

# A Computational Auditory Scene Analysis System for Robust Speech Recognition

Sundararajan Srinivasan<sup>1</sup>, Yang Shao<sup>2</sup>, Zhaozhang Jin<sup>2</sup> and DeLiang Wang<sup>2,3</sup>

<sup>1</sup>Biomedical Engineering Department

<sup>2</sup>Department of Computer Science and Engineering

<sup>3</sup>Center for Cognitive Science

The Ohio State University

Columbus, OH 43210, USA

{srinivso, shaoy, jinzh, dwang}@cse.ohio-state.edu

## Abstract

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 T-F masks are used in conjunction with missing-data methods for recognition. Systematic evaluations on a speech separation challenge task show significant improvement over the baseline performance.

**Index Terms:** speech segregation, computational auditory scene analysis, binary time-frequency mask, robust speech recognition.

## 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 [1]. 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 [2, 3, 4]. 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 [5]. These model-based systems rely heavily on the *a priori* information of sound sources. As a result, they face difficulty in handling novel sources. For example, systems that assume and model the presence of multiple speech sources only, do not lend themselves easily to handling of 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) sys-

tem that makes weak assumptions about the various sound sources in the mixture.

From an information processing perspective, the notion of an ideal binary T-F mask has been proposed as the computational goal of CASA [6]. 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; a T-F unit denotes the signal at a particular time and frequency. Previous studies have shown that such masks can provide robust recognition results [7, 8]. Hence, we propose a CASA system that estimates this mask to facilitate the recognition of target speech in the presence of interference.

According to Bregman [9], human auditory scene analysis (ASA) takes place in two main steps: segmentation and grouping. Segmentation [10] decomposes the auditory scene into groups of contiguous T-F units or segments, each of which should originate from a single sound source. Grouping involves combining the segments that are likely to arise from the same source together into a single stream [9]. Grouping itself comprises of simultaneous and sequential organizations. Simultaneous organization involves grouping of segments at a particular time. Sequential organization refers to grouping across time.

In this paper, we present a two-stage monaural CASA system that follows the ASA account of auditory organization as described above. The input speech mixture is analyzed by an auditory filterbank in successive time frames. The system then generates segments based on a multi-scale onset and offset analysis [11]. In simultaneous grouping, 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 segments across time frames. 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 is an estimate of the ideal binary mask. This mask is used in conjunction with missing-data methods [7, 12] for recognizing the target speech utterance.

The rest of the paper is organized as follows. The next section contains a detailed presentation of our proposed system. The system has been systematically evaluated on the speech separation task that involves the recognition of a target speech utterance in the presence of either a competing speaker or speech-shaped noise [13]. The evaluation results are presented in Section 3. Fi-

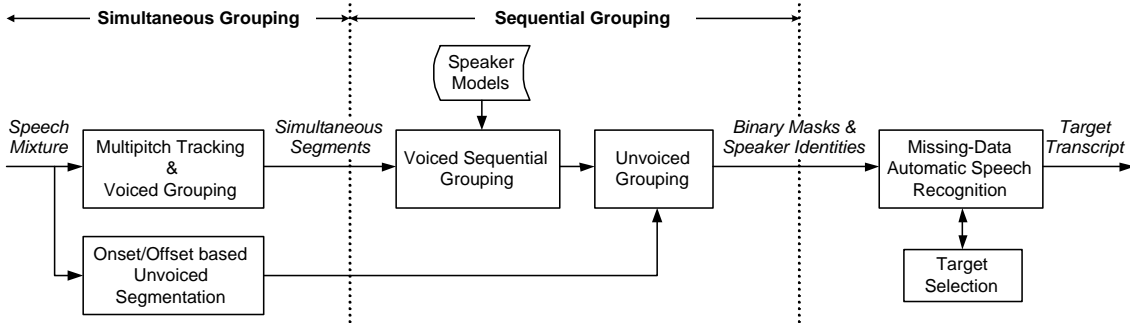


Figure 1: Schematic diagram of the proposed two-stage CASA system. In the simultaneous grouping stage, the system uses periodicity similarity to group voiced components in each frame. In addition, the system performs an onset/offset analysis to produce segments for unvoiced speech. The second stage involves sequential grouping of the voiced and the unvoiced segments across time frames. The system generates binary T-F masks and speaker identities, which are used by a missing-data ASR to recognize the target utterance.

nally, conclusions and future work are given in Section 4.

## 2. System description

Our model for monaural speech segregation is a two-stage CASA system shown in Fig. 1. The input to the system is a mixture of target and interference. In this section, we describe our system in terms of its processing of the two-talker mixtures from the speech separation task [13]. In the simultaneous grouping stage, the system estimates pitch tracks of individual sources in the mixture in order to segregate the corresponding voiced portions in each time frame. Additionally, it segments the unvoiced portions based on an onset and offset analysis. The utterances are derived from a closed set of 34 talkers of both genders. Hence, in the sequential grouping stage, speaker models are used to group the segregated voiced segments and the unvoiced segments across time to produce binary T-F masks, which are used in conjunction with missing-data methods to recognize the target utterance.

### 2.1. Simultaneous grouping

The input mixture is downsampled from its original frequency (see Section 3) to 20 kHz. The mixture signal is first analyzed using a 128-channel gammatone filterbank whose center frequencies are quasi-logarithmically spaced from 50 Hz to 8 kHz [14]. The output is then decomposed into time frames by applying a sliding window of 20 ms with 10 ms overlap. The resulting T-F energy decomposition is referred to as cochleagram and is used for further processing by the CASA system.

The human auditory system segregates a target speech source from various interferences using several cues, including differences in pitch and onsets [9]. Here, we first adapt the speech separation system in [15] to segregate the voiced portions of individual speakers. The system in [15] estimates multiple pitch tracks and their associated simultaneous voiced segments. In the low-frequency range, the system generates segments based on temporal continuity and cross-channel correlation, and groups them based on periodicity similarity. In the high-frequency range, the signal envelope fluctuates at the pitch rate and amplitude modulation (AM) rates are used for grouping [14].

We then seek to recover unvoiced segments for subsequent grouping with the estimated simultaneous voiced segments. We employ a multi-scale onset/offset analysis system [11] for segmentation. Specifically, the system first detects onsets and offsets, and

then generates segments by matching corresponding onset and offset fronts at multiple scales. A final set of segments are then produced by integrating over all scales. Unlike natural conversations, the blind mixing of speech in the speech separation task blurs and merges onset-offset fronts, thus creating segments that contain both voiced and unvoiced speech and that are not speaker homogeneous. We extract the unvoiced segments by removing those portions that are overlapped with the simultaneous segments.

### 2.2. Sequential grouping

The simultaneous voiced segments are supposedly composed of T-F units from the same speaker. However, the segments from the same speaker are still separated in time. Thus, a CASA system requires sequential grouping to organize these segments into speaker streams. For this purpose, we adapt and employ our sequential organization algorithm [16] based on speaker models. 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 the speaker pair that yields the highest aggregated SID score [16].

Studies have shown that voiced speech plays a dominating role in SID and sequential grouping (e.g. [16]). Therefore in this task, we first apply the model-based sequential grouping algorithm to organize the simultaneous segments, producing two binary masks and corresponding speaker identities. Originally, the algorithm calculates frame-level SID likelihoods for posterior maximization. However, here the simultaneous segments are composed of T-F units, resulting in missing frequency components in a frame. To accommodate this constraint, we employ a missing-data method to compute the likelihoods similar to [17]. Specifically, we apply the bounded marginalization method [7].

The unvoiced segments are subsequently grouped with the masks using the above sequential grouping algorithm except that now we use the fixed speaker pair that has been detected. We find that the onset/offset analysis does not capture all speech segments. Therefore to refine the binary masks, we apply a watershed algorithm [18] (available in the Matlab toolboxes) to the cochleagram of the mixture and extract segments that comprise of T-F units with

similar energy values. First, a watershed segment is absorbed by either of the aforementioned masks if it is strongly (greater than two-thirds) overlapped with the masks. This step assumes that a small segment of connected T-F units with close energy values is produced by the same speaker. Second, if a segment is not merged, then its overlapped portions, if any, are removed. Finally, the remaining segments are grouped with the refined masks using the sequential grouping algorithm and the detected speaker pair.

### 2.3. Recognition strategy

The output of the sequential grouping stage is a pair of binary T-F masks and the corresponding speaker identities. The speaker identities are used to infer the gender information. The binary masks are then used in a spectrogram reconstruction method of missing-data recognition [12] to reconstruct the spectral values in the T-F units labeled 0 in each mask using a gender-dependent prior speech model. To apply this method, a mixture spectrogram is first generated by applying a short-time Fourier transform, consisting of 10 ms time frames and 256 DFT coefficients, to the signal. Recall that our CASA system produces a binary mask that corresponds to a 128-channel gammatone filterbank. For consistency, this mask is mapped into the DFT domain prior to reconstruction [19]. From the reconstructed spectrograms, we compute the Mel-Frequency Cepstral Coefficients (MFCCs) for use in recognition. 12 cepstral coefficients and a logarithmic frame energy term, along with delta and acceleration coefficients, are extracted each frame. These feature vectors are then used by an ASR system to transcribe the target. For target selection, see below.

## 3. Experimental results

The proposed system has been evaluated on the speech separation task [13]. The goal in this task is 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 follow a sentence grammar of “<\$command> <\$color> <\$preposition> <\$letter> <\$number> <\$adverb>” (e.g. “Place blue at F 2 now.”). There are 4 choices each for \$command, \$color, \$preposition and \$adverb, 25 choices for \$letter (A-Z except W), and 10 choices for \$number (0-9 and zero). The two-talker task is to identify the letter and the number spoken by the talker who said “white”. The SSN task is to identify the color, the letter and the number [13]. The training data is drawn from a closed set of 34 talkers of both genders 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 data is generated by mixing clean utterances with speech-shaped noise at 4 SNRs: -12, -6, 0 and 6 dB. The test sets have 600 utterances in each TMR/SNR conditions. Since our CASA system does not have parameters to tune, we do not report results on the development set.

Each of our 34 speaker models utilizes the cochleagram feature as described in Section 2, and comprises of 64 mixtures of Gaussians. Although the speaker genders are not explicitly provided, they can be inferred from the training data. Hence, we train gender-dependent prior models for use in reconstruction. These models comprise of 1024 Gaussian mixtures. As mentioned in Section 2.3, during testing, the detected speaker identities are used

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, and R. Impr. is the relative improvement.

TMR(dB)/Sys.	DG	SG	ST	Avg.	R. Impr.	
6	Baseline	66.00	65.92	66.52	66.17	8.31
	Proposed	82.00	77.37	57.69	71.67	
3	Baseline	51.25	49.44	51.58	50.83	21.98
	Proposed	79.25	68.72	40.95	62.00	
0	Baseline	36.00	34.64	32.58	34.33	52.69
	Proposed	70.00	63.97	27.15	52.42	
-3	Baseline	19.25	22.07	18.55	19.83	105.1
	Proposed	58.50	49.16	17.65	40.67	
-6	Baseline	9.50	10.34	9.50	9.75	216.2
	Proposed	45.00	33.80	15.61	30.83	
-9	Baseline	3.25	4.75	3.62	3.83	389.6
	Proposed	29.00	16.76	11.09	18.75	

to infer the gender of segregated streams. Whole-word HMM-based gender-independent ASR models are first trained with 8 states for each word and 32 Gaussian mixtures with diagonal covariance in each state using HTK. We then perform supervised gender adaptation using maximum likelihood linear regression and maximum-a-posteriori adaptation [1].

For the two-talker task, because the identity of target is unknown after segregation we select the target stream as follows. We recognize both segregated streams using the following two grammars: “<\$command> white <\$preposition> <\$letter> <\$number> <\$adverb>” and “<\$command> <\$non-white> <\$preposition> <\$letter> <\$number> <\$adverb>” (\$non-white refers to the 3 colors excluding white). For each stream, we then obtain a normalized score by subtracting two recognition likelihoods. Finally the stream with the larger score is chosen as the target.

Table 1 summarizes the performance of the proposed CASA system on the two-talker task. Performance is measured in terms of percentage accuracy score for the relevant keywords at each TMR condition [13]. We report the results for the DG, the SG and the ST subconditions, together with 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. This is also shown by the relative improvement (R. Impr.) in accuracy. Larger improvements are observed in the DG and the SG conditions. However, the system does not perform nearly as much in the ST condition, which is not a realistic condition. This is primarily due to our use of speaker models for sequential grouping. Note that for the ST condition, neither speaker characteristics nor grammar are distinctive for segregation. Figure 2 compares the system performance with (w/) and without (w/o) the ST conditions. Note that baseline performance is nearly the same

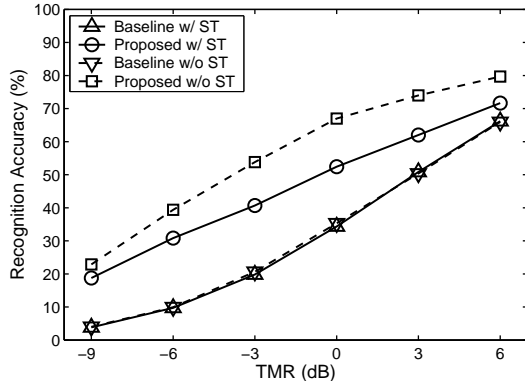


Figure 2: Recognition accuracy on the two-talker task. The solid triangle line represents our baseline recognition results. The dashed inverted-triangle line shows the baseline performance without the same talker (ST) data. The results of our proposed CASA system is given as the solid circle line. Its accuracy without the ST condition is presented as the dashed square line.

with and without ST conditions. Our CASA system achieves further absolute improvement of about 10% on average in the without ST condition over the with ST condition.

For the SSN task, the speaker model-based sequential grouping algorithm is not applicable. Hence, we directly use the simultaneous voiced segments for recognition. Note that the absence of unvoiced portions makes the resulting mask sparse and therefore for this task we use the bounded marginalization method [7] on the cochleagram feature. The missing-data ASR model is gender-independent and trained similar to the ASR training described before. Table 2 presents the performance of our system in terms of percentage recognition accuracy across different SNRs. The clean condition is included to indicate the performance of our missing-data ASR without the separation system. Across all SNR conditions, our CASA system shows a significant improvement over the baseline recognizer. This confirms the ability of our separation system to generalize well to the SSN condition.

## 4. Conclusion

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 noise. We have systematically evaluated our system on the speech separation task and obtained significant improvement over the baseline performance across all TMR/SNR conditions. 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. Moreover, the segregation does not depend on the target vocabulary. A resulting advantage is the generality of our system in terms of dealing with both speech and non-speech interferences.

**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

Table 2: Recognition accuracy (in %) on the SSN task of the proposed CASA system. For comparison, the baseline performance is also shown.

SNR(dB)	Baseline System	Proposed System	R. Impr.
Clean	93.94	-	-
6	29.50	76.78	160.2
0	16.22	66.22	308.3
-6	12.50	39.06	212.5
-12	13.00	19.22	47.85

Guoning Hu for discussion and much assistance. We acknowledge the SLATE Lab (E. Fosler-Lussier) and the Ohio Supercomputer Center for providing computing resources.

## 5. References

- [1] X. Huang, A. Acero, and H. Hon, *Spoken Language Processing*. Upper Saddle River, NJ: Prentice Hall PTR, 2001.
- [2] T. Kristjansson, H. Attias, and J. Hershey, "Single microphone source separation using high resolution signal reconstruction," in *Proc. ICASSP '04*, vol. 2, 2004, pp. 817–820.
- [3] G-J Jang and T-W Lee, "A probabilistic approach to single channel blind signal separation," in *Advances in Neural Information Processing Systems 15*, S. Becker, S. Thrun, and K. Obermayer, Eds. Cambridge, MA: MIT Press, 2003, pp. 1173–1180.
- [4] B. Raj, R. Singh, and P. Smaragdis, "Recognizing speech from simultaneous speakers," in *Proc. Interspeech '05*, 2005, pp. 3317–3320.
- [5] A. N. Deoras and M. Hasegawa-Johnson, "A factorial HMM approach to simultaneous recognition of isolated digits spoken by multiple talkers on one audio channel," in *Proc. ICASSP '04*, vol. 1, 2004, pp. 861–864.
- [6] D. L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech separation by humans and machines*, P. Divenyi, Ed., Norwell, MA, 2005, pp. 181–197.
- [7] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Comm.*, vol. 34, pp. 267–285, 2001.
- [8] N. Roman, D. L. Wang, and G. J. Brown, "Speech segregation based on sound localization," *J. Acoust. Soc. Am.*, vol. 114, pp. 2236–2252, 2003.
- [9] A. S. Bregman, *Auditory scene analysis*. Cambridge, MA: The MIT Press, 1990.
- [10] D. L. Wang and G. J. Brown, "Separation of speech from interfering sounds based on oscillatory correlation," *IEEE Trans. on Neural Networks*, vol. 10, no. 3, pp. 684–697, 1999.
- [11] G. Hu and D. L. Wang, "Auditory segmentation based on onset and offset analysis," *IEEE Trans. on Audio, Speech and Language Processing*, 2006, in press.
- [12] B. Raj, M. L. Seltzer, and R. M. Stern, "Reconstruction of missing features for robust speech recognition," *Speech Communication*, vol. 43, pp. 275–296, 2004.
- [13] M. Cooke and T-W. Lee. Speech separation and recognition competition. Available at <http://www.dcs.shef.ac.uk/~martin/SpeechSeparationChallenge.htm>
- [14] G. Hu and D. L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. on Neural Networks*, vol. 15, pp. 1135–1150, 2004.
- [15] G. Hu, "Monaural speech organization and segregation," Ph.D. dissertation, Biophysics Program, The Ohio State University, 2006.
- [16] Y. Shao and D. L. Wang, "Model-based sequential organization in cochannel speech," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, pp. 289–298, 2006.
- [17] —, "Robust speaker recognition using binary time-frequency masks," in *Proc. ICASSP '06*, vol. 1, 2006, pp. 645–648.
- [18] L. Vincent and P. Soille, "Watersheds in digital spaces: An efficient algorithm based on immersion simulations," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 13, no. 6, pp. 583–598, 1991.
- [19] S. Srinivasan and D. L. Wang, "A supervised learning approach to uncertainty decoding for robust speech recognition," in *Proc. ICASSP '06*, vol. 1, 2006, pp. 297–300.