Final AFOSR Project Performance Report

DeLiang Wang

(Principal Investigator)

The Ohio State University

March 2008

This PI was awarded the AFOSR grant "Monaural Speech Segregation by Integrating Primitive and Schema-based Analysis" (Grant No.: FA9550-04-1-0117). The project was funded for the period of 2/15/04 to 12/31/07 with the total amount of $672K. This report summarizes the progress made throughout the project period.

1. RESEARCH PROGRESS

The natural auditory environment typically contains multiple simultaneous events. A remarkable feat of auditory perception is the ability to disentangle the acoustic mixture and group the components of the same event into a stream. This aspect of human audition is called auditory scene analysis (ASA), which has a primitive (bottom-up) process and a schema-based (top-down) process. A major task of auditory scene analysis is monaural segregation of speech from interfering sounds. This project seeks to develop an auditory scene analysis approach to monaural speech segregation.

Consistent with the stated objectives of the project, the project has made considerable progress along the following four directions. First, we have proposed a schema-based model for phonemic restoration, which refers to the perceptual synthesis of the phonemes that are masked by appropriate replacement sounds by utilizing lexical context. Second, we have developed an approach to address the problem of sequential organization, which is based on trained speaker models. Third, we have proposed an approach for segmentation of auditory scenes based on event detection, in an attempt to address the segregation of unvoiced speech. Fourth, we have developed a comprehensive system for segregating unvoiced speech, a long standing challenge in computational auditory scene analysis (CASA). In addition, encouraging progress has been made on enhancing reverberant speech and modeling of multitalker speech perception.

The major findings along the above directions are described in more detail in the following five subsections.

1.1 A Schema-based Model for Phonemic Restoration

In 1970, R. Warren discovered that, when a masking sound (cough) fully replaced the first "s" of the word "legislatures" in the sentence "The state governors met with their respective legislatures convening in the capital city," listeners reported the hearing of the

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
03-02-2008	Final	

4. TITLE AND SUBTITLE

Monaural Speech Segregation by Integrating Primitive and Schema-based Analysis

5a. CONTRACT NUMBER

5b. GRANT NUMBER
FA9550-04-1-0117

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

DeLiang Wang

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

The Ohio State University

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

AFOSR/NL
875 N Randolph St,
Arlington, VA 22203

10. SPONSOR/MONITOR'S ACRONYM(S)
AFOSR

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-SR-AR-TR-08-0306

12. DISTRIBUTION / AVAILABILITY STATEMENT

Distribution A: Approved for Public Release

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The natural auditory environment typically contains multiple simultaneous events. A remarkable feat of auditory perception is the ability to disentangle the acoustic mixture and group the components of the same event into a stream. This aspect of human audition is called auditory scene analysis (ASA), which has a primitive (bottom-up) process and a schema-based (top-down) process. A major task of auditory scene analysis is monaural segregation of speech from interfering sounds. This project seeks to develop an auditory scene analysis approach to monaural speech segregation.

Consistent with the stated objectives of the project, the project has made considerable progress along the following four directions. First, we have proposed a schema-based model for phonemic restoration, which refers to the perceptual synthesis of the phonemes that are masked by appropriate replacement sounds by utilizing lexical context. Second, we have developed an approach to address the problem of sequential organization, which is based on trained speaker models. Third, we have proposed an approach for segmentation of auditory scenes based on event detection, in an attempt to address the segregation of unvoiced speech. Fourth, we have developed a comprehensive system for segregating unvoiced speech, a long standing challenge in computational auditory scene analysis (CASA). In addition, encouraging progress has been made on enhancing reverberant speech and modeling of multitalker speech perception.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER *(include area code)*

masked phoneme. When phonemic restoration happens, subjects were unable to localize the masking sound within a sentence accurately; that is, they could not identify the position of the masker in the sentence. When "s" was replaced with silence instead, phonemic restoration was not observed. Subsequent perceptual studies have shown that phonemic restoration is dependent on the linguistic skills of listeners, the characteristics of the masking sound, and temporal continuity of speech.

Existing models for phonemic restoration, however, use only temporal continuity. These models poorly restore unvoiced phonemes and are also limited in their ability to restore voiced phonemes. We have proposed a schema-based model for phonemic restoration. Our model employs lexical knowledge in the form of a speech recognizer and a sub-lexical representation in word templates realizing the role of speech schemas. The model corresponds to a multi-stage system, where the input is utterances with words containing masked phonemes. The maskers used in our experiments are broadband sound sources. Phonemes are masked by adding a noise source to the signal waveform. In the first stage input waveform with masked phonemes is converted into a spectrogram by Fourier analysis. A binary mask for the spectrogram is generated in this stage to identify reliable and unreliable parts. If a time-frequency (T-F) unit in the spectrogram contains predominantly speech energy, it is labeled reliable; it is labeled unreliable otherwise. We then identify the spectro-temporal regions which predominantly contain energy from speech in a two step process. In the first step, two features are calculated at each frame, spectral flatness and normalized energy. Masked, unvoiced and silent frames, all have high spectral flatness, but the energy in masked frames is higher than that in unvoiced and silent frames. These features are then fed to a perceptron classifier which labels each frame as being either clean (reliable) or noisy (unreliable). In the second step, the frequency units in a noisy frame are further analyzed for possible temporal continuity with neighboring clean frames using a Kalman filter. Spectral regions in the noisy frames which exhibit strong continuity with the spectral regions of the neighboring clean frames are also labeled clean.

The second stage is missing data automatic speech recognition (ASR) based on hidden Markov model (HMM), which provides word level recognition of the input signal by utilizing only the reliable spectro-temporal regions. Thus, the input to the missing data ASR is the spectrogram of the input signal along with a corresponding binary mask. As for restoration, template-based speech recognizers use spectral templates to model each word. These templates could be used as a base for restoration. We train a word-level template corresponding to each HMM model in missing data ASR. During training, signals are converted into a cepstral representation and then time normalized by dynamic time warping (DTW). We then compute their average. Two sets of templates are considered, speaker-independent and speaker-dependent. In certain applications, it may be possible to identify the target speaker for enhancement. In such applications, these speaker-dependent templates could be applied.

Based on the results of recognition, word templates corresponding to the noisy words are selected. A template thus activated is warped to be of the same duration as the noisy word using DTW. The T-F units of the template corresponding to the unreliable T-F units then replace the unreliable units of the noisy word.

A template is an average representation of each word. Thus, the restored phoneme may not be in consonance with the speaking style of the remaining utterance. In order to

maintain the overall naturalness of the utterance, we perform pitch based smoothing. The last stage of the model is the overlap-and-add method of resynthesis. Resynthesized waveforms are used for informal listening and performance evaluation.

Figure 1 shows an example of phonemic restoration by our model. Figure 1a shows the spectrogram of the word "eight". Figure 1b shows the spectrogram of the word mixed with a burst of white noise that masks the phoneme /t/ in the word. Figure 1c shows the result of our model where the masked phoneme is largely restored. For comparison, Figure 1d shows the result from a previously proposed restoration model that is based on Kalman filtering.

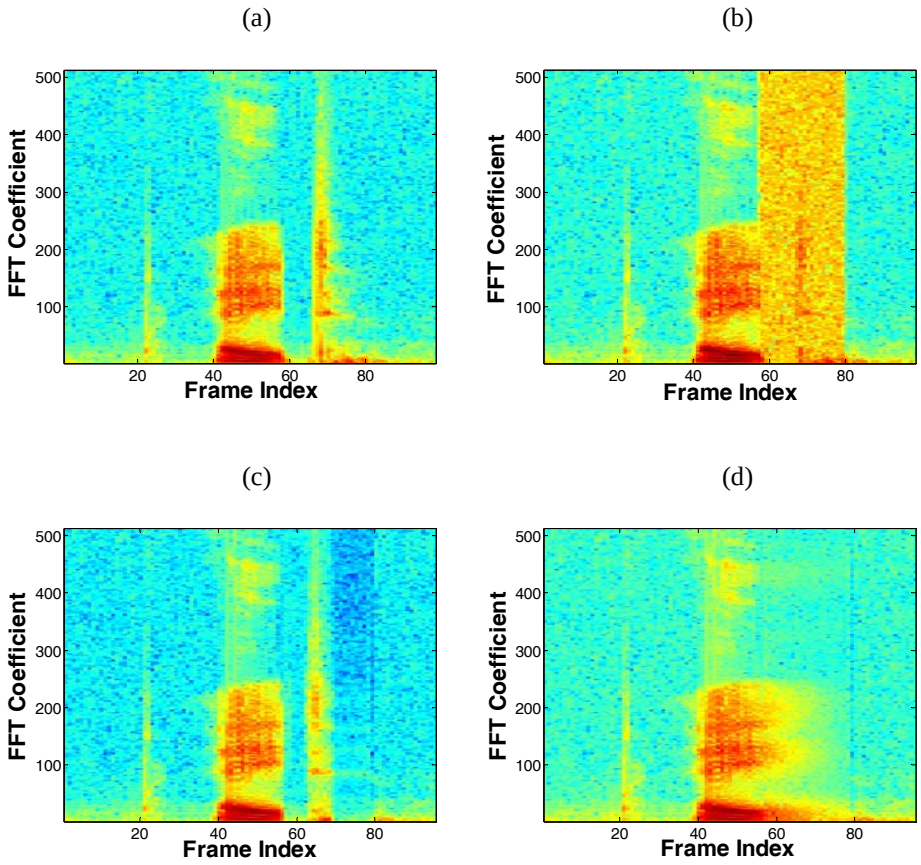

Figure 1. Illustration of phonemic restoration. (a). The spectrogram of the word "eight". (b). The spectrogram obtained from (a) with the phoneme /t/ masked by a burst of white noise. (c). The restoration of the masked phoneme from the proposed model. (d). The retoration of the masked phoneme from a Kalman filter model.

The proposed model is able to restore both voiced and unvoiced phonemes with a high degree of naturalness. Systematic testing shows that our model outperforms a previous model that employs Kalman filtering for synthesizing or extrapolating the masked sound. The model is described in a 2005 paper by S. Srinivasan and D.L. Wang,

entitled "A schema-based model for phonemic restoration," published in *Speech Communication* (see Sect. 2.4 for detailed reference).

1.2 Sequential Organization Based on Speaker Models

A human listener has the ability to follow a speaker's voice while others are talking simultaneously. This aspect of ASA is referred to as sequential organization, which integrates sound components across time into the same perceptual stream. Our study on sequential organization focuses on cochannel speech, or mixtures of two voices. In the study, we have explored speaker characteristics, particularly speaker models, for grouping the T-F energy of the same speaker into a single stream in cochannel speech. As a result of successful sequential organization, speaker recognition from cochannel mixtures should also improve. That is, robust speaker identification in cochannel conditions is a major benefit of performing sequential organization.

Our proposed system consists of three stages. First, a multipitch tracking algorithm is adapted and applied to cochannel speech and pitch contours for both speakers are produced. The algorithm filters the mixture signal into multiple frequency channels through an auditory filterbank; it then selects "clean" channels and peaks within each clean channel as pitch candidates at each time frame. Multiple pitch hypotheses are formed; the hypotheses are further integrated across the frequency channels. Afterwards, pitch contours are decoded as a sequence of most likely pitch hypotheses using an HMM framework.

The second stage is used to extract usable speech from a cochannel mixture based on pitch information. Due to the nature of human voice, a speech utterance contains voiced portions, unvoiced portions and silence. Therefore, there are some portions (segments) of cochannel speech that contain only one speaker's voiced part or one speaker's voiced part plus another speaker's unvoiced part, the latter usually having much lower energy. The voiced spectra of these frames are minimally corrupted, and can be used to derive speaker features for speaker identification. So they form usable speech and are retained, while the portions with overlapping pitch contours as well as silent portions are removed, resulting in a set of usable speech segments.

For any two segments in the usable speech set, whether they are from the same speaker is unknown. In the third stage, our model-based sequential grouping algorithm groups the segments into two speaker streams by searching for the optimal hypothesis in the joint speaker and grouping space. Our formulation is extended from the traditional SID (speaker identification) probabilistic framework. Exhaustive search in the space is computationally prohibitive with increasing number of segments. Thus we propose a hypothesis pruning algorithm to remove hypotheses of low likelihoods, which drastically reduces computation time while resulting in comparable performance with exhaustive search. As a byproduct, speaker identities are also determined.

Our model-based approach for sequential organization assigns the extracted usable speech segments into speaker streams. Our usable speech extraction method produces segments useful for cochannel SID across various target-to-interferer ratios. Evaluation results show that the proposed hypothesis pruning algorithm achieves SID performance close to the ceiling performance with prior pitch information or exhaustive search, and it

performs significantly better than alternative approaches to speaker assignment. Our sequential grouping algorithm can also handle the situation where only one speaker is present in a cochannel mixture. A paper describing our approach was published in a 2006 paper by Y. Shao and D.L. Wang, entitled "Model-based sequential organization in cochannel speech," in *IEEE Transactions on Audio, Speech and Language Processing*.

1.3 Auditory Segmentation

Perceptual theories of auditory scene analysis suggest that ASA takes place in two conceptual stages: Segmentation and grouping. The first stage decomposes an auditory scene into a collection of auditory elements (segments), each of which should originate from the same sound source. The second stage selectively groups the segments into distinct streams, corresponding to different auditory events. We believe that auditory segmentation is a key computational stage in speech segregation.

We have investigated auditory segmentation by analyzing onsets and offsets of auditory events. Onsets and offsets are important ASA cues, and there is strong evidence for onset detection by auditory neurons. Onsets and offsets, corresponding to sudden intensity changes, tend to delineate auditory events. Quantitatively speaking, onsets and offsets correspond to the peaks and valleys of the time derivative of the intensity. However, because of intensity fluctuations within individual events, many peaks and valleys of the derivative do not correspond to real onsets and offsets. Therefore, we smooth the intensity over time to reduce the fluctuations in the smoothing stage. The degree of smoothing is called the *scale* – the larger the scale is, the smoother the intensity becomes. In the stage of onset/offset detection and matching, the system detects onsets and offsets in each filter channel and merges detected onsets and offsets into onset and offset fronts if they occur at close times. It then matches individual onset and offset fronts to form segments.

As a result of smoothing, event onsets and offsets of small T-F regions may be blurred at a larger (coarser) scale. Consequently, the system may miss small events or generate segments combining different events, a case of under-segmentation. On the other hand, at a smaller (finer) scale, the system may be sensitive to insignificant intensity fluctuations within individual events. Consequently, the system tends to separate a continuous event into several segments, a case of over-segmentation. Therefore, it is difficult to obtain satisfactory segmentation with a single scale. Our system handles this issue by integrating onset/offset information across different scales in an orderly manner in the stage of multiscale integration, which yields the final set of segments.

Figure 2 illustrates the performance of our multiscale analysis system for segmentation. The bounding contours of obtained segments for the mixture in Figure 2(f) are shown in Figure 2(a)-(d) for four scales of analysis. The background is represented by blue. Compared with the ideal binary mask in Figure 2(e), which labels all T-F regions where target has stronger energy than interference, the obtained segments capture a majority of target speech. Some segments for the interference are also formed. Note that the segmentation stage does not distinguish segments corresponding to target and those corresponding to interference, which is the task of grouping.

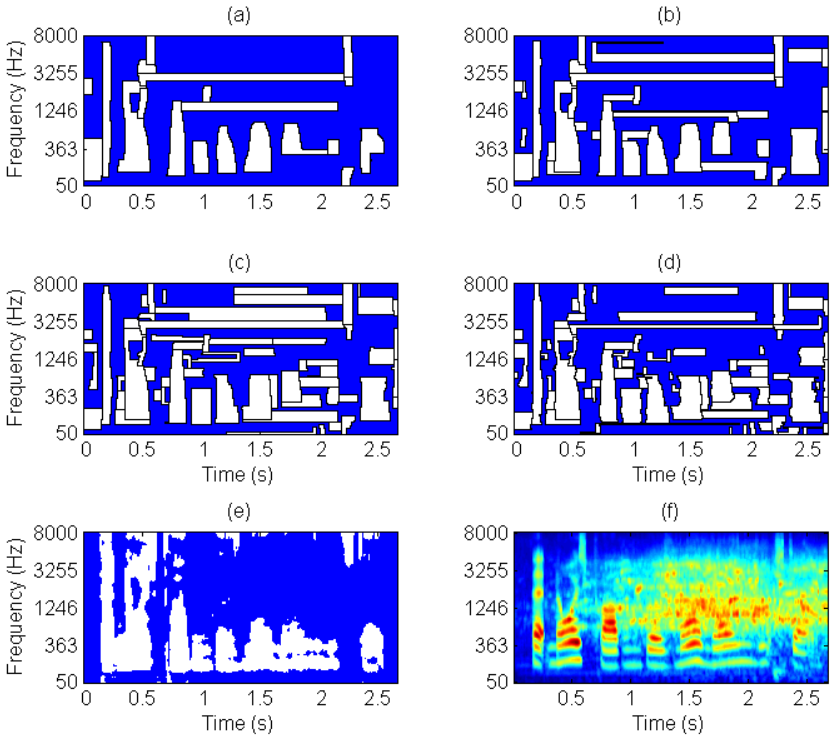

Figure 2. Auditory segmentation. Obtained segments correspond to white regions with black bounding contours, and the background is indicated by blue. (a)-(d) Results of segmentation at four different scales for the input shown in (f). (e). The ideal binary mask for the input. (f). The cochleagram of an input mixture which corresponds to a speech utterance mixed with a crowd noise at 0-dB SNR.

Extensive evaluation shows that much target speech, including unvoiced speech, is correctly segmented, and target speech and interference are well separated into different segments. This work was published in a 2007 paper by G. Hu and D.L. Wang, entitled "Auditory segmentation based on onset and offset analysis," in *IEEE Transactions on Audio, Speech and Language Processing*.

1.4. Unvoiced Speech Segregation

In English, unvoiced speech is composed of a subset of stops, fricatives, and affricates. With the exception of the fricative /h/, stops, fricatives, and affricates are called obstruents in phonetics. To simplify terminology, we refer to all of them as *expanded obstruents*. Unvoiced speech segregation is a great deal more difficult than voiced speech segregation because of two reasons. First, unvoiced speech lacks the harmonic cue and is often noise-like acoustically. Second, sound energy of unvoiced speech is usually much weaker than that of voiced speech; as a result, unvoiced speech is more susceptible to interference. Our approach to unvoiced speech segregation first segments an input mixture using the onset/offset based method (see Sect. 1.3), which is

applicable to both unvoiced and voiced speech, and then groups segments dominated by unvoiced speech. Due to the lack of an effective technique for sequential grouping, our study focuses on segregating unvoiced speech from non-speech interference.

A segment may be dominated by voiced target, unvoiced target, or interference. Our goal is to group segments dominated by unvoiced target. As voiced speech is expected to be easier to segregate, we first employ voiced speech segregation and use its results to identify the segments dominated by voiced speech. We consider a segment to be dominated by voiced target if more than half of its total energy is included in the voiced time frames of the segment, and more than half of its energy in the voiced frames is included in segregated voiced speech. All the segments dominated by voiced target are grouped into a voiced stream. Note that the voiced stream may include some unvoiced speech because an unvoiced consonant is often coarticulated with a neighboring voiced phoneme, hence included in a segment dominated by voiced target.

Once segments dominated by voiced speech are grouped, the remaining segments will be dominated by either unvoiced speech or interference. Consequently, we formulate unvoiced speech segregation as a classification problem. Let s denote a remaining segment, which lasts from frame m_1 to m_2, and $\mathbf{X}_s = [X_s(m_1), X_s(m_1+1), \ldots, X_s(m_2)]$ its corresponding T-F region on the cochleagram. $H_0(m_1, m_2)$ denotes the hypothesis that s is dominated by speech and $H_1(m_1, m_2)$ the hypothesis that it is dominated by interference. Furthermore, let $H_{0,a}(m_1, m_2)$ be the hypothesis that this region is dominated by an expanded obstruent and $H_{0,b}(m_1, m_2)$ by any other speech sound. We classify s as dominated by unvoiced speech if:

$$P(H_{0,a}(m_1, m_2) \mid \mathbf{X}_s) > P(H_1(m_1, m_2) \mid \mathbf{X}_s) \tag{1}$$

Because the durations of segments are varied, direct evaluation of the probabilities in the above inequality is unfeasible computationally. Therefore, we assume that each time frame is statistically independent. With this frame independence assumption, (1) becomes,

$$\prod_{m=m_1}^{m_2} P(H_{0,a}(m) \mid X_s(m)) > \prod_{m=m_1}^{m_2} P(H_1(m) \mid X_s(m)) \tag{2}$$

By applying the Bayes rule and a further assumption that the prior and the posterior probabilities of a frame do not depend on the frame index within a given segment, we have,

$$\left[\frac{P(H_{0,a})}{P(H_1)}\right]^{m_2-m_1+1} \prod_{m=m_1}^{m_2} \frac{p(X_s(m) \mid H_{0,a})}{p(X_s(m) \mid H_1)} > 1 \tag{3}$$

The prior probability ratio of $P(H_{0,a})$ and $P(H_1)$ obviously depends on the SNR of the acoustic mixture, and this relationship can be approximated by a linear function. Moreover, one can estimate mixture SNR from segregated voiced speech.

In (3), the likelihood ratio between $p(X_s(m)|H_{0,a})$ and $p(X_s(m)|H_1)$ is estimated by training a multilayer perceptron (MLP) whose desired output is 1 if the corresponding

frame is dominated by an expanded obstruent and 0 otherwise. Note that the trained MLP gives a good estimate of the probability. With the MLP estimate of the likelihood ratio and the SNR-based estimate of the prior probability ratio, (3) is used to label a segment as either expanded obstruent or interference. All the segments labeled as unvoiced speech are grouped to the voiced stream to produce the final segregated speech stream.

We have systematically evaluated segregation performance in terms of an SNR metric. The evaluation uses a test corpus containing 20 target utterances randomly selected from the test part of the TIMIT database mixed with 15 nonspeech intrusions. The intrusions have not been used during training, and represent a broad range of nonspeech sounds encountered in typical acoustic environments. Each speech utterance is mixed with every intrusion at the SNR levels of -5 dB, 0 dB, 5 dB, 10 dB, and 15 dB. Hence the test corpus contains 300 mixtures at each SNR level and 1500 mixtures in total. Figure 3 shows the systematic results at different SNRs. Figure 3(a) and 3(b) display the average SNR of segregated target and the corresponding SNR gain. Figures 3(c) and 3(d) display the results at unvoiced frames separately. The figure clearly shows that our system produces significant SNR improvements. To put our performance in perspective, Figure 3 also shows the SNR results of a spectral subtraction method. It is clear from the figure that our system performs substantially better for both voiced and unvoiced speech than spectral subtraction, with the only exception that occurs for unvoiced speech segregation at the input SNR of 15 dB. The amount of improvement increases with decreasing mixture SNR.

Our study on unvoiced speech segregation represents the first systematic effort on addressing this challenge. The results have been published in a series of ICASSP papers by G. Hu and D.L. Wang, and an extensive paper describing this work, authored by G. Hu and D.L. Wang, has been accepted by the *Journal of the Acoustical Society of America*.

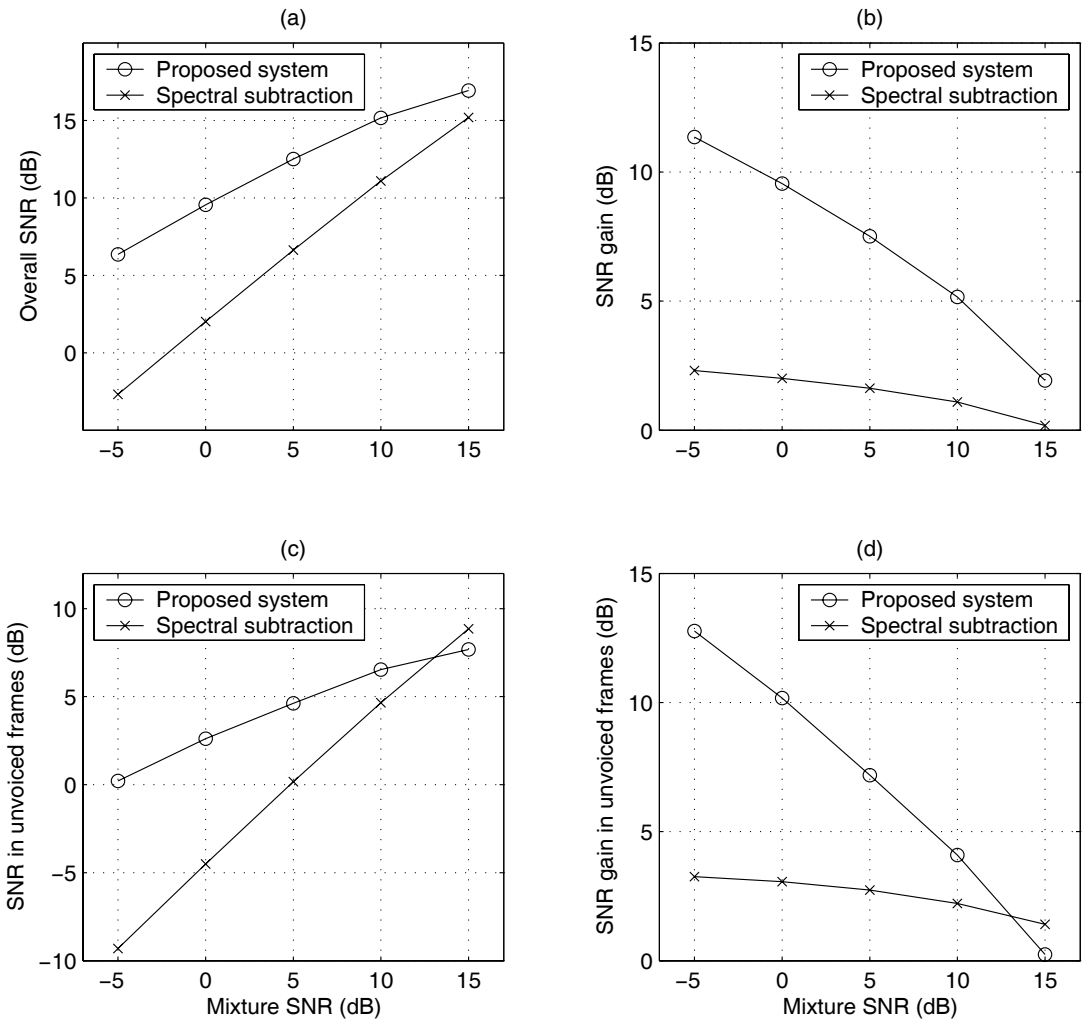

Figure 3. SNR results of the proposed system and spectral subtraction. (a) SNRs of segregated speech at different mixture SNR levels. (b) SNR gains of segregated targets. (c) SNRs of segregated targets at unvoiced frames. (d) SNR gains of segregated targets at unvoiced frames.

1.5. Other Advances

A main cause of speech degradation in practically all listening situations is room reverberation. Although human listening is, to a considerable degree, little affected by room reverberation – indeed increased loudness as a result of reverberation may even enhance speech intelligibility – reverberation causes major performance degradation for machine listening. Consequently, an effective reverberant speech enhancement system is essential for many speech technology applications including speech and speaker recognition. Also, hearing-impaired listeners suffer from reverberation effects to a much greater extent than normal-hearing listeners. Hence a system that enhances reverberant speech could contribute to the design of more effective hearing aids. Under noise-free

conditions, the perceived quality of reverberant speech is determined by two distinct perceptual attributes: coloration and long-term reverberation. They correspond to two physical variables: signal-to-reverberant energy ratio (SRR) and reverberation time, respectively. Based on this analysis, we have proposed a two-stage approach to enhance reverberant speech recorded monaurally. In the first stage, an inverse filter is estimated in order to reduce coloration effects or increase SRR. The second stage employs spectral subtraction to minimize the influence of long-term reverberation. Our two-stage algorithm has been systematically evaluated, and the evaluation results show that the algorithm achieves substantial enhancement of reverberant speech. We have also carried out a quantitative comparison with a recent enhancement algorithm on a corpus of reverberant speech and our algorithm yields significantly better performance. A paper describing the two-stage algorithm, by M. Wu and D.L. Wang, was published by *IEEE Transactions on Audio, Speech and Language Processing* in 2006.

In everyday listening, both background noise and reverberation degrade the speech signal. Psychoacoustic research suggests that human speech perception under reverberant conditions relies mostly on monaural processing. While speech segregation based on periodicity has achieved considerable progress in handling additive noise, little research in monaural segregation has been devoted to reverberant scenarios. Reverberation smears the harmonic structure of speech signals, and our evaluations using a pitch-based segregation algorithm show that an increase in the room reverberation time causes degraded performance due to weakened periodicity in the target signal. We have developed a two-stage monaural separation system that combines the inverse filtering of the room impulse response corresponding to target location and pitch-based speech segregation. The key idea in the first stage is to estimate a filter that inverts the room impulse response corresponding to the target source. The effect of applying this inverse filter on the reverberant mixture is two-fold: It improves the harmonic structure of the target signal while smearing those signals originating at other locations. We have found that this effect provides a better input signal for the pitch-based segregation stage. The second stage adapts a voiced speech segregation system. A systematic evaluation of this two-stage system shows that it results in considerable SNR gains across different conditions. To our knowledge, this is the first study that addresses monaural speech segregation with room reverberation. A paper describing this algorithm, by N. Roman and D.L. Wang, was published by the *Journal of the Acoustical Society of America* in 2006.

Listeners' ability to understand a target speaker in the presence of one or more simultaneous competing speakers is subject to two types of masking: Energetic and informational. Energetic masking occurs when target and interfering signals overlap in time and frequency resulting in portions of the target becoming inaudible. Informational masking occurs when the listener is unable to segregate the target from interference, while both are audible. We have proposed a novel model of multitalker speech perception that accounts for both of the above types of masking. Human perception in the presence of energetic masking is modeled using a speech recognizer that treats the masked time-frequency (T-F) units of target as missing data. Specifically, when target speech is presented together with interference, some T-F regions will contain predominantly target energy (reliable) and the rest are subject to energetic masking by interference. We use a missing data recognition method that treats the latter T-F regions as unreliable during recognition. To apply missing data recognition requires a binary T-F mask that provides

information about which T-F regions, of the mixture signal, are reliable and which are unreliable. The task of generating such a mask is essentially the task of segregating the target from the mixture. Therefore to model informational masking, we employ a voiced speech segregation system in order to estimate a binary mask that selects the T-F regions of the mixture where target dominates interference. The similarities between target and interference characteristics affect the performance of speech segregation and hence contribute to informational masking in our model. Using this model we have quantitatively simulated several aspects of listeners' performance in multitalker conditions, including the differential effects of energetic and informational masking on multitalker perception. The performance of our model is in broad agreement with perceptual results. A preliminary report on this work, by S. Srinivasan and D.L. Wang, was published in the *Proceedings of 2005 INTERSPEECH*, and a comprehensive version has been conditionally accepted by the *Journal of the Acoustical Society of America*.

2. OTHER INFORMATION

2.1 Development of Human Resources

The project in various stages has supported four doctoral students as graduate research assistants: Nicoleta Roman, Guoning Hu, Soundar Srinivasan, and Yang Shao. The support has enabled them to complete their doctoral studies.

Roman's work on location-based speech segregation and pitch-based segregation of reverberant speech, led to a Ph.D. dissertation completed in August 2005. Her dissertation, entitled "Auditory-based algorithms for sound segregation in multisource and reverberant environments", is posted on the PI's laboratory webpage at http://www.cse.ohio-state.edu/pnl/theses.html. An executive summary of the dissertation is given in Appendix 1.

Hu's work on segregation of both voiced and unvoiced speech led to a Ph.D. dissertation entitled "Monaural speech organization and segregation" completed in June 2006. His dissertation will be soon posted on the same website, pending provisional patent filing. An executive summary of the dissertation is given in Appendix 2.

Srinivasan's work on integrating CASA and robust speech recognition led to a Ph.D. dissertation entitled "Integrating computational auditory scene analysis and automatic speech recognition," completed in September 2006. His dissertation has been posted on the above website, and an executive summary is given in Appendix 3.

Shao's work on model-based sequential grouping and robust speaker recognition led to a Ph.D. dissertation entitled "Sequential organization in computational auditory scene analysis," completed in September 2007. His dissertation will soon be posted on the above website, and an executive summary is given in Appendix 4.

This grant has helped the PI to update a graduate-level course entitled "Computational audition", and enhance the existing graduate-level courses "Survey of Artificial Intelligence", "Introduction to Neural Networks" and "Brain Theory and Neural Networks". Additionally, the PI has participated in a great deal of curriculum and seminar activity for training undergraduate students.

2.2 Honors/Awards

The PI was elected in 2004 to IEEE Fellow "For contributions to advancing oscillatory correlation theory and its application to auditory and visual scene analysis." IEEE, standing for the Institute of Electrical and Electronics Engineers, is the largest professional organization in the world.

The PI received the 2005 Lumley Research Award from the OSU College of Engineering. This is the third consecutive time the PI received this recognition (over a period of 13 years).

Guoning Hu received a Student Research Award from the OSU Biophysics Graduate Program in 2005.

The PI received the Outstanding Paper Award from IEEE Computational Intelligence Society in 2007 for his paper entitled "The time dimension for scene analysis," published in *IEEE Transactions on Neural Networks* in 2005.

2.3 Transition or Collaborative Activities

The PI has collaborated with Dr. Douglas Brungart of AFRL (Dayton OH) on psychoacoustic evaluation of computational multitalker analysis systems. The collaboration has led to a number of results concerning energetic and informational masking as well as the effectiveness of ideal binary masking, a notion developed in the PI's laboratory. A paper summarizing the results was published by the *Journal of the Acoustical Society of America* in 2006, and a second paper is currently under revision for the *Journal of the Acoustical Society of America*.

Aetion Technologies, a small-business company located in Columbus, Ohio, has partnered with the PI in a Phase I STTR project funded by AFOSR. This Phase I project conducted a feasibility study to determine what improvements were needed in order to apply the speech segregation algorithms developed in the PI's laboratory in practical situations.

The PI had a 2-year project from AFRL/IF in Rome, New York, to study speaker recognition in co-channel conditions (ended in June 2006). For the project, we have applied the results from this AFOSR project to perform speech segregation as a preprocessing step in order to solve the robust speaker recognition problem. The PI visited the Rome Lab in May 2005 (hosted by Drs. Stanley Wenndt and John Grieco) and presented a 2-day tutorial on computational auditory scene analysis, and again in June 2006 (hosted by Dr. Stanley Wenndt) to present the results achieved in the speaker recognition project.

From October 2006 to June 2007, the PI visited Oticon A/S, a major hearing aid manufacturer located in Copenhagen, Denmark. He has collaborated with the Oticon signal processing group as well as the Oticon Eriksholm Research Center in an effort to evaluate the potential benefits of speech separation algorithms in improving speech intelligibility of hearing impaired listeners.

2.4 Publications

Edited Books

Wang D.L. and Brown G.J. (ed.): *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. IEEE Press/Wiley, 2006.

King I., Wang J., Chan L., and Wang D.L. (ed.): *Neural Information Processing*. Springer Lecture Notes in Computer Science, Springer, 2006.

Journal articles

Hu G. and Wang D.L. (2004): "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Transactions on Neural Networks*, vol. 15, pp. 1135-1150.

Palomäki K.J., Brown G.J., and Wang D.L. (2004): "A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation," *Speech Communication*, vol. 43, pp. 361-378.

Srinivasan S. and Wang D.L. (2005): "A schema-based model for phonemic restoration," *Speech Communication*, vol. 45, pp. 63-87.

Wang D.L., Kristjansson A., and Nakayama K. (2005): "Efficient visual search without top-down or bottom-up guidance," *Perception & Psychophysics*, vol. 67, pp. 239-253.

Wang D.L. (2005): "The time dimension for scene analysis," *IEEE Transactions on Neural Networks*, vol. 16, pp. 1401-1426.

Shao Y. and Wang D.L. (2006): "Model-based sequential organization in cochannel speech," *IEEE Transactions on Audio, Speech, & Language Processing*, vol. 14, pp. 289-298.

Wu M. and Wang D.L. (2006): "A two-stage algorithm for one-microphone reverberant speech enhancement," *IEEE Transactions on Audio, Speech, & Language Processing*, vol. 14, pp. 774-784.

Wu M. and Wang D.L. (2006): "A pitch-based method for the estimation of short reverberation time," *Acta Acustica united with Acustica*, vol. 92, pp. 337-339.

Srinivasan S., Roman N., and Wang D.L. (2006): "Binary and ratio time-frequency masks for robust speech recognition," *Speech Communication*, vol. 48, pp. 1486-1501.

Brungart D.S., Chang P.S., Simpson B.D., and Wang D.L. (2006): "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," *Journal of the Acoustical Society of America*, vol. 120, pp. 4007-4018.

Roman N. and Wang D.L. (2006): "Pitch-based monaural segregation of reverberant speech," *Journal of the Acoustical Society of America*, vol. 120, pp. 4040-4051.

Hu G. and Wang D.L. (2007): "Auditory segmentation based on onset and offset analysis," *IEEE Transactions on Audio, Speech, & Language Processing*, vol. 15, pp. 396-405.

Li Y. and Wang D.L. (2007): "Separation of singing voice from music accompaniment for monaural recordings," *IEEE Transactions on Audio, Speech, and Language Processing*, pp. 1475-1487.

Srinivasan S. and Wang D.L. (2007): "Transforming binary uncertainties for robust speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 2130-2140.

Pedersen M.S., Wang D.L., Larsen J., and Kjems U. (2008): Two-microphone separation of speech mixtures. *IEEE Transactions on Neural Networks*, vol. 19, pp. 475-492.

Wang D.L. and Chang P.S. (2008): An oscillatory correlation model of auditory streaming. *Cognitive Neurodynamics*, vol. 2, pp. 7-19.

Roman N. and Wang D.L. (2008): Binaural tracking of multiple moving sources. *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, pp. 728-739.

Shao Y., Srinivasan S., Jin Z., and Wang D.L. (2008): "A computational auditory scene analysis system for speech segregation and robust speech recognition," *Computer Speech and Language*, in press.

Hu G. and Wang D.L. (2008): "Segregation of unvoiced speech from nonspeech interference," *Journal of the Acoustical Society of America*, in press.

Book chapters

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*, Kluwer Academic, Norwell MA, pp. 181-197.

Brown G.J. and Wang D.L. (2005): "Separation of speech by computational auditory scene analysis," In: Benesty J., Makino S., and Chen J. (ed.), *Speech Enhancement*, Springer, New York, pp. 371-402.

Brown G.J. and Wang D.L. (2006): “Timing is of the essence: Neural oscillator models of auditory grouping,” In: Greenberg S. and Ainsworth W. (ed.), *Listening to Speech: An Auditory Perspective*, Lawrence Erlbaum, Mahwah NJ, pp. 375-392.

Hu G. and Wang D.L. (2006): “An auditory scene analysis approach to monaural speech segregation,” In: Hänsler E. and Schmidt G. (ed.), *Selected Methods for Acoustic Echo and Noise Control*, Springer, Berlin, pp. 485-515.

Wang D.L. and Brown G.J. (2006): “Fundamentals of computational auditory scene analysis,” In: Wang D.L. and Brown G.J. (ed.), *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, Wiley/IEEE Press, Hoboken NJ, pp. 1-44.

Wang D.L. (2006): “Feature-based speech segregation,” In: Wang D.L. and Brown G.J. (ed.), *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, Wiley/IEEE Press, Hoboken NJ, pp. 81-114.

Stern R.M., Brown G.J., and Wang D.L. (2006): “Binaural sound localization,” In: Wang D.L. and Brown G.J. (ed.), *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, Wiley/IEEE Press, Hoboken NJ, pp. 147-185.

Brown G.J. and Wang D.L. (2006): “Neural and perceptual modeling,” In: Wang D.L. and Brown G.J. (ed.), *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, Wiley/IEEE Press, Hoboken NJ, pp. 351-387.

Wang D.L. (2007): “Computational scene analysis,” In: Duch W. and Mandziuk J. (ed.), *Challenges for Computational Intelligence*, Springer, Berlin, pp. 163-191.

Conference papers

Roman N., Wang D.L., and Brown G.J. (2004): "A classification-based cocktail-party processor," in *Advances in Neural Information Processing Systems* (*NIPS* 2003), vol. 16, Cambridge MA: MIT Press.

Roman N. and Wang D.L. (2004): "Binaural sound segregation for multisource reverberant environments," in *Proceedings of ICASSP-04*, pp. II.373-376.

Hu G. and Wang D.L. (2004): “Auditory segmentation based on event detection,” *Proceedings of ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing*.

Srinivasan S., Roman N., and Wang D.L. (2004): “On binary and ratio time-frequency masks for robust speech recognition,” *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*.

Shao Y. and Wang D.L. (2004): "Model-based sequential organization for cochannel speaker identification," *Proceedings of ICSLP-04*.

Wu M. and Wang D.L. (2005): "A two-stage algorithm for enhancement of reverberant speech," in *Proceedings of ICASSP-05*, pp. I.1085-1088.

Hu G. and Wang D.L. (2005): "Separation of fricatives and affricates," in *Proceedings of ICASSP-05*, pp. I.1001-1004.

Li Y. and Wang D.L. (2005): "Detecting pitch of singing voice in polyphonic audio," in *Proceedings of ICASSP-05*, pp. III.17-20.

Srinivasan S. and Wang D.L. (2005): "Robust speech recognition by integrating speech separation and hypothesis testing," in *Proceedings of ICASSP-05*, pp. I.89-92.

Srinivasan S. and Wang D.L. (2005): "Modeling the perception of multitalker speech," *Proceedings of Interspeech*-05, pp. 1265-1268.

Roman N. and Wang D.L. (2005): "A pitch-based model for separation of reverberant speech," *Proceedings of Interspeech*-05, pp. 2109-2112.

Pedersen M.S., Wang D.L., Larsen J., and Kjems U. (2005): "Overcomplete blind source separation by combining ICA and binary time-frequency masking," *Proceedings of IEEE Workshop on Machine Learning for Signal Processing*, pp. 15-20.

Pedersen M.S., Wang D.L., Larsen J., and Kjems U. (2006): "Separating underdetermined convolutive speech mixtures," *Proceedings of International Conference on Independent Component Analysis and Blind Source Separation*, *Springer Lecture Notes in Computer Science*, vol. 3889, pp. 674-681.

Wang D.L. and Hu G. (2006): "Unvoiced speech segregation," *Proceedings of ICASSP*-06, pp. V.953-956.

Shao Y. and Wang D.L. (2006): "Robust speaker recognition using binary time-frequency masks," *Proceedings of ICASSP*-06, pp. I.645-648.

Srinivasan S. and Wang D.L. (2006): "A supervised learning approach to uncertainty decoding for robust speech recognition," *Proceedings of ICASSP*-06, pp. I.297-300.

Roman N., Srinivasan S., and Wang D.L. (2006): "Speech recognition in multisource reverberant environments with binaural inputs," *Proceedings of ICASSP*-06, pp. I.309-312.

Srinivasan S., Shao Y., Jin Z., and Wang D.L. (2006): "A computational auditory scene analysis system for robust speech recognition," *Proceedings of Interspeech*-06, pp. 73-76.

Li Y. and Wang D.L. (2006): "Singing voice separation from monaural recordings," *Proceedings of the International Symposium on Music Information Retrieval* (ISMIR), pp. 176-179.

Li Y. and Wang D.L. (2007): "Pitch detection in polyphonic music using instrument tone models," *Proceedings of ICASSP*-07, pp. II.481-484.

Shao Y., Srinivasan S., and Wang D.L. (2007): "Incorporating auditory feature uncertainties in robust speaker identification," *Proceedings of ICASSP*-07, pp. IV.277-280.

Srinivasan S., Roman N., and Wang D.L. (2007): "Exploiting uncertainties for binaural speech recognition," *Proceedings of ICASSP*-07, pp. IV.789-792.

Jin Z. and Wang D.L. (2007): "A supervised learning approach to monaural segregation of reverberant speech," *Proceedings of ICASSP*-07, pp. IV.921-924.

Li Y. and Wang D.L. (2008): "Musical sound separation using pitch-based labeling and binary time-frequency masking," *Proceedings of ICASSP*-08, in press.

Li Y. and Wang D.L. (2008): "On the optimality of ideal binary time-frequency masks," *Proceedings of ICASSP*-08, in press.

Shao Y. and Wang D.L. (2008): "Robust speaker identification using auditory features and computational auditory scene analysis," *Proceedings of ICASSP*-08, in press.

Report of Inventions and Subcontracts

DeLiang Wang
(Principal Investigator)

March 2008

Department of Computer Science & Engineering and Center for Cognitive Science
The Ohio State University

The project, entitled "Monaural Speech Segregation by Integrating Primitive and Schema-based Analysis" (FA9550-04-1-0117), was funded by the Air Force Office of Scientific Research from February 2004 to December 2007.

This is to report that a provisional patent application entitled "A method for accurate pitch estimation and voice separation," has been filed as a result of this AFOSR grant.

Appendix 1. Executive Summary of Nicoleta Roman's Ph.D. Dissertation

At a cocktail party, listeners can selectively attend to a single voice and filter out other interferences. This perceptual ability has motivated a new field of study known as computational auditory scene analysis (CASA) which aims to build speech separation systems that incorporate auditory principles. The psychological process of figure-ground segregation suggests that the target signal should be segregated as foreground while the remaining stimuli are treated as background. Accordingly, the computational goal of CASA should be to estimate an ideal time-frequency (T-F) binary mask, which selects the target if it is stronger than the interference in a local T-F unit. This dissertation investigates four aspects of CASA processing: location-based speech segregation, binaural tracking of multiple moving sources, binaural sound segregation in reverberation, and monaural segregation of reverberant speech. For localization, the auditory system utilizes the interaural time difference (ITD) and interaural intensity difference (IID) between the ears. It is observed that within a narrow frequency band, modifications to the relative strength of the target source with respect to the interference trigger systematic changes for ITD and IID resulting in a characteristic clustering. Consequently, this dissertation proposes a supervised learning approach to estimate the ideal binary mask. A systematic evaluation shows that the resulting system produces masks very close to the ideal binary ones and large speech intelligibility improvements.

In realistic environments, source motion requires consideration. Binaural cues are strongly correlated with locations in T-F units dominated by one source resulting in channel-dependent conditional probabilities. Consequently, the dissertation proposes a multi-channel integration method of these probabilities in order to compute the likelihood function in a target space. Finally, a hidden Markov model is employed for forming continuous tracks and automatically detecting the number of active sources. Reverberation affects the ITD and IID cues. A binaural segregation system is therefore proposed that combines target cancellation through adaptive filtering and a binary decision rule to estimate the ideal binary mask. A major advantage of the proposed system is that it imposes no restrictions on the interfering sources. Quantitative evaluations show that our system outperforms related beamforming approaches.

Psychoacoustic evidence suggests that monaural processing play a vital role in segregation. It is known that reverberation smears the harmonicity of speech signals. This dissertation therefore proposes a two-stage separation system that combines inverse filtering of target room impulse response with pitch-based segregation. As a result of the first stage, the harmonicity of a signal arriving from target direction is partially restored while signals arriving from other locations are further smeared, and this leads to improved segregation and considerable signal-to-noise ratio gains.

Appendix 2. Executive Summary of Guoning Hu's Ph.D. Dissertation

In a natural environment, speech often occurs simultaneously with acoustic interference. Many applications, such as automatic speech recognition and telecommunication, require an effective system that segregates speech from interference in the monaural (one-microphone) situation. While this task of monaural speech segregation has proven to be very challenging, human listeners show a remarkable ability to segregate an acoustic mixture and attend to a target sound, even with one ear. This perceptual process is called auditory scene analysis (ASA). Research in ASA has inspired considerable effort in constructing computational ASA (CASA) based on ASA principles. Current CASA systems, however, face a number of challenges in monaural speech segregation.

This dissertation presents a systematic and extensive effort in developing a CASA system for monaural speech segregation that addresses several major challenges. The proposed system consists of four stages: Peripheral analysis, feature extraction, segmentation, and grouping. In the first stage, the system decomposes the auditory scene into a time-frequency representation via bandpass filtering and time windowing. The second stage extracts auditory features corresponding to ASA cues, such as periodicity, amplitude modulation, onset and offset. In the third stage, the system segments an auditory scene based on a multiscale analysis of onset and offset. The last stage includes an iterative algorithm that simultaneously estimates the pitch of a target utterance and segregates the voiced target based on a pitch estimate. Finally, our system sequentially groups voiced and unvoiced portions of the target speech for non-speech interference, and this grouping task is performed using feature-based classification.

Systematic evaluation shows that the proposed system extracts a majority of target speech without including much interference. Extensive comparisons demonstrate that the system has substantially advanced the state-of-the-art performance in voiced speech segregation, and represents the first systematic study of unvoiced speech segregation.

Appendix 3. Executive Summary of Soundararajan Srinivasan's Ph.D. Dissertation

Speech perception studies indicate that robustness of human speech recognition is primarily due to our ability to segregate a target sound source from other interferences. This perceptual process of auditory scene analysis (ASA) is of two types, primitive and schema-driven. This dissertation investigates several aspects of integrating computational ASA (CASA) and automatic speech recognition (ASR). While bottom-up CASA are used as front-end for ASR to improve its robustness, ASR is used to provide top-down information to enhance primitive segregation.

Listeners are able to restore masked phonemes by utilizing lexical context. We present a schema-based model for phonemic restoration. The model employs missing-data ASR to decode masked speech and activates word templates via dynamic time warping. A systematic evaluation shows that the model restores both voiced and unvoiced phonemes with a high spectral quality.

Missing-data ASR requires a binary mask from bottom-up CASA that identifies speech-dominant time-frequency regions of a noisy mixture. We propose a two-pass system that performs segregation and recognition in tandem. First, an n-best lattice, consistent with bottom-up speech separation, is generated. Second, the lattice is re-scored using a model-based hypothesis test to improve mask estimation and recognition accuracy concurrently.

By combining CASA and ASR, we present a model that simulates listeners' ability to attend to a target speaker when degraded by energetic and informational masking. Missing-data ASR is used to account for energetic masking and the output degradation of CASA is used to model informational masking. The model successfully simulates several quantitative aspects of listener performance.

The degradation in the output of CASA-based front-ends leads to uncertain ASR inputs. We estimate feature uncertainties in the spectral domain and transform them into the cepstral domain via nonlinear regression. The estimated uncertainty substantially improves recognition accuracy.

We also investigate the effect of vocabulary size on conventional and missing-data ASRs. Based on binaural cues, for conventional ASR, we extract the speech signal using a Wiener filter and for missing-data ASR, we estimate a binary mask. We find that while missing-data ASR outperforms conventional ASR on a small vocabulary task, the relative performance reverses on a larger vocabulary task.

Appendix 4. Executive Summary of Yang Shao's Ph.D. Dissertation

A human listener has the ability to follow a speaker's voice while others are speaking simultaneously. In particular, the listener can organize the time-frequency (T-F) energy of the same speaker into a single stream. This aspect of auditory perception is termed auditory scene analysis (ASA). ASA comprises two organization processes: segmentation and grouping. Segmentation decomposes the auditory scene into T-F segments. Grouping combines the segments from the same source into a single perceptual stream. Within the grouping process, simultaneous organization integrates segments that overlap in time, and sequential organization groups segments across time.

Inspired by ASA research, computational auditory scene analysis (CASA) aims to organize sound based on ASA principles. CASA systems seek to segregate target speech from a complex auditory scene. However, almost all the existing systems focus on simultaneous organization. This dissertation presents a systematic effort on sequential organization. The goal is to organize T-F segments from the same speaker that are separated in time into a single stream. This study proposes to employ speaker characteristics for sequential organization.

This study first explores bottom-up methods for sequential grouping. Subsequently, a speaker-model-based sequential organization framework is proposed and shown to yield better grouping performance than feature-based methods. Specifically, a computational objective is derived for sequential grouping in the context of cochannel speaker recognition. Cochannel speech occurs when two utterances are transmitted in a single communication channel. This formulation leads to a grouping system that searches for the optimal grouping of separated speech segments. To reduce search space and computation time, a hypothesis pruning method is then proposed and it achieves performance close to that of exhaustive search. Systematic evaluations show that the proposed system improves not only grouping performance but also speech recognition accuracy.

The model-based grouping system is then extended to handle multi-talker as well as non-speech intrusions using generic models. This generalization is shown to function well regardless of interference types and the number of interfering sources. The grouping system is further extended to deal with noisy inputs from unknown speakers. Specifically, it employs a speaker quantization method that extracts representative speakers from a large speaker space and performs sequential grouping using obtained generic models. The resulting grouping performance is only moderately lower than that with known speaker models.

In addition to sequential grouping, this dissertation presents a systematic effort in robust speaker recognition. A novel usable speech extraction method is proposed that significantly improves recognition performance. Then, missing-data recognition is combined with the use of CASA as a front-end processor. Substantial performance improvements are achieved in speaker recognition evaluations under various noisy conditions. Finally, a general solution is proposed for robust speaker recognition in the presence of additive noise. Novel speaker features are derived from auditory filtering and cepstral analysis, and are used in conjunction with an uncertainty decoder that accounts for mismatch introduced in front-end processing. Systematic evaluations show that the proposed system achieves significant performance improvement over the use of typical speaker features and a state-of-the-art robust front-end processor for noisy speech.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE: 3/21/08	3. REPORT TYPE AND DATES COVERED
		Final technical report

4. TITLE AND SUBTITLE

Monaural speech segregation by integrating primitive and schema-based analysis

5. FUNDING NUMBERS

FA9550-04-1-0117

6. AUTHOR(S)

DeLiang Wang: Principal Investigator

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

The Ohio State University
Research Foundation
1960 Kenny Road
Columbus, OH 43210-1063

8. PERFORMING ORGANIZATION REPORT NUMBER

746317

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

AFOSR
Attn: Dr. Willard Larkin,
AFOSR/NL, Room 713
4015 Wilson Blvd.
Arlington, VA 22203-1945

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution is limited.

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 Words)***

The natural auditory environment typically contains multiple simultaneous events. A remarkable feat of auditory perception is the ability to disentangle the acoustic mixture and group the components of the same event into a stream. This fundamental aspect of human perception is called auditory scene analysis, which is divided into a primitive (bottom-up) process and a schema-based (top-down) process. A particularly important task of auditory scene analysis is monaural segregation of speech from interfering sounds. This project seeks to develop an auditory scene analysis approach to monaural speech segregation. Guided by perceptual data, the project will integrate primitive and schema-based analysis. Starting with accepted models for cochlear filtering and hair cell transduction, the proposed system has several stages of computation, including auditory segmentation, multi-pitch tracking, event detection, and labeling of time-frequency units based on pitch and onset. Labeled segments will be grouped into short streams via simultaneous organization, and sequential organization will further link successive streams across intermittent sections of unvoicing, silence, or other interference. Then a schema-based process will be employed to group voiced and unvoiced sections of the same utterance. The resulting system will be tested using real recordings of natural speech and interfering sounds.

14. SUBJECT TERMS

Auditory scene analysis, speech segregation, primitive analysis, schema-based analysis, computational audition

15. NUMBER OF PAGES

23

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. Z39-18
298-102