Technical Report OSU-CISRC-8/07-TR62
Department of Computer Science and Engineering
The Ohio State University
Columbus, OH 43210-1277

Ftpsite: **ftp.cse.ohio-state.edu**
Login: **anonymous**
Directory: **pub/tech-report/2007**
File: **TR62.pdf**
Website: **http://www.cse.ohio-state.edu/research/techReport.shtml**

A Computational Auditory Scene Analysis System for Speech Segregation and Robust Speech Recognition

Yang Shao
Department of Computer Science and Engineering
The Ohio State University, Columbus, OH 43210, USA
shaoy@cse.ohio-state.edu

Soundararajan Srinivasan*
Biomedical Engineering Department
The Ohio State University, Columbus, OH 43210, USA
srinivasan.36@osu.edu

Zhaozhang Jin
Department of Computer Science and Engineering
The Ohio State University, Columbus, OH 43210, USA
jinzh@cse.ohio-state.edu

DeLiang Wang
Department of Computer Science and Engineering & Center for Cognitive Science
The Ohio State University, Columbus, OH 43210, USA
dwang@cse.ohio-state.edu

Correspondence should be directed to D.L. Wang: Department of Computer Science and Engineering, The Ohio State University, 2015 Neil Ave., Columbus, OH 43210-1277.
Phone: (614)292-6827 URL: http://www.cse.ohio-state.edu/~dwang/

——

* Present address: Research and Technology Center, Robert Bosch LLC, Pittsburgh, PA 15212, USA

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.

Index Terms – 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 recognizers (ASRs) are typically trained on clean speech and face the mismatch problem when 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 subjective 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 (Wang and Brown, 1999) decomposes the auditory scene into groups of contiguous time-frequency (T-F) units or segments, each of which originates from a single sound source. 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 the 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 a particular 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

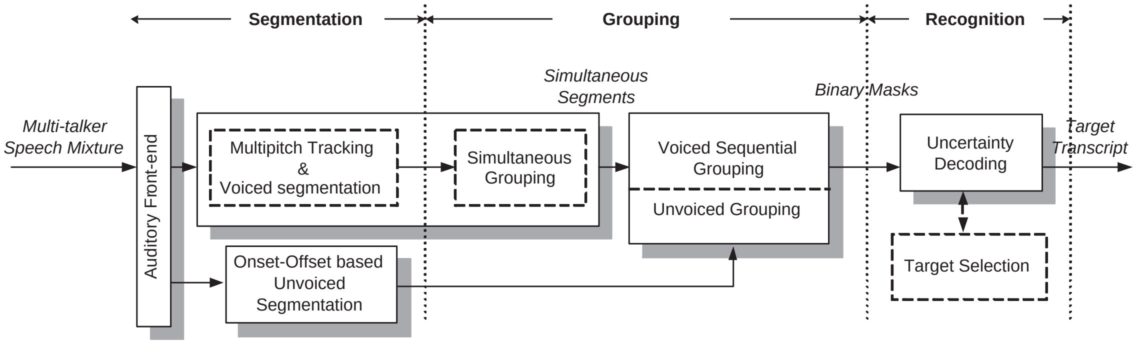


Figure 1: *Schematic diagram of the proposed two-stage CASA system and its application to ASR. The auditory front-end decomposes input signal into a T-F representation called a cochleagram (see Section 2). In the segmentation stage, our system generates both voiced segments and unvoiced segments. Then in the the grouping stage, the system first uses periodicity similarity to group voiced components by a simultaneous grouping process. Subsequently, a sequential grouping algorithm organizes the simultaneous segments and the 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.*

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 Figure 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 (Hu and Wang, 2006). In the simultaneous grouping stage, the system estimates pitch tracks of individual sources in the mixture and within each time frame, voiced components of individual sound sources are segregated based on periodicity similarity. This is followed by a sequential grouping stage that utilizes speaker characteristics to group the resulting simultaneous segments across time. Specifically, we first sequentially group the segregated voiced portions. Unvoiced segments are then grouped with the corresponding voiced "streams". The output of our CASA system are estimates of the ideal binary masks corresponding to the underlying sources. The masks are then used in an uncertainty decoding approach to robust speech recognition (Srinivasan and Wang, 2007). 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 (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.

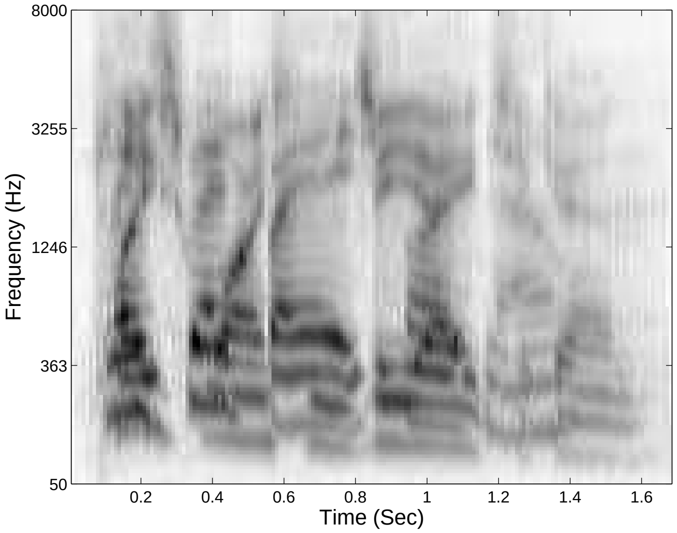


Figure 2: *Cochleagram of a two-talker utterance mixed at 0dB signal-to-noise ratio (SNR). Darker color indicates stronger energy within the corresponding time-frequency unit.*

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:

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

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. This yields a corresponding frame rate of 10 ms, 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, analogous to the widely used spectrogram. Please 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. Figure 2 shows a cochleagram of a two-talker mixture utterance. The signals from these two talkers are mixed at 0dB 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 et al., 2007).

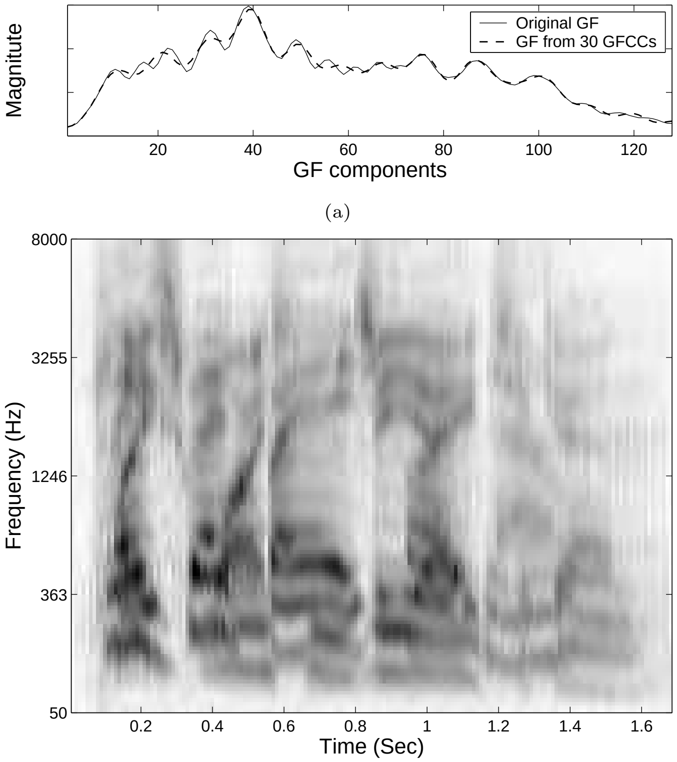
Rigorously speaking, the newly derived features are not cepstral coefficients because a cepstral analysis requires a log operation between the first and the second frequency analysis for the deconvolution purpose (Oppenheim et al., 1999). Here we regard these features as cepstral coefficients because of the functional similarities between the above transformation and that of a typical cepstral analysis. Figure 3 illustrates a GFCC transformed GF and a cochleagram. The top plot shows a comparison of a GF frame at 1 *sec* in Figure 2 and the resynthesized GF from its 30 GFCCs. The bottom plot presents the resynthesized cochleagram from Figure 2 using 30 GFCCs. As can be seen from Figure 2, the lowest 30-order GFCCs retain the majority information in a 128-dimensional GF. This is due to the "energy compaction" property of DCT (Oppenheim et al., 1999). Hence we use the 30-dimensional GFCCs as feature vectors in our study. Besides the static features, dynamic features such as delta coefficients are calculated to incorporate temporal information for ASR (Young et al., 2000).

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 are only produced by one source and the units that belong to different streams are produced by different sources. One could estimate this separation by directly grouping individual T-F units. However, a local unit is 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. This information could provide robust foundations for grouping.

3.1 Voiced segmentation

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. Additionally, the periodic signal usually lasts for some time, within which it has good temporal continuity. Thus, we perform segmentation of voiced portions by merging T-F units using cross-channel correlation and temporal continuity based on the Hu and Wang (2006) system. Specifically, neighboring T-F units with sufficiently high cross-channel correlation in a correlogram response are merged to form segments in the low frequency range. A correlogram



(a)

(b)

Figure 3: *Illustrations of energy compaction by GFCCs. Plot (a) shows a GF frame at time 1 sec in Figure 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 Figure 2 using 30 GFCCs.*

is a periodicity representation, consisting of autocorrelations of filter responses across all the filter channels (Wang and Brown, 1999). In the high frequency range, where a Gammatone filter responds to multiple harmonics, we merge the T-F units on the basis of cross-channel correlation of response envelopes. Along the time dimension, we employ the temporal continuity to merge neighboring units if they show high cross-channel correlations (Hu and Wang, 2006). Figure 4 illustrates voiced segments obtained from the signal in Figure 2.

3.2 Unvoiced segmentation

This section describes how our system generates unvoiced segments. 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 the harmonic structure, and as a result is more difficult to segment. Here we employ an onset/offset based segmentation method by Hu and Wang (2007). This method has three processing stages: Smoothing, onset/offset detection, and multiscale integration. In the first stage, the system smoothes the Gammatone filter responses using a Gaussian smoothing process. 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 defined as vertical contours connecting onset and offset

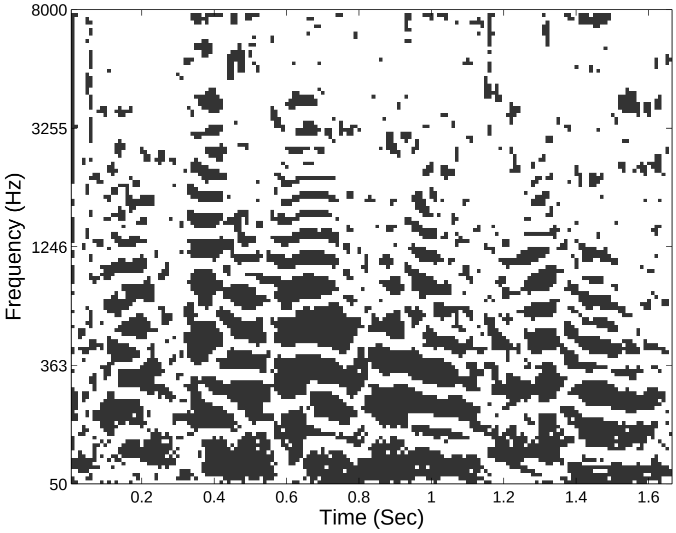


Figure 4: *Estimated voiced segments of the two-talker mixture in Figure 2. White color shows the background. The dark regions represent the voiced segments.*

candidates across frequency. Segments are generated 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 fine scale. Thus, a cochleagram may be under-segmented or over-segmented because of detection errors. Hence, it is difficult to produce satisfactory segmentation on a single scale. Therefore, a final set of segments are produced by integrating over 4 different scales (for further details see, Hu and Wang (2007)).

Since onsets and offsets correspond to sudden intensity increases and decreases which could be introduced by voiced speech or unvoiced speech, the obtained segments usually contain both speech types. Additionally, unlike natural conversations, the case of blind mixing of sources leads to blurring and merging of onset-offset fronts. Thus, matching onset and offset fronts creates segments that are not source homogeneous. Here, we extract the unvoiced segments from the onset/offset segments by removing those portions that are overlapped with the voiced segments.

4 Grouping

The voiced and unvoiced segments are separated in both frequency and time. The goal of the grouping process is to assign these segments into corresponding streams. Since voiced speech exhibits consistent periodicity across frequency channels at a particular time, we first group the voiced segments across frequency, namely simultaneous grouping. The resulting simultaneous segments are further grouped across time to form homogeneous streams in sequential grouping. The unvoiced segments are then merged with the streams to generate the final grouped outputs.

4.1 Simultaneous grouping

Since pitch is a useful cue for simultaneous grouping (Bregman, 1990; Wang and Brown, 1999), we employ the Hu and Wang (2006) system to estimate pitch periods and then use them to group the voiced segments across frequency. The outputs of the system are pitch contours and their corresponding simultaneous segments. These segments are represented in the form of binary T-F masks, with 1 labeling a T-F unit belonging to a segment and 0 labeling everything else. These masks are actually estimates of the ideal binary masks corresponding to individual sources in the mixture (Hu and Wang, 2006) but separated in time.

The system first produces an initial estimate of pitch contours for up to two sources for the entire utterance based on the aforementioned correlogram. Then, T-F units are labeled according to their consistency of periodicity with the pitch estimates. Specifically, for low-frequency channels where harmonics are resolved, if a unit shows similar response at an estimated pitch period, the corresponding T-F unit is labeled consistent with the pitch estimate; it is labeled inconsistent otherwise. For high-frequency channels that respond to several harmonics, an amplitude modulation model is used to determine whether a unit response shows beating at the pitch period and thus pitch-consistent (Hu and Wang, 2006). Subsequently, a voiced segment is grouped into a simultaneous segment if more than half of its units are labeled consistent with the pitch estimate. The segments are further expanded by absorbing neighboring units that have the same label.

We regard a set of grouped voiced segments as a simultaneous segment, represented by a binary mask with the T-F units labeled as foreground (target-dominant or 1) if they are consistent with the pitch estimates and others as background (interference-dominant or 0). Figure 5 shows a collection of simultaneously grouped segments obtained from the signal in Figure 2. The background is shown in white, and the different gray regions represent different simultaneous segments. Note that some voiced segments have been dropped from Figure 4 because they are not consistent with detected pitch. The simultaneous segments have expanded to include more high-frequency units than the voiced segments by absorbing the neighboring units with similar periodicity.

4.2 Sequential grouping

The simultaneous segments are still separated in time and inter-mingled with segments from other speakers when multiple speakers are present. Thus, a CASA system requires a sequential grouping process to organize these segments into corresponding streams. For this purpose, we adapt and employ the sequential organization algorithm by Shao and Wang (2006). This 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 segment assignment by maximizing the posterior probability of an assignment given the simultaneous segments. As a by-product, it also detects the two 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 segment belonging to a speaker model. Finally, the optimal segment assignment is chosen by

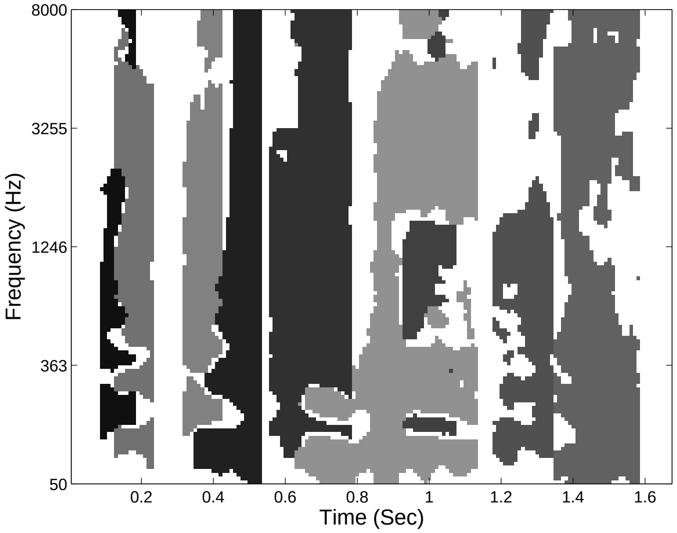


Figure 5: *Estimated simultaneous segments of the two-talker mixture in Figure 2. White color shows the background. Different gray-colored segments indicate that the voiced segments have been grouped across frequency but still separated in time.*

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,

$$\hat{g}, \hat{\lambda}_{\mathrm{I}}, \hat{\lambda}_{\mathrm{II}} = \underset{\lambda_{\mathrm{I}}, \lambda_{\mathrm{II}} \in \Lambda, g \in G}{\arg\max} P(\lambda_{\mathrm{I}}, \lambda_{\mathrm{II}}, g|S) \tag{2}$$

$S = \{S_1, S_2, \ldots, S_i, \ldots, S_N\}$ is the set of N simultaneous segments S_i derived from preceding CASA processing. g is a segment 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 a g and a segment set S, we know how the individual segments are assigned to two streams. Please note that g does not represent speaker identities but only denote that the segments marked with the same label are from the same speaker. Thus, our objective of sequential grouping may be stated as finding a segment assignment $\hat{g}$ that maximizes the posterior probability. As a side-product, the underlying speaker identities are also detected.

The posterior probability in (2) can be rewritten as

$$P(\lambda_{\mathrm{I}}, \lambda_{\mathrm{II}}, g|S) = \frac{P(\lambda_{\mathrm{I}}, \lambda_{\mathrm{II}}, g, S)}{P(S)} = P(S|g, \lambda_{\mathrm{I}}, \lambda_{\mathrm{II}})P(g|\lambda_{\mathrm{I}}, \lambda_{\mathrm{II}})\frac{P(\lambda_{\mathrm{I}}, \lambda_{\mathrm{II}})}{P(S)} \tag{3}$$

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 equation (3) into (2) and remove the constant terms. The objective then becomes finding an assignment and two speakers that have the maximum probability of assigned simultaneous segments given the corresponding speaker models as follows.

$$\hat{g}, \hat{\lambda}_{\mathrm{I}}, \hat{\lambda}_{\mathrm{II}} = \operatorname*{arg\,max}_{\lambda_{\mathrm{I}}, \lambda_{\mathrm{II}} \in \Lambda, g \in G} P(S|g, \lambda_{\mathrm{I}}, \lambda_{\mathrm{II}}) \tag{4}$$

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

Given a labeling g, we denote S^0 as the subset of segments labeled 0, and S^1 the subset labeled 1. Since S^0 and S^1 are complementary, the probability term in (4) can be written as follows,

$$P(S|g, \lambda_{\mathrm{I}}, \lambda_{\mathrm{II}}) = P(S^0, S^1|\lambda_{\mathrm{I}}, \lambda_{\mathrm{II}}) \tag{5}$$

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

Assuming that any two segments, S_i and S_j, are independent of each other given the models and that segments with different labels are produced by different speakers, the conditional probability in (5) can be written as

$$\begin{aligned} P(S^0, S^1|\lambda_{\mathrm{I}}, \lambda_{\mathrm{II}}) &= P(S^0|\lambda_{\mathrm{I}}, \lambda_{\mathrm{II}})P(S^1|\lambda_{\mathrm{I}}, \lambda_{\mathrm{II}}) \\ &= \prod_{S_i \in S^0} P(S_i|\lambda_{\mathrm{I}}) \prod_{S_j \in S^1} P(S_j|\lambda_{\mathrm{II}}) \end{aligned} \tag{6}$$

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 segments given a speaker pair. This search requires evaluations of SID scores of a segment given a speaker model. In our recent study, we have found that incorporating our novel Gammatone frequency cepstral coefficients (GFCCs) with binary masks and uncertainty decoding yields substantially better SID performance than the state-of-the-art robust features (Shao et al., 2007). Here, we calculate the likelihood score of a segment using this method.

Studies have shown that voiced speech plays a dominant role in sequential grouping and speaker recognition, e.g. Shao and Wang (2006). Therefore, we first apply the above sequential grouping algorithm to organize the simultaneous segments, producing two binary masks (streams) and corresponding speaker identities. Secondly, the unvoiced segments are grouped with the two streams using the above sequential grouping algorithm except that the system uses the detected speaker pair associated with the masks. We find that the unvoiced segments are typically much smaller than simultaneous segments, resulting in poor likelihood estimation by uncertainty decoding. Therefore likelihoods are calculated by a missing data method of marginalization which ignores the missing T-F units (Shao and Wang, 2006). Figure 6 presents the separated speaker streams after grouping simultaneous segments and unvoiced segments. The two speaker streams are shown in two different gray colors.

We find that the onset/offset analysis does not capture all the speech segments. Therefore,

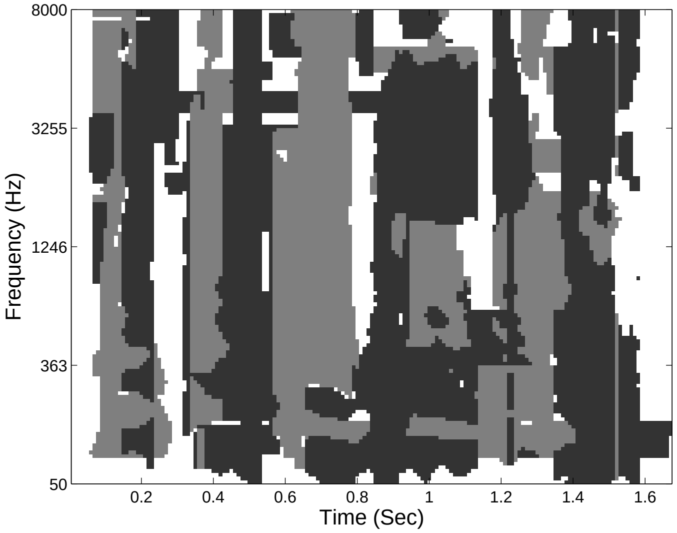


Figure 6: *The estimated speaker streams after sequential grouping of simultaneous segments and unvoiced segments. White color shows the background. The two gray-colored regions represent two separated speaker streams.*

to refine the binary masks, we apply a watershed algorithm (Vincent and Soille, 1991) to the cochleagram of the mixture and extract fragments that comprise T-F units with similar energy values. A resulting segment is absorbed by either of the aforementioned masks if it is largely (greater than two-thirds) overlapped with them. This step assumes that a small segment of connected T-F units with close energy values is produced by the same speaker. Subsequently, if a segment has not been merged, its overlapped portions, if any, are removed from itself. Finally, the remaining segments are grouped with the refined masks using the sequential grouping algorithm and the detected speaker pair.

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 ASR recognizer of Cooke et al. (2001). This recognizer treats the noise-dominant T-F units as missing or unreliable and marginalizes them 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 components 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 Gaussians,

$$p(X) = \sum_{k=1}^{M} p(k)p(X|k), \tag{7}$$

where $M = 2048$ is the number of mixture components, k is the mixture index, $p(k)$ is the mixture 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 mixture into their reliable and unreliable components as:

$$\mu_k = \begin{bmatrix} \mu_{r,k} \\ \mu_{u,k} \end{bmatrix} \ , \ \Sigma_k = \begin{bmatrix} \Sigma_{rr,k} & \Sigma_{ru,k} \\ \Sigma_{ur,k} & \Sigma_{uu,k} \end{bmatrix}. \tag{8}$$

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

We first estimate the unreliable components given the reliable ones as

$$E_{X_u|X_r}(X_u) = \sum_{k=1}^{M} p(k|X_r)\hat{X}_{u,k}, \tag{9}$$

where $p(k|X_r)$ is the *a posteriori* probability of the k'th mixture given the reliable data and $\hat{X}_{u,k}$ is the expected value of X_u given the k'th mixture. $p(k|X_r)$ 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 mixture is computed as

$$\hat{X}_{u,k} = \mu_{u,k} + \Sigma_{ur,k}\Sigma_{rr,k}^{-1}(X_r - \mu_{r,k}). \tag{10}$$

Besides estimating the unreliable T-F units, we are also interested in computing the uncertainty in our estimates. The variance associated with the reconstructed vector $\hat{X}$ can also be computed in a similar fashion to the computation of the mean in (9) as:

$$\hat{\Sigma}_{\hat{X}} = \sum_{k=1}^{M} p(k|X_r)\left\{ \left(\begin{bmatrix} X_r \\ \hat{X}_{u,k} \end{bmatrix} - \mu_k \right) \times \left(\begin{bmatrix} X_r \\ \hat{X}_{u,k} \end{bmatrix} - \mu_k \right)^T + \begin{bmatrix} 0 & 0 \\ 0 & \hat{\Sigma}_{u,k} \end{bmatrix} \right\}, \tag{11}$$

where

$$\hat{\Sigma}_{u,k} = \Sigma_{uu,k} - \Sigma_{ur,k}\Sigma_{rr,k}^{-1}\Sigma_{ru,k}. \tag{12}$$

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

$$p(z|k,q) = N(z; \mu_{k,q}, \Sigma_{k,q}) \tag{13}$$

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

$$\int_{-\infty}^{\infty} p(z|k,q)p(\hat{z}|z)dz = N(\hat{z}; \mu_{k,q}, \Sigma_{k,q} + \Sigma_{\hat{z}}). \tag{14}$$

It can be seen in (14) that the uncertainty decoder increases the variance of individual Gaussian mixture components to account for the mask estimation errors (Deng et al., 2005; Srinivasan and Wang, 2007). Here, the reconstructed GF feature $\hat{X}$ and its uncertainty $\hat{\Sigma}_{\hat{X}}$ are transformed into the GFCC domain as $\hat{z}$ and $\hat{\Sigma}_{\hat{z}}$ and then used in the uncertainty decoder. $\hat{\Sigma}_{\hat{z}}$ is the estimate of $\Sigma_{\hat{z}}$ in (14).

6 Experimental Results

We evaluate our system on the speech separation 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). The signals are sampled at 25 kHz and every utterance follows a sentence grammar of

$command $color $preposition $letter $number $adverb.

There are 4 choices each for $command, $color, $preposition and $adverb, 25 choices for $letter (A-Z except W), and 10 choices for $number (1-9 and zero). For example, a valid utterance could be "Place blue at F 2 now". The possible choices in each position are roughly uniformly distributed in the corpus. The two-talker task is to identify the letter and the number spoken by the talker who said the keyword color, "white". The SSN task is to identify the color, the letter and the number.

The training data is drawn from a closed set of 34 talkers and consists of 17,000 utterances. The two-talker test data contains pairs of sentences mixed at 6 different target-to-masker ratios (TMRs): −9, −6, −3, 0, 3 and 6 dB. One third of this data consists of same talker (ST) mixtures, another third comprises of mixtures of different talkers of the same gender (SG), and the remaining third consists of different gender (DG) mixtures. The SSN test set is generated by mixing clean utterances with speech-shaped noise at 4 SNRs: −12, −6, 0 and 6 dB. Each TMR contains 600 utterances. The SSC corpus also contains a development set. However, since our CASA system does not have parameters to tune, we do not report results on the development set.

To build speaker models, we utilize the GFCC feature as described in Section 2. Each of the 34 speaker models comprises 64 mixtures of Gaussians. The speech prior model is trained on GF features and comprises 2048 Gaussian mixtures. This prior model and the binary masks are used in the cochleagram domain to reconstruct the missing T-F units. The reconstructed GFs are then transformed into the GFCC domain using DCT. For recognition, we form the 60-dimensional feature vector of GFCC_D, including delta coefficients calculated using a sliding window of 5 frames (Young et al., 2000). GF uncertainties are also transformed into the cepstral domain since DCT is a linear transformation. Whole-word HMM-based speaker-independent ASR models are then trained on clean speech; each word model comprises 8 states and 32 Gaussian mixtures with diagonal covariance in each state. The uncertainty decoder also uses diagonal covariance for uncertainties. During the recognition process, given the estimated uncertainties and the clean ASR, the uncertainty decoder calculates the likelihood of the reconstructed GFCC_D features and transcribes the speech.

Since our CASA system 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 (*W* and *NW*):

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

and

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

where $non-white only has 3 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 *W*. The stream with a larger score is chosen as the target, i.e., stream 1 (s_1) is chosen as the target when

$$P_W(s_1) - P_{NW}(s_1) > P_W(s_2) - P_{NW}(s_2), \tag{15}$$

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. (15) is the same as,

$$P_W(s_1) + P_{NW}(s_2) > P_{NW}(s_1) + P_W(s_2). \tag{16}$$

The evaluation results of the proposed CASA system on the two-talker task is summarized in Table 1. The performance is measured in terms of recognition accuracy of the relevant keywords at each TMR conditions (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 also show the performance of our baseline system without segregation. The proposed system improves significantly over the baseline system in terms of average accuracy across all TMR conditions. Larger improvements are observed in the DG and the SG conditions. However, the system does not perform nearly as well in the ST condition, which is not a realistic 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. Figure 7 compares the system performance with (w/) and without (w/o) the ST conditions. Note that baseline performance is nearly

Table 1: *Recognition accuracy (in %) of the baseline system and the proposed CASA system on the two-talker task. DG, SG and ST refer to subconditions of "different gender", "same gender" and "same talker" respectively. Avg. is the mean accuracy.*

TMR(dB)/System		DG	SG	ST	Avg.
6	Baseline	66.00	65.92	66.52	66.17
	Proposed	80.75	76.81	54.98	70.08
3	Baseline	51.25	49.44	51.58	50.83
	Proposed	78.50	72.63	39.14	62.25
0	Baseline	36.00	34.64	32.58	34.33
	Proposed	74.50	67.31	25.34	54.25
−3	Baseline	19.25	22.07	18.55	19.83
	Proposed	63.50	53.07	20.59	44.58
−6	Baseline	9.50	10.34	9.50	9.75
	Proposed	48.00	36.31	17.19	33.17
−9	Baseline	3.25	4.75	3.62	3.83
	Proposed	32.00	22.34	11.99	21.75

the same for the with and the without ST conditions. Our CASA system achieves further absolute improvement of over 11% on average in the without ST condition over the with ST condition.

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

Table 2: *Speaker identification (SID) accuracies in the two-talker task. "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.*

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

For the SSN task, the speaker model-based sequential grouping algorithm is not applicable. Hence, we directly use the simultaneous segment as the final binary mask output. As in the two-talker condition, 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 other speech sources or speech-shaped

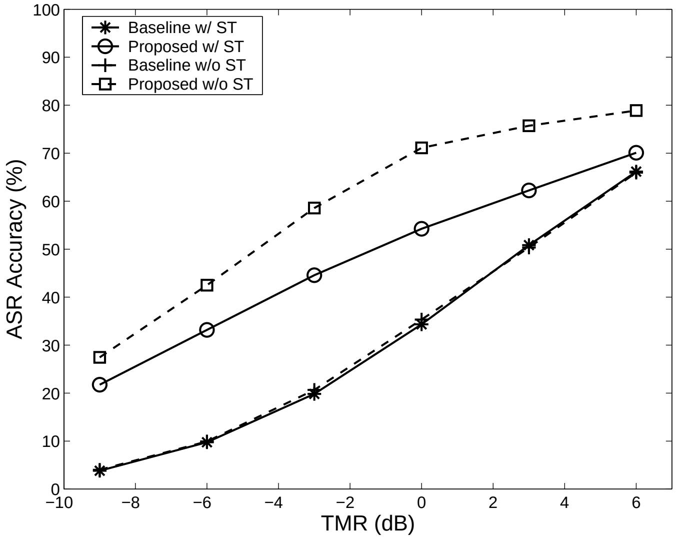


Figure 7: *Recognition accuracy on the two-talker task. The solid star line represents our baseline recognition results. The dashed plus line shows the baseline performance without the same talker (ST) data. The results of our proposed CASA system is given as the solid circle line. Its accuracy without the ST condition is presented as the dashed square line.*

Table 3: *Recognition accuracy (in %) on the SSN task using the proposed CASA system. For comparison, the baseline performance is also shown.*

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	-

noise. The proposed CASA system is an end-to-end system that follows the ASA account of auditory organization and produces 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 0dB 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, AM, 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 we have utilized text-independent speaker models to represent speaker characteristics. Moreover, the segregation 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. Further, 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.

Similar to the generality of our CASA system, the uncertainty decoding framework of Srinivasan and Wang (2007) 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.

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

Cooke, M., Lee, T., 2006. Speech separation and recognition competition.
URL `http://www.dcs.shef.ac.uk/~martin/SpeechSeparationChallenge.htm`

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. on Speech, and Audio Processing 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: Proc. ICASSP '04. Vol. 1. pp. 861–864.

Ephraim, Y., 1992. A Bayesian estimation approach for speech enhancement using hidden Markov models. IEEE Trans. on Signal Processing 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. on Speech, and Audio Processing 4, 352–359.

Hu, G., Wang, D., 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. on Audio, Speech, and Language Processing 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 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: Proc. ICASSP '04. Vol. 2. pp. 817–820.

Moore, B. C. J., 2003. An introduction to the Psychology of Hearing, 5th Edition. Academic Press, San Diego, CA.

Oppenheim, A. V., Schafer, R. W., Buck, J. R., 1999. Discrete-time signal processing, 2nd Edition. Prentice-Hall, Inc., Upper Saddle River, NJ.

Patterson, R. D., Holdsworth, J., Allerhand, M., 1992. Auditory models as perprocessors for speech recognition. In: Schouten, M. E. H. (Ed.), The auditory processing of speech: From sounds to words. Mouton de Gruyter, Berlin, Germany, Ch. 1, pp. 67–83.

Raj, B., Seltzer, M. L., Stern, R. M., 2004. Reconstruction of missing features for robust speech recognition. Speech Communication 43, 275–296.

Raj, B., Singh, R., Smaragdis, P., 2005. Recognizing speech from simultaneous speakers. In: Proc. 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., Srinivasan, S., Wang, D. L., 2007. Incorporating auditory feature uncertainties in robust speaker identification. In: Proc. ICASSP '07. Vol. IV. pp. 277–280.

Shao, Y., Wang, D. L., 2006. Model-based sequential organization in cochannel speech. IEEE Trans. on Audio, Speech and Language Processing 14, 289–298.

Srinivasan, S., Roman, N., Wang, D. L., 2006. Binary and ratio time-frequency masks for robust speech recognition. Speech Communication 48, 1486–1501.

Srinivasan, S., Wang, D. L., 2007. Transforming binary uncertainties for robust speech recognition. IEEE Transactions on Audio, Speech and Language Processing, in press.

Varga, A. P., Moore, R. K., 1990. Hidden Markov model decomposition of speech and noise. In: Proc. ICASSP '90. pp. 845–848.

Vincent, L., Soille, P., 1991. Watersheds in digital spaces: An efficient algorithm based on immersion simulations. IEEE Trans. on Pattern Analysis and Machine Intelligence 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., Brown, G. J., 1999. Separation of speech from interfering sounds based on oscillatory correlation. IEEE Trans. on Neural Networks 10 (3), 684–697.

Young, S., Kershaw, D., Odell, J., Valtchev, V., Woodland, P., 2000. The HTK Book (for HTK Version 3.0). Microsoft Corporation.