted as the dashed line. Plot (b) presents the resynthesized cochleagram from Fig. 2 using 30 GFCCs.

mation provides a robust foundation for grouping and achieves better SNR improvement of segregated speech than T–F unit information (Hu and Wang, 2004).

3.1. Voiced segmentation

To perform segmentation, an input mixture is first passed through the Gammatone filterbank in (1). Then, the response of each filter is transduced by the Meddis model of inner hair cells (Meddis, 1988), whose output represents the firing rate of an auditory nerve fiber. We use $h(c,n)$ to denote the hair cell output of filter channel c at time step n. In each filter channel, the output is further divided into 20 ms time frames with a 10 ms frame shift. To capture periodicity of the signal, we construct a correlogram, which is a periodicity representation that consists of autocorrelations of filter responses across all the filter channels (Wang and Brown, 2006). For a T–F unit of channel c in time frame m, its autocorrelation function of the hair cell response is

$$A(c,m,\tau) = \frac{1}{2T}\sum_{n=0}^{2T-1} h(c,mT-n)h(c,mT-n-\tau). \quad (3)$$

Here, τ refers to time lag, and the maximum lag corresponds to 50 Hz. T is the number of samples in the 10 ms time shift; a time window lasts 20 ms and contains $2T$ samples. Exact values of these three parameters can be easily derived when input sampling frequency is known. $h(c,n)$ is zero-padded at the beginning for the calculation of the first frame of A. In addition, envelope correlogram is calculated by replacing the hair cell output in (3) with envelope of the hair cell output, which is obtained using low-pass filtering (Hu and Wang, 2006).

Please cite this article in press as: Shao, Y. et al., A computational auditory scene analysis system for speech ..., Computer Speech and Language (2008), doi:10.1016/j.csl.2008.03.004

The periodic nature of voiced speech provides useful cues for segmentation by CASA systems. For example, a harmonic usually activates a number of adjacent auditory channels because the pass-bands of adjacent filters have significant overlap, which leads to a high cross-channel correlation. For the same T–F unit, the cross-channel correlation is calculated as

$$C(c,m) = \frac{1}{L}\sum_{\tau=0}^{L-1} \widehat{A}(c,m,\tau)\widehat{A}(c+1,m,\tau+1). \quad (4)$$

$\widehat{A}(c,m,\tau)$ denotes normalized $A(c,m,\tau)$ with zero mean and unity variance. L refers to the aforementioned maximum delay in sampling steps.

Additionally, the periodic signal usually lasts for some time, within which it has good temporal continuity. Hence, we perform segmentation of voiced portions by merging T–F units using cross-channel correlation and temporal continuity (Hu and Wang, 2006). Specifically, neighboring T–F units in both time and frequency are merged to form segments in the low frequency range (below 1 kHz) if the cross-channel correlation of the corresponding correlogram responses is higher than the threshold of 0.985 (Hu and Wang, 2006). In the high-frequency range (above 1 kHz), where a Gammatone filter responds to multiple harmonics, we merge the T–F units on the basis of cross-channel correlation of the envelope correlogram. The same threshold is used as in the low-frequency range. Fig. 4 illustrates voiced segments obtained from the signal in Fig. 2.

3.2. Unvoiced segmentation

In a speech utterance, unvoiced speech constitutes a smaller portion of overall energy than voiced speech but it contains important phonetic information for speech recognition. Unvoiced speech lacks harmonic structure, and as a result is more difficult to segment. Here we employ an onset/offset based segmentation method proposed by Hu and Wang (2007). Onsets and offsets correspond to sudden intensity changes, reflecting boundaries of auditory events. They are usually derived from peaks and valleys of the time derivative of the intensity. However, these peaks and valley do not always correspond to real onsets and offsets. Hence, the intensity is typically smoothed over time to reduce noisy fluctuations. The intensity is also smoothed over frequency to enhance the alignment of onsets and offsets. The degree of smoothing is called the scale. The larger the scale is, the smoother the intensity becomes, and vice versa.

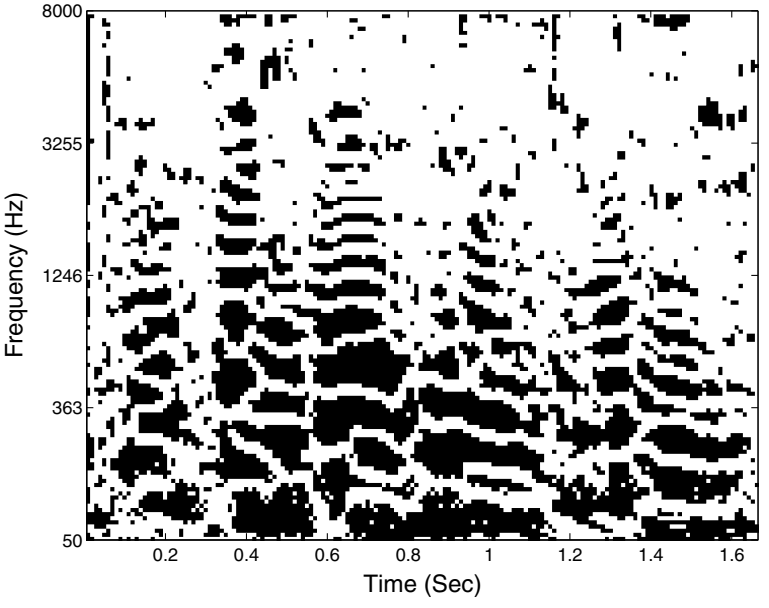

Fig. 4. Estimated voiced segments of the two-talker mixture in Fig. 2. White color shows the background. The dark regions represent the voiced segments.

Please cite this article in press as: Shao, Y. et al., A computational auditory scene analysis system for speech ..., Computer Speech and Language (2008), doi:10.1016/j.csl.2008.03.004

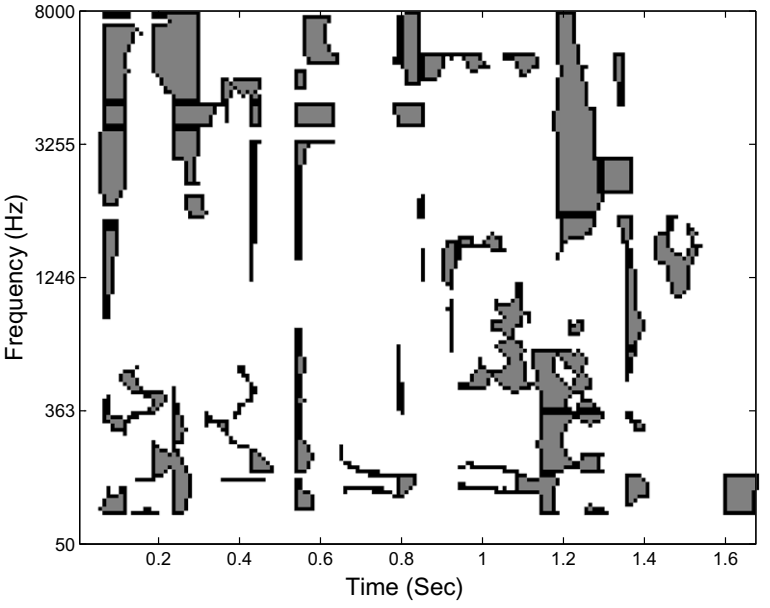

Fig. 5. Extracted unvoiced segments of the two-talker mixture shown in Fig. 2. White color shows the background. Gray regions represent unvoiced segments bounded by black lines.

Specifically, this segmentation method has three steps: Smoothing, onset/offset detection, and multi-scale integration. An input mixture is first passed through the Gammatone filterbank. Then, the output from each filter channel is half-wave rectified, low-pass filtered and down-sampled to extract temporal envelope (Hu and Wang, 2007). In the first processing stage, the system smoothes the temporal envelope by convolving it with a Gaussian function. This function has zero mean and its standard deviation determines the smoothing scale (Weickert, 1997; Hu and Wang, 2007). In the second stage, the system detects onsets and offsets in each filter channel and then merges simultaneous onsets and offsets from adjacent channels into onset and offset fronts, which are vertical contours connecting onset and offset candidates across frequency. Segments are formed by matching individual onset and offset fronts. The smoothing operation may blur event onsets and offsets of small T–F regions at a coarse scale, resulting in misses of some true onsets and offset. On the other hand, the detection process may be sensitive to insignificant intensity fluctuations within events at a small scale. Thus, a cochleagram may be under-segmented or over-segmented at a fixed scale. Hence, it is difficult to produce satisfactory segmentation on a single scale. Therefore, a final set of segments are identified by integrating over four different scales (see Hu and Wang (2007) for details).

Since onsets and offsets correspond to sudden intensity increases and decreases which could be introduced by either voiced speech or unvoiced speech, segments obtained by the onset/offset analysis usually contain both speech types. Additionally, unlike natural conversations, the case of blind mixing of speakers in the SSC task leads to blurring and merging of onset–offset fronts. Thus, matching onset and offset fronts creates some segments that are not speaker homogeneous, i.e. the same segment contains T–F units from more than one source. Here, we extract unvoiced segments from onset/offset segments by removing those portions that are overlapped with the voiced segments. Specifically, if a T–F unit within an onset/offset segment is also included in a voiced segment, it is regarded as voiced speech and removed from the former. The remaining T–F units are regarded as unvoiced speech, and contiguous regions of such units are deemed unvoiced segments. Fig. 5 illustrates unvoiced segments obtained from the signal in Fig. 2.

4. Grouping

Voiced and unvoiced segments generated at the segmentation stage are separated in both frequency and time. The goal of the grouping process is to assign these segments into two streams, corresponding to two

Please cite this article in press as: Shao, Y. et al., A computational auditory scene analysis system for speech ..., Computer Speech and Language (2008), doi:10.1016/j.csl.2008.03.004

underlying sources in an input mixture. Since voiced speech exhibits consistent periodicity across frequency channels at a particular time, we first group voiced segments across frequency, namely simultaneous grouping. Grouping outputs at this stage are simultaneous streams, each of which is composed of voiced segments that should be mainly produced by one source (speaker). We do not perform simultaneous grouping of unvoiced segments because there is no reliable grouping cue for under multi-talker conditions. Since simultaneous streams from the same speaker are still separated, they are further grouped across time to form speaker homogeneous streams by a sequential grouping process. Unvoiced segments are then merged with speaker streams to generate final outputs.

4.1. Simultaneous grouping

Since pitch is a useful cue for simultaneous grouping (Bregman, 1990; Wang and Brown, 1999), we employ the pitch-based speech segregation method of Hu and Wang (2006) for simultaneous grouping. Outputs of this method are pitch contours and their corresponding simultaneous streams.

For an input mixture, an initial estimate of pitch contours for up to two speakers is produced for the entire mixture based on the aforementioned correlogram. Then, T–F units are labeled according to their consistency of periodicity with the pitch estimates. For each estimated pitch contour, we check the consistency of T–F units within the corresponding time frames against the specific estimates and conduct unit labeling. For low-frequency channels where harmonics are resolved, if a unit shows a similar response at an estimated pitch period, the corresponding T–F unit is labeled consistent with the pitch estimate; it is labeled inconsistent otherwise. Specifically, for a T–F unit at channel c and frame m, the similarity is determined by comparing a correlogram ratio, $A(c,m,\tau_S)/A(c,m,\tau_M)$, with a predefined threshold (Hu and Wang, 2006). τ_S indicates the lag period corresponding to the pitch estimate and τ_M denotes the lag that gives the maximum value of A over a plausible pitch range of 80–500 Hz. Since a high-frequency channel responds to multiple harmonics, its response is amplitude-modulated and the response envelope fluctuates at the pitch frequency (Wang, 2006). Thus, to determine whether a high-frequency T–F unit is pitch-consistent, the same similarity criterion is applied on the envelope correlogram (Hu and Wang, 2006). Subsequently, a voiced segment is grouped into a simultaneous stream if more than half of its units are labeled consistent with the pitch estimate. Since voiced segmentation is conducted based on cross-channel correlation and pitch labeling is based on the periodicity criterion, there are some T–F units labeled as pitch-consistent but have not been included in voiced segments. Thus, a simultaneous stream is further expanded by absorbing those units that are immediate neighbors to the stream and that are consistent with pitch estimate.

We regard a set of grouped voiced segments as a simultaneous stream, represented by a binary mask. If T–F units of the stream are consistent with a pitch estimate, the corresponding units in the mask are labeled as foreground (target-dominant or 1) and others as background (interference-dominant or 0). Thus, simultaneous streams are represented in the form of binary T–F masks. These masks correspond to individual streams in the mixture (Hu and Wang, 2006). Note that for each estimated pitch contour, there is a corresponding simultaneous stream. Fig. 6 shows a collection of such streams obtained from the signal in Fig. 2. The background is shown in white, and the different gray regions represent different simultaneous streams. It is evident that some voiced segments have been dropped from Fig. 4. Some of these segments exhibit strong cross-channel correlation and local periodicity, but they are actually not voiced speech and do not agree with any pitch estimates. Others exhibit distorted periodicity because of the mixing of two speakers and they are not consistent with either detected pitch estimate. As observed from the figure, compared with those voiced segments in Fig. 4, simultaneous streams have grown bigger and absorbed more T–F units at high-frequency channels than low-frequency channels.

4.2. Sequential grouping

Simultaneous streams from the same speaker are still separated in time and interleaved with speech from other speakers when multiple speakers are present. Thus, a CASA system requires a sequential grouping process to further organize simultaneous streams into streams that correspond to underlying speakers in an input mixture. For this purpose, we adapt the sequential organization algorithm by Shao and Wang (2006). This

Please cite this article in press as: Shao, Y. et al., A computational auditory scene analysis system for speech ..., Computer Speech and Language (2008), doi:10.1016/j.csl.2008.03.004

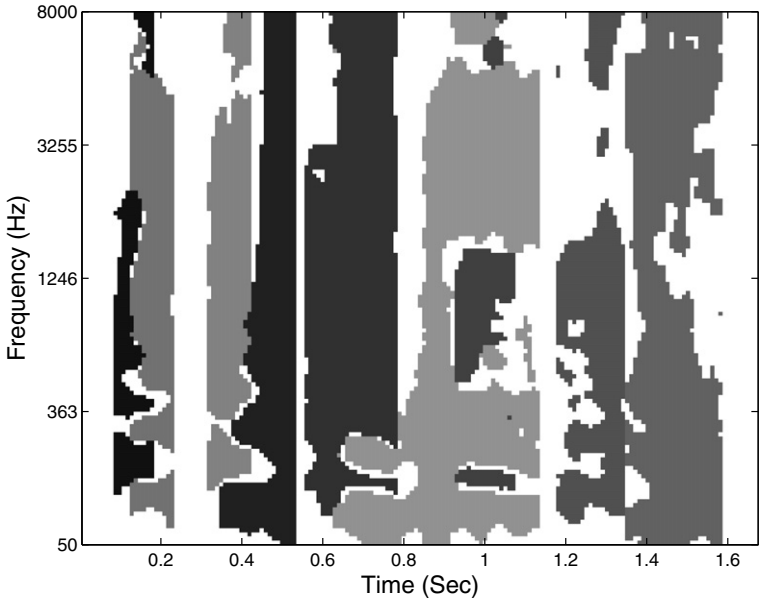

Fig. 6. Simultaneous streams estimated from the two-talker mixture in Fig. 2. White color shows the background. Different gray-colored regions indicate that the speech from the same speaker has been grouped across frequency but is still separated in time.

algorithm performs sequential organization based on speaker characteristics, which are modeled as text-independent models in a typical speaker recognition system (Furui, 2001). The algorithm searches for the optimal stream assignment by maximizing a posterior probability of an assignment given simultaneous streams. As a by-product, it also detects underlying speakers from the input mixture. Specifically, for each possible pair of speakers, it searches for the best assignment using speaker identification (SID) scores of a stream belonging to a speaker model. Finally, the optimal stream assignment is chosen by the speaker pair that yields the highest aggregated SID score (Shao and Wang, 2006).

The goal of SID is to find the speaker model that maximizes the posterior probability for an observation sequence O given a set of K speakers $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_K\}$ (Furui, 2001) Assuming an input signal comprises two sources (two-talker) from the speaker set, we establish our computational goal of sequential grouping as,

$$\hat{g}, \hat{\lambda}_{\mathrm{I}}, \hat{\lambda}_{\mathrm{II}} = \underset{\lambda_{\mathrm{I}}, \lambda_{\mathrm{II}} \in \Lambda, g \in G}{\arg\max}\, P(\lambda_{\mathrm{I}}, \lambda_{\mathrm{II}}, g|S) \tag{5}$$

$S = \{S_1, S_2, \ldots, S_i, \ldots, S_N\}$ is the set of N simultaneous streams S_i derived from preceding CASA processing. g is a stream labeling vector and its components take binary values of 0 or 1, referring to up to two speaker streams. G is the assignment space. By combining g and a stream set S, we know how the individual streams are assigned to two speaker streams. Note that g does not represent speaker identities but only denotes that the streams marked with the same label are from the same speaker. Thus, our objective of sequential grouping may be stated as finding a stream assignment $\hat{g}$ that maximizes the posterior probability. As a by-product, the underlying speaker identities are also detected.

The posterior probability in (5) can be rewritten as

$$P(\lambda_{\mathrm{I}}, \lambda_{\mathrm{II}}, g|S) = \frac{P(\lambda_{\mathrm{I}}, \lambda_{\mathrm{II}}, g, S)}{P(S)} = P(S|g, \lambda_{\mathrm{I}}, \lambda_{\mathrm{II}}) P(g|\lambda_{\mathrm{I}}, \lambda_{\mathrm{II}}) \frac{P(\lambda_{\mathrm{I}}, \lambda_{\mathrm{II}})}{P(S)} \tag{6}$$

Since a two-talker mixture may be blindly mixed, the assignment is independent of specific models. Thus, $P(g|\lambda_{\mathrm{I}}, \lambda_{\mathrm{II}})$ becomes $P(g)$ which may depend on the SNR level of a mixture because an extremely high or low SNR might lead to a bias of either 0 or 1 in g. Without prior knowledge, we assume it to be uniformly distributed.

Assuming independence of speaker models and applying the same assumption from speaker identification studies that prior probabilities of speaker models are the same, we insert Eq. (6) into (5) and remove the con-

stant terms. The objective then becomes finding an assignment and two speakers that have the maximum probability of assigned simultaneous streams given the corresponding speaker models as follows.

$$\hat{g}, \hat{\lambda}_{\mathrm{I}}, \hat{\lambda}_{\mathrm{II}} = \underset{\lambda_{\mathrm{I}}, \lambda_{\mathrm{II}} \in \Lambda, g \in G}{\arg\max}\ P(S|g, \lambda_{\mathrm{I}}, \lambda_{\mathrm{II}}) \tag{7}$$

The conditional probability on the right-hand-side of the equation is essentially a joint SID score of assigned streams.

Given a labeling g, we denote S^0 as the subset of streams labeled 0, and S^1 the subset labeled 1. Since S^0 and S^1 are complementary, the probability term in (7) can be written as follows,

$$P(S|g, \lambda_{\mathrm{I}}, \lambda_{\mathrm{II}}) = P(S^0, S^1|\lambda_{\mathrm{I}}, \lambda_{\mathrm{II}}) \tag{8}$$

Here, the g term is dropped for simplification because the two subsets already incorporate the labeling information.

Assuming that any two streams, S_i and S_j, are independent of each other given the models and that streams with different labels are produced by different speakers, the conditional probability in (8) can be written as

$$P(S^0, S^1|\lambda_{\mathrm{I}}, \lambda_{\mathrm{II}}) = P(S^0|\lambda_{\mathrm{I}}, \lambda_{\mathrm{II}})P(S^1|\lambda_{\mathrm{I}}, \lambda_{\mathrm{II}}) = \prod_{S_i \in S^0} P(S_i|\lambda_{\mathrm{I}}) \prod_{S_j \in S^1} P(S_j|\lambda_{\mathrm{II}}) \tag{9}$$

Thus, the goal of sequential grouping leads to a search of the speaker and the label space for a specific label (grouping) and a pair of speakers that maximizes the joint probability of assigned streams given a speaker pair. This search requires evaluations of SID scores of a stream given a speaker model. However, as defined by a binary T–F mask, frequency components of a time frame within a stream are not all clean, which leads to a missing data problem for likelihood calculation (Cooke et al., 2001). Recently, we find that incorporating GFCCs with uncertainty decoding (Srinivasan and Wang, 2007) yields substantially better SID performance than state-of-the-art robust features and alternative missing-data methods (Shao et al., 2007). This approach enhances corrupted features by reconstructing missing T–F units from a speech prior model, and errors from reconstruction are also incorporated in the likelihood of the enhanced features. Hence, we calculate the likelihood score of a stream using this approach.

Studies such as Lovekin et al. (2001) and Shao and Wang (2006) have shown that voiced speech plays a dominant role in sequential grouping and speaker recognition. Therefore, we first apply the above sequential grouping algorithm to organize simultaneous streams, which are mostly voiced speech. Outputs of the algorithm are two binary masks (streams) and corresponding speaker identities. Fig. 7a presents two estimated speaker streams in different gray colors after simultaneous streams are grouped.

Lovekin et al. (2001) have also shown that unvoiced speech contributes positively to speaker recognition but it is not as discriminative as voiced speech. Here, unvoiced segments are grouped with the two streams using the same sequential grouping algorithm. However, since likelihoods of unvoiced speech are not as discriminative, the grouping algorithm employs the already detected speaker pair associated with the organized streams instead of searching for it. In addition, we find that unvoiced segments are typically much smaller than simultaneous streams, resulting in poor likelihood estimation by uncertainty decoding. Therefore, likelihoods are calculated using a missing data method of marginalization (Cooke et al., 2001; Shao and Wang, 2006). Simply put, instead of reconstructing missing T–F units, this method ignores them in the likelihood calculation. Fig. 7b presents two speaker streams after unvoiced segments are also grouped.

We find that the above processes do not capture all the speech segments. To further refine organized streams, we apply a watershed algorithm (Vincent and Soille, 1991) to the cochleagram of the mixture and extract segments that may have been missed by the preceding processes. Given an intensity map such as a cochleagram, the watershed technique identifies segment regions surrounding local minima, and it has been widely used for image segmentation. Additionally, watershed segments that contain less than eight T–F units are removed from further processing. Subsequently, a watershed segment is first absorbed by the two speaker streams if it is largely (greater than or equal to two-thirds) overlapped with either of them, since a speaker's voice tends to occupy contiguous time frames and activate neighboring filters due to large overlaps between neighboring Gammatone bandwidths. Then, if a segment has not been merged with either stream, i.e. the overlapped region is less than two-thirds of its total area, those overlapped T–F units are removed from

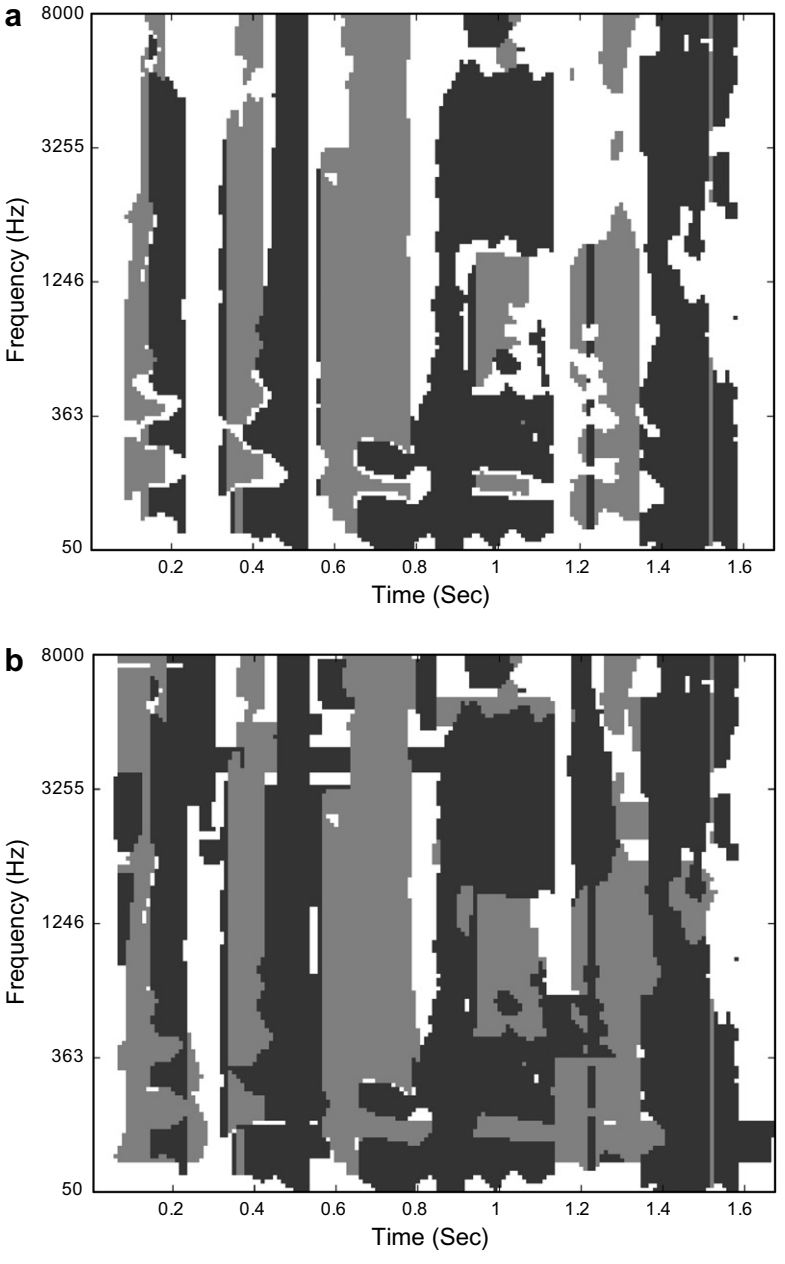

Fig. 7. Estimated speaker streams. Plot (a) is obtained by sequential grouping of simultaneous streams. Plot (b) shows streams after unvoiced segments are also grouped. White color shows the background. The two gray-colored regions represent two separated speaker streams.

the segment. Finally, watershed segments are sequentially grouped to the organized streams just like unvoiced segments.

5. Uncertainty decoding

After processing a two-talker mixture by our CASA system, we have obtained two binary T–F masks. How do we use these masks for speech recognition? One solution is to use a missing data speech recognizer of Cooke et al. (2001). This recognizer treats noise-dominant T–F units as missing or unreliable and marginalizes them

Please cite this article in press as: Shao, Y. et al., A computational auditory scene analysis system for speech ..., Computer Speech and Language (2008), doi:10.1016/j.csl.2008.03.004

during recognition. However, this constrains the recognition to be performed in the T–F domain instead of the cepstral domain. Additionally, mask estimation errors degrade the recognition performance. Estimation of these errors would enable their use in an uncertainty decoder (Deng et al., 2005) for improved recognition results. Hence, we propose the use of our novel auditory cepstral feature, GFCC, in conjunction with the uncertainty decoder for speech recognition. Specifically, GFs are reconstructed from the estimated binary masks by utilizing the statistical information contained in a speech prior. At the same time reconstruction uncertainties are also estimated. These GFs and uncertainties are transformed into the GFCC domain as described in Section 2. Finally the resulting GFCCs and the uncertainties are fed into an uncertainty decoder for recognition.

Given a binary T–F mask, a noisy spectral vector Y at a particular time frame is partitioned into reliable and unreliable constituents as Y_r and Y_u, where $Y = Y_r \cup Y_u$. The reliable features are the T–F units labeled 1 (target-dominant) in the binary mask while the unreliable features are the ones labeled 0 (interference-dominant). Assuming that the reliable features Y_r approximate well the true ones X_r, a Bayesian decision is then employed to estimate the remaining features X_u given the reliable ones and a prior speech model. As proposed by Raj et al. (2004) and Srinivasan and Wang (2007), we model the speech prior as a mixture of Gaussian densities, widely known as Gaussian mixture model (GMM) (Huang et al., 2001).

$$p(X) = \sum_{k=1}^{M} p(k)p(X|k), \tag{10}$$

where $M = 2048$ is the number of Gaussians in the mixture, k indexes Gaussian component, $p(k)$ refers to component weight, and $p(X|k) = N(X; \mu_k, \Sigma_k)$. The binary mask is also used to partition the mean and covariance of each Gaussian into their reliable and unreliable parts as:

$$\mu_k = \begin{bmatrix} \mu_{r,k} \\ \mu_{u,k} \end{bmatrix}, \quad \Sigma_k = \begin{bmatrix} \Sigma_{rr,k} & \Sigma_{ru,k} \\ \Sigma_{ur,k} & \Sigma_{uu,k} \end{bmatrix}. \tag{11}$$

Note that $\Sigma_{ru,k}$ and $\Sigma_{ur,k}$ denote the cross-covariance between the reliable and unreliable parts.

We first estimate the unreliable features given the reliable ones as

$$E_{X_u|X_r}(X_u) = \sum_{k=1}^{M} p(k|X_r)\hat{X}_{u,k}, \tag{12}$$

where $p(k|X_r)$ is the*a posteriori*probability of the k'th component given the reliable data and $\widehat{X}_{u,k}$ is the expected value of X_u given the k'th Gaussian function. $p(k|X_r)$ is estimated using the Bayesian rule and the marginal distribution $p(X_r|k) = N(X_r; \mu_{r,k}, \Sigma_{rr,k})$ (Srinivasan and Wang, 2007). The expected value in the unreliable T–F units corresponding to the k'th Gaussian is computed as

$$\widehat{X}_{u,k} = \mu_{u,k} + \Sigma_{ur,k}\Sigma_{rr,k}^{-1}(X_r - \mu_{r,k}). \tag{13}$$

Besides estimating unreliable T–F units, we are also interested in computing the uncertainty in our estimates. The variance associated with the reconstructed vector $\widehat{X}$ can also be computed in a similar fashion to the computation of the mean in (12) as:

$$\widehat{\Sigma}_{\widehat{X}} = \sum_{k=1}^{M} p(k|X_r)\left\{\left(\begin{bmatrix} X_r \\ \widehat{X}_{u,k} \end{bmatrix} - \mu_k\right) \times \left(\begin{bmatrix} X_r \\ \widehat{X}_{u,k} \end{bmatrix} - \mu_k\right)^T + \begin{bmatrix} 0 & 0 \\ 0 & \widehat{\Sigma}_{u,k} \end{bmatrix}\right\}, \tag{14}$$

where

$$\widehat{\Sigma}_{u,k} = \Sigma_{uu,k} - \Sigma_{ur,k}\Sigma_{rr,k}^{-1}\Sigma_{ru,k}. \tag{15}$$

The observation density in each state of a hidden Markov model (HMM) based ASR system is usually modeled as a GMM. Therefore,

$$p(z|k,q) = N(z; \mu_{k,q}, \Sigma_{k,q}) \tag{16}$$

is the likelihood of observing a speech frame z given state q and mixture component k; $\mu_{k,q}$ and $\Sigma_{k,q}$ are the mean and the variance of the Gaussian mixture component. When noisy speech is processed by unbiased

speech enhancement algorithms, it is shown by Deng et al. (2005) that the observation likelihood should be computed as

$$\int_{-\infty}^{\infty} p(z|k,q)p(\hat{z}|z)dz = N(\hat{z}; \mu_{k,q}, \Sigma_{k,q} + \Sigma_{\hat{z}}). \quad (17)$$

It can be seen in (17) that the uncertainty decoder increases the variance of individual Gaussian components to account for feature reconstruction errors (Deng et al., 2005; Srinivasan and Wang, 2007). Here, the reconstructed GF feature $\widehat{X}$ and its uncertainty $\widehat{\Sigma}_{\widehat{X}}$ are transformed into the GFCC domain as $\hat{z}$ and $\widehat{\Sigma}_{\hat{z}}$ and then used in the uncertainty decoder. $\widehat{\Sigma}_{\hat{z}}$ is the estimate of $\Sigma_{\hat{z}}$ in (17).

6. Experimental results

We evaluate our system using the SSC task (Cooke and Lee, 2006). This task aims to recognize speech from a target talker in the presence of another competing speaker (two-talker) or speech-shaped noise (SSN).

To build speaker models, we utilize the GFCC feature as described in Section 2. Each of the 34 speaker models in the SSC task comprises a mixture of 64 Gaussians. The speech prior model is trained on GF features and is a mixture of 2048 Gaussian densities. This prior model and the binary masks are used in the cochleagram domain to reconstruct missing T–F units. The reconstructed GFs are then transformed into the GFCC domain using the DCT. For recognition, we form the 60-dimensional feature vector of GFCC_D, which includes delta coefficients. GF uncertainties are transformed into the cepstral domain since DCT is a linear transformation. Uncertainties of delta coefficients are also derived. Whole-word HMM-based speaker-independent ASR models are then trained on clean speech; each word model comprises 8 states and a 32-component Gaussian mixture with diagonal covariance in each state. The uncertainty decoder also uses diagonal covariance for uncertainties. During the recognition process, given estimated uncertainties and clean ASR models, the uncertainty decoder calculates the likelihood of reconstructed GFCC_D features and transcribes the speech.

Since our speech segregation does not rely on the content information in an utterance, the system does not know which separated stream contains "white" in the two-talker task. In order to select the target, we employ a normalized scoring method. We let our uncertainty decoder recognize both segregated streams using two different grammars (TW and NW):

TW : $command white $preposition $letter $number $adverb

and

NW : $command $non-white $preposition $letter $number $adverb,

where $non-white only has three choices of colors except white. A normalized score is calculated for each stream by subtracting the recognition likelihood score of NW from the one using grammar TW. The stream with a larger score is chosen as the target, i.e., stream 1 (s_1) is chosen as the target when

$$P_{\mathrm{TW}}(s_1) - P_{\mathrm{NW}}(s_1) > P_{\mathrm{TW}}(s_2) - P_{\mathrm{NW}}(s_2), \quad (18)$$

or stream 2 (s_2) if otherwise. This selection metric is actually the same as evaluating the joint likelihood score of one stream containing the keyword "white" while the other containing $non-white. Eq. (18) is the same as,

$$P_{\mathrm{TW}}(s_1) + P_{\mathrm{NW}}(s_2) > P_{\mathrm{NW}}(s_1) + P_{\mathrm{TW}}(s_2). \quad (19)$$

We first evaluate our proposed system on the two-talker task of SSC. On average, for a two-talker mixture of two-second long, the system produces transcription in 1–2 min on a Dell server with a dual Xeon 3.4 GHz processor and 4 GB memory. Speech segregation consumes approximately 90% of the total time and feature reconstruction and uncertainty decoding account for the rest. Evaluation results are summarized in Table 1. The performance is measured in terms of recognition accuracy of the relevant keywords at each TMR condition (Cooke and Lee, 2006). We report the results for the different gender (DG), the same gender (SG) and the same talker (ST) subcategories as well as the overall mean score (Avg.).

For comparison, we show the performance of our baseline system without segregation. The baseline system employs the same speaker-independent ASR models as the proposed system but uses a conventional MFCC

Table 1
Recognition accuracy (in %) of the baseline system and the stages of the proposed CASA system on the two-talker task

TMR(dB)/System		DG	SG	ST	Avg.
6	Baseline	66.00	65.92	66.52	66.17
	Voiced only	62.00	66.48	59.50	62.42
	Voiced + unvoiced	72.00	70.67	54.75	65.25
	Proposed overall	80.75	76.81	54.98	70.08
3	Baseline	51.25	49.44	51.58	50.83
	Voiced only	54.25	56.15	45.48	51.58
	Voiced + unvoiced	68.25	66.20	41.63	57.83
	Proposed overall	78.50	72.63	39.14	62.25
0	Baseline	36.00	34.64	32.58	34.33
	Voiced only	42.00	46.09	33.03	39.92
	Voiced + unvoiced	61.24	58.10	28.51	48.25
	Proposed overall	74.50	67.31	25.34	54.25
−3	Baseline	19.25	22.07	18.55	19.83
	Voiced only	36.00	32.96	23.98	23.67
	Voiced + unvoiced	52.25	46.92	20.18	39.08
	Proposed overall	63.50	53.07	20.59	44.58
−6	Baseline	9.50	10.34	9.50	9.75
	Voiced only	25.00	23.18	15.16	20.83
	Voiced + unvoiced	39.25	33.52	13.57	28.08
	Proposed overall	48.00	36.31	17.19	33.17
−9	Baseline	3.25	4.75	3.62	3.83
	Voiced only	17.75	16.20	11.99	15.17
	Voiced + unvoiced	30.00	24.02	11.76	21.50
	Proposed overall	32.00	22.34	11.99	21.75

DG, SG and ST refer to subconditions of "different gender", "same gender" and "same talker" respectively. Avg. is the mean accuracy.

feature instead of GFCC. This MFCC feature consists of 39 dimensions, including MFCCs, normalized energy, delta and acceleration coefficients (Huang et al., 2001). Speech recognition takes the grammar TW and is conducted using the HTK toolkit (Young et al., 2000). In addition, we include evaluation results for two intermediate stages of our system to illustrate the relative contributions of different stages. Specifically, we have evaluated speaker streams after sequential grouping of simultaneous streams. Since they are mainly voiced speech, their results are shown in the'Voiced only' rows in the table. Results that evaluate speaker streams after unvoiced segments are assigned (i.e. without watershed segments) are presented in the 'Voiced + unvoiced' rows.

On average, the proposed system improves significantly over the baseline system in terms of average accuracy across all TMR conditions. Except for the ST and 6 dB conditions, recognition accuracy is improved over the baseline after simultaneous streams are organized by sequential grouping. The performance is further improved after unvoiced segments and watershed segments are grouped. In general, large improvements are observed in the DG and the SG conditions. However, the proposed system does not perform well in the ST condition. This is primarily due to our use of speaker characteristics for sequential grouping. Note that for the ST condition, speaker characteristics are not distinctive for segregation. Fig. 8 compares the system performance with $(w/)$ and without (w/o) the ST condition. Note that baseline performance is nearly the same in these two conditions. Our CASA system on average achieves a further absolute improvement of over 11% when the ST condition is excluded.

Since our sequential grouping algorithm also identifies the underlying speakers, we also present the evaluation results of SID performance in Table 2. Under most of the TMR conditions, we achieve an accuracy of over 90% in recognizing the target speaker.

For the SSN task of SSC, since there is only one speaker in an input mixture and SSN is not voiced, we treat all the pitch contours and simultaneous streams as produced by the target speaker. Hence, we use a binary

Please cite this article in press as: Shao, Y. et al., A computational auditory scene analysis system for speech ..., Computer Speech and Language (2008), doi:10.1016/j.csl.2008.03.004

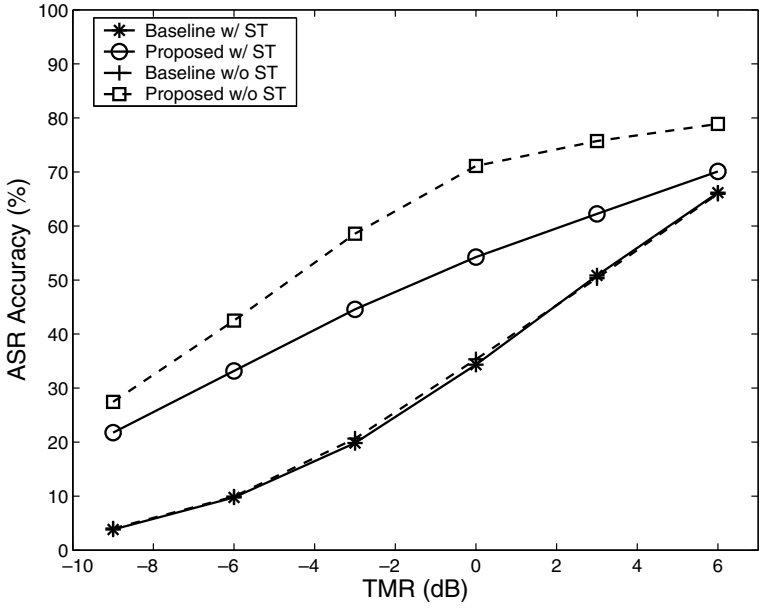

Fig. 8. Average recognition accuracy on the two-talker task. The solid star line represents the baseline recognition results. The dashed plus line shows the baseline performance without the same talker (ST) data. The results of the proposed CASA system are given as the solid circle line. Its accuracy without the ST condition is presented as the dashed square line.

Table 2
Speaker identification (SID) accuracy (in %) on the two-talker task

TMR(dB)	−9	−6	−3	0	3	6
Both SID	12.83	33.50	57.50	65.33	63.17	46.17
Target SID	57.17	89.50	98.17	99.5	99.83	99.33

"Both SID" shows the accuracies when both speakers in a mixture are identified correctly. "Target SID" presents the accuracies when the target speaker is identified as either of the SID outputs.

Table 3
Speech recognition accuracy (in %) on the SSN task using the proposed CASA system

SNR(dB)	−12	−6	0	6	Clean
Baseline system	13.00	12.50	16.22	29.50	93.94
Proposed system	18.78	33.39	57.94	75.40	–

For comparison, the baseline performance is also shown.

mask that aggregates all the simultaneous streams as the final output. In the two-talker task, GFCC_D features and associated uncertainties obtained from reconstructed GF features are fed to the uncertainty decoder for recognition. Table 3 presents the performance of our system in terms of recognition accuracy in percentage. Across all the SNR conditions, our CASA system shows a significant improvement over the baseline recognizer.

7. Conclusions and discussions

In this paper, we have presented a CASA system capable of segregating and recognizing the contents of a target utterance in the presence of another speech utterance or speech-shaped noise. The proposed CASA system is an end-to-end system that largely follows the ASA account of auditory organization and produces

Please cite this article in press as: Shao, Y. et al., A computational auditory scene analysis system for speech ..., Computer Speech and Language (2008), doi:10.1016/j.csl.2008.03.004

streams that correspond to different sound sources in a mixture. The contents of the target stream are then recognized using an uncertainty decoding framework. We have systematically evaluated our system on the speech separation task and obtained significant improvement over the baseline performance across all TMR/SNR conditions. For example, at 0 dB TMR two-talker condition, the absolute improvement in word accuracy is about 20%. Additionally, the accuracy of identification of both speakers is 65.33%, and of the target speaker is 99.5%.

The proposed system is primarily based on features such as periodicity and onset/offset. These properties are not specific to the target source to be segregated or even to speech sounds. In other words, the system does not use *a priori* knowledge of sound sources in the mixture, except in sequential grouping where the system knows that there are two speakers in each mixture and utilizes text-independent speaker models to represent speaker characteristics. In addition, our system does not depend on the nature and the size of the target vocabulary, the recognition task or even the language of the speech sources in the mixture. A resulting advantage is the generality of our system in terms of dealing with both speech and non-speech interferences. Furthermore, the ASA inspired architecture of the proposed system and adoption of the ideal binary mask as the computational goal make the system readily scalable to handle multiple interference sources.

For the sequential grouping process to tackle multi-talker conditions, one could extend the model-based algorithm by replacing the speaker pair with a speaker triplet, a speaker quadruplet, etc., in (5). This will lead to a computational objective similar to (7) after applying the derivation in Section 4.2. This extension still makes an explicit assumption of the speaker number in a mixture. To remove this assumption, one could further extend the formulation by including another search that evaluates different speaker numbers. Such direct extension, however, is not scalable to situations with a large number of speakers. Alternatively, one could view a mixture as that of target speech and background, and construct the latter as a generic model Shao (2007).

Similar to the generality of our CASA segregation, the uncertainty decoding framework does not require interference conditions to be known *a priori*. Hence, it is used in conjunction with our CASA system for robust recognition. Although the proposed uncertainty estimation approach provides promising results, other approaches for estimation of this uncertainty could also be explored. For example, it might be beneficial to directly estimate the uncertainties corresponding to the static and the delta coefficients. This would enable us to exploit the differences in the *a priori* accuracies of these coefficients (Srinivasan and Wang, 2007).

Finally, Hu (2006) has shown that segmentation and simultaneous grouping achieve consistent SNR improvement for various intrusion types, including non-stationary noise and music. Furthermore, Shao (2007) has demonstrated the effectiveness of the sequential grouping algorithm under speech, stationary and non-stationary noise conditions. In addition, Srinivasan and Wang (2007) have also shown that the uncertainty decoder improves recognition accuracy in the presence of different noise types and for larger vocabularies. Therefore, our system is expected to generalize reasonably to other noise conditions that are not included in the SSC task.

Acknowledgements

This research was supported in part by an AFOSR Grant (FA9550-04-1-0117), an AFRL Grant (FA8750-04-1-0093) and an NSF Grant (IIS-0534707). We are grateful to Guoning Hu for discussion and much assistance. We acknowledge the SLATE Lab (E. Fosler-Lussier) for providing computing resources. A preliminary version of this work was presented in 2006 Interspeech.

References

Allen, J.B., 2005. Articulation and Intelligibility. Morgan & Claypool, San Rafael, CA.

Bregman, A.S., 1990. Auditory Scene Analysis. The MIT Press, Cambridge, MA.

Cooke, M., Lee, T., 2006. Speech separation and recognition competition. Available from: <http://www.dcs.shef.ac.uk/martin/SpeechSeparationChallenge.htm>.

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

Deng, L., Droppo, J., Acero, A., 2005. Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion. IEEE Trans. Speech Audio Process. 13, 412–421.

Deoras, A.N., Hasegawa-Johnson, M., 2004. A factorial HMM approach to simultaneous recognition of isolated digits spoken by multiple talkers on one audio channel. In: Proceedings of ICASSP'04, vol. 1. pp. 861–864.
Ephraim, Y., 1992. A Bayesian estimation approach for speech enhancement using hidden Markov models. IEEE Trans. Signal Process. 40 (4), 725–735.
Furui, S., 2001. Digital Speech Processing, Synthesis, and Recognition. Marcel Dekker, New York.
Gales, M.J.F., Young, S.J., 1996. Robust continuous speech recognition using parallel model combination. IEEE Trans. Speech Audio Process. 4, 352–359.
Hu, G., 2006. Monaural speech organization and segregation. Ph.D. Thesis, Biophysics Program, The Ohio State University.
Hu, G., Wang, D.L., 2004. Monaural speech segregation based on pitch tracking and amplitude modulation. IEEE Trans. Neural Networks 15, 1135–1150.
Hu, G., Wang, D.L., 2006. An auditory scene analysis approach to monaural speech segregation. In: Hansler, E., Schmidt, G. (Eds.), Topics in Acoustic Echo and Noise Control. Springer, Heidelberg, pp. 485–515.
Hu, G., Wang, D.L., 2007. Auditory segmentation based on onset and offset analysis. IEEE Trans. Audio Speech Language Process. 15, 396–405.
Huang, X., Acero, A., Hon, H., 2001. Spoken Language Processing. Prentice Hall PTR, Upper Saddle River, NJ.
Jang, G., Lee, T., 2003. A probabilistic approach to single channel blind signal separation. In: Becker, S., Thrun, S., Obermayer, K. (Eds.), Advances in Neural Information Processing Systems, vol. 15. MIT Press, Cambridge, MA, pp. 1173–1180.
Kristjansson, T., Attias, H., Hershey, J., 2004. Single microphone source separation using high resolution signal reconstruction. In: Proceedings of ICASSP'04, vol. 2. pp. 817–820.
Lovekin, J.M., Yantorno, R.E., Krishnamachari, K.R., Benincasa, D.S., Wenndt, S.J., 2001. Developing usable speech criteria for speaker identification technology. In: Proceedings of ICASSP'01, pp. 421–424.
Meddis, R., 1988. Simulation of auditory neural transduction: further studies. The Journal of the Acoustical Society of America 83, 1056–1063.
Moore, B.C.J., 2003. An Introduction to the Psychology of Hearing, fifth ed. Academic Press, San Diego, CA.
Oppenheim, A.V., Schafer, R.W., Buck, J.R., 1999. Discrete-time Signal Processing, second ed. Prentice-Hall, Inc., Upper Saddle River, NJ.
Patterson, R.D., Holdsworth, J., Allerhand, M., 1992. Auditory models as preprocessors for speech recognition. In: Schouten, M.E.H. (Ed.), The Auditory Processing of Speech: From Sounds to Words. Mouton de Gruyter, Berlin, Germany, pp. 67–83 (Chapter 1).
Raj, B., Seltzer, M.L., Stern, R.M., 2004. Reconstruction of missing features for robust speech recognition. Speech Commun. 43, 275–296.
Raj, B., Singh, R., Smaragdis, P., 2005. Recognizing speech from simultaneous speakers. In: Proceedings of Interspeech'05, pp. 3317–3320.
Roman, N., Wang, D.L., Brown, G.J., 2003. Speech segregation based on sound localization. J. Acoust. Soc. Am. 114, 2236–2252.
Roweis, S.T., 2005. Automatic speech processing by inference in generative models. In: Divenyi, P. (Ed.), Speech Separation by Humans and Machines. Kluwer Academic, Norwell, MA, pp. 97–134.
Shao, Y., 2007. Sequential organization in computational auditory scene analysis. Ph.D. Thesis, Computer Science and Engineering, The Ohio State University.
Shao, Y., Wang, D.L., 2006. Model-based sequential organization in cochannel speech. IEEE Trans. Audio Speech Language Process. 14, 289–298.
Shao, Y., Srinivasan, S., Wang, D.L., 2007. Incorporating auditory feature uncertainties in robust speaker identification. In: Proceedings of ICASSP'07, vol. IV, pp. 277–280.
Srinivasan, S., Wang, D.L., 2007. Transforming binary uncertainties for robust speech recognition. IEEE Trans. Audio, Speech Lang. Process. 15 (7), 2130–2140.
Srinivasan, S., Roman, N., Wang, D.L., 2006. Binary and ratio time–frequency masks for robust speech recognition. Speech Commun. 48, 1486–1501.
Varga, A.P., Moore, R.K., 1990. Hidden Markov model decomposition of speech and noise. In: Proceedings of ICASSP'90, pp. 845–848.
Vincent, L., Soille, P., 1991. Watersheds in digital spaces: an efficient algorithm based on immersion simulations. IEEE Trans. Pattern Anal. Mach. Intell. 13 (6), 583–598.
Wang, D.L., 2005. On ideal binary mask as the computational goal of auditory scene analysis. In: Divenyi, P. (Ed.), Speech Separation by Humans and Machines. Norwell, MA, pp. 181–197.
Wang, D.L., 2006. Feature-based speech segregation. In: Wang, D.L., Brown, G.J. (Eds.), Computational Auditory Scene Analysis: Principles, Algorithms, and Applications. Wiley-IEEE Press, Hoboken, NJ, pp. 81–114.
Wang, D.L., Brown, G.J., 1999. Separation of speech from interfering sounds based on oscillatory correlation. IEEE Trans. Neural Networks 10 (3), 684–697.
Wang, D.L., Brown, G.J. (Eds.), 2006. Computational Auditory Scene Analysis: Principles, Algorithms, and Applications. Wiley-IEEE Press, Hoboken, NJ.
Weickert, J., 1997. A review of nonlinear diffusion filtering. In: Romeny, B.H., Florack, L.J.K.a.M.V. (Eds.), Scale-space Theory in Computer Vision. Springer, Berlin, pp. 3–28.
Young, S., Kershaw, D., Odell, J., Valtchev, V., Woodland, P., 2000. The HTK Book (for HTK Version 3.0). Microsoft Corporation.

Please cite this article in press as: Shao, Y. et al., A computational auditory scene analysis system for speech ..., Computer Speech and Language (2008), doi:10.1016/j.csl.2008.03.004