AFRL-AFOSR-VA-TR-2016-0256

# Speech Segregation based on Binary Classification

**Deliang Wang**
**OHIO STATE UNIVERSITY THE**
**1960 KENNY RD**
**COUMBUS, OH 432101016**

**07/15/2016**
**Final Report**



Air Force Research Laboratory
AF Office Of Scientific Research (AFOSR)/RTB2

Arlington, Virginia 22203
Air Force Materiel Command

# REPORT DOCUMENTATION PAGE

*Form Approved*
*OMB No. 0704-0188*



**PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.**

| 1. REPORT DATE *(DD-MM-YYYY)* | 2. REPORT TYPE | 3. DATES COVERED *(From - To)* |
|---|---|---|
| 15-07-2016 | Final performance report | 5/2012 - 4/2016 |

**4. TITLE AND SUBTITLE**
Speech Segregation based on Binary Classification

**5a. CONTRACT NUMBER**

**5b. GRANT NUMBER**
FA9550-12-1-0130

**5c. PROGRAM ELEMENT NUMBER**

**6. AUTHOR(S)**
DeLiang Wang (Principal Investigator)

**5d. PROJECT NUMBER**

**5e. TASK NUMBER**

**5f. WORK UNIT NUMBER**

**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**

**9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)**
Dr. Patrick Bradshaw
AFOSR
875 North Randolph Street 4027
Arlington, VA 22203

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

**11. SPONSOR/MONITOR'S REPORT NUMBER(S)**

**12. DISTRIBUTION/AVAILABILITY STATEMENT**
DISTRIBUTION A: Distribution approved for public release.

**13. SUPPLEMENTARY NOTES**

**14. ABSTRACT**
This AFOSR project aimed to develop a classification-based approach to address the speech segregation challenge. The supervised approach is in sharp contrast to traditional speech segregation approaches. There are four major accomplishments made in this project. First, a supervised approach based on neural networks was developed to perform pitch tracking in very noisy conditions. Second, different training targets were examined for supervised speech segregation, leading to the adoption of the ideal ratio mask (IRM). A subsequent listening evaluation shows increased intelligibility in noise for human listeners following IRM estimation. Third, an algorithm was proposed to recognize speakers in cochannel (two-talker) conditions. This algorithm uses deep neural networks for cochannel speaker identification, and achieves the state-of-the-art results in both anechoic and reverberant conditions. Fourth, a spectral mapping method was developed to address the issue of robustness to room reverberation. This supervised method learns a mapping from the magnitude spectrogram of reverberant speech to that of anechoic speech, as well as from the spectrogram of reverberant-noisy speech to that of anechoic-clean speech.

**15. SUBJECT TERMS**
Binary classification, time-frequency masking, supervised speech segregation, speech intelligibility, room reverberation

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT | b. ABSTRACT | c. THIS PAGE | | | DeLiang Wang |
| Unclassified | Unclassified | Unclassified | Unclassified | 21 | 19b. TELEPHONE NUMBER *(Include area code)* 614-292-6827 |

Reset

**Standard Form 298** (Rev. 8/98)
Prescribed by ANSI Std. Z39.18
Adobe Professional 7.0

DISTRIBUTION A: Distribution approved for public release.

## INSTRUCTIONS FOR COMPLETING SF 298

**1. REPORT DATE.** Full publication date, including day, month, if available. Must cite at least the year and be Year 2000 compliant, e.g. 30-06-1998; xx-06-1998; xx-xx-1998.

**2. REPORT TYPE.** State the type of report, such as final, technical, interim, memorandum, master's thesis, progress, quarterly, research, special, group study, etc.

**3. DATES COVERED.** Indicate the time during which the work was performed and the report was written, e.g., Jun 1997 - Jun 1998; 1-10 Jun 1996; May - Nov 1998; Nov 1998.

**4. TITLE.** Enter title and subtitle with volume number and part number, if applicable. On classified documents, enter the title classification in parentheses.

**5a. CONTRACT NUMBER.** Enter all contract numbers as they appear in the report, e.g. F33615-86-C-5169.

**5b. GRANT NUMBER.** Enter all grant numbers as they appear in the report, e.g. AFOSR-82-1234.

**5c. PROGRAM ELEMENT NUMBER.** Enter all program element numbers as they appear in the report, e.g. 61101A.

**5d. PROJECT NUMBER.** Enter all project numbers as they appear in the report, e.g. 1F665702D1257; ILIR.

**5e. TASK NUMBER.** Enter all task numbers as they appear in the report, e.g. 05; RF0330201; T4112.

**5f. WORK UNIT NUMBER.** Enter all work unit numbers as they appear in the report, e.g. 001; AFAPL30480105.

**6. AUTHOR(S).** Enter name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. The form of entry is the last name, first name, middle initial, and additional qualifiers separated by commas, e.g. Smith, Richard, J, Jr.

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES).** Self-explanatory.

**8. PERFORMING ORGANIZATION REPORT NUMBER.** Enter all unique alphanumeric report numbers assigned by the performing organization, e.g. BRL-1234; AFWL-TR-85-4017-Vol-21-PT-2.

**9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES).** Enter the name and address of the organization(s) financially responsible for and monitoring the work.

**10. SPONSOR/MONITOR'S ACRONYM(S).** Enter, if available, e.g. BRL, ARDEC, NADC.

**11. SPONSOR/MONITOR'S REPORT NUMBER(S).** Enter report number as assigned by the sponsoring/ monitoring agency, if available, e.g. BRL-TR-829; -215.

**12. DISTRIBUTION/AVAILABILITY STATEMENT.** Use agency-mandated availability statements to indicate the public availability or distribution limitations of the report. If additional limitations/ restrictions or special markings are indicated, follow agency authorization procedures, e.g. RD/FRD, PROPIN, ITAR, etc. Include copyright information.

**13. SUPPLEMENTARY NOTES.** Enter information not included elsewhere such as: prepared in cooperation with; translation of; report supersedes; old edition number, etc.

**14. ABSTRACT.** A brief (approximately 200 words) factual summary of the most significant information.

**15. SUBJECT TERMS.** Key words or phrases identifying major concepts in the report.

**16. SECURITY CLASSIFICATION.** Enter security classification in accordance with security classification regulations, e.g. U, C, S, etc. If this form contains classified information, stamp classification level on the top and bottom of this page.

**17. LIMITATION OF ABSTRACT.** This block must be completed to assign a distribution limitation to the abstract. Enter UU (Unclassified Unlimited) or SAR (Same as Report). An entry in this block is necessary if the abstract is to be limited.

# Final AFOSR Project Performance Report

DeLiang Wang

(Principal Investigator)

*The Ohio State University*

July 2016

This PI was awarded the AFOSR grant "Speech segregation based on binary classification" (Grant No.: FA9550-12-1-0130). The project was funded for the period of 5/1/12 to 4/30/16 with the total amount of $932,284. This report summarizes the progress made throughout the 4-year project period.

## 1. RESEARCH PROGRESS

One of the biggest challenges in speech and audio processing is speech segregation, which is the problem of separating target speech from its acoustic background. The goal of this AFOSR project was to develop a speech segregation system that can potentially improve speech intelligibility in noise for human listeners. Motivated by the perceptual principles of auditory scene analysis and the speech intelligibility studies of ideal time-frequency masking, the project sought to develop a classification-based approach to tackle the speech segregation challenge. The supervised approach is in sharp contrast to traditional speech segregation approaches.

Consistent with the stated objectives, the project made substantial advances along the following four directions. First, we have developed a supervised approach to pitch tracking in very noisy conditions on the basis of neural networks. Second, we have investigated different training targets for supervised speech segregation, leading to the adoption of the ideal ratio mask (IRM). A subsequent listening evaluation shows increased intelligibility in noise for human listeners following IRM estimation. Third, we have proposed an algorithm for recognizing speakers in cochannel (two-talker) conditions. This algorithm uses deep neural networks (DNNs) for cochannel speaker identification, and achieves the state-of-the-art results in both anechoic and reverberant conditions. Fourth, we have developed a spectral mapping method to address the issue of robustness to room reverberation. This supervised method learns a mapping from the magnitude spectrogram of reverberant speech to that of anechoic speech, as well as from the spectrogram of reverberant-noisy speech to that of anechoic-clean speech. Although not highlighted in this report, this project has also contributed to considerable progress in the development of an unsupervised approach to cochannel speech separation, generalization of the binary classification approach, and the study of acoustic-phonetic features other than pitch for the purpose of supervised speech segregation.

The major findings along each of above four directions are described in more detail in the following subsections.

### 1.1 Neural Network Based Pitch Tracking in Very Noisy Speech

Pitch, or fundamental frequency (F0), is one of the most important characteristics of speech signals. A pitch tracking algorithm robust to background interference is critical to many applications, including speaker identification and speech separation. Although pitch tracking has been studied for decades, it is still challenging to estimate pitch from speech in the presence of strong noise, where the harmonic structure of speech is severely corrupted. In this work, we perform pitch estimation or tracking using supervised learning, where probabilistic pitch states are directly learned from noisy speech data. More specifically, we proposed two alternative neural networks modeling pitch state distribution given noisy observations. The first one is a feedforward DNN that is trained on static frame-level acoustic features. The second one is a recurrent deep neural network (RNN) that is trained on sequential frame-level features and capable of learning temporal dynamics. Both DNNs and RNNs produce accurate probabilistic outputs of pitch states, which are then connected into pitch contours by Viterbi decoding as part of a hidden Markov model (HMM).

The proposed pitch tracking algorithms first extract spectral domain features in each frame; specifically, we compute the log power spectrogram and then normalize it to the long-term speech spectrum in order to attenuate background noise. To simplify the representation of pitch, we quantize the plausible F0 range into $M$ frequency bins, corresponding to $M$ pitch states. DNNs and RNNs are then employed to compute the posterior probability of the pitch state for each frequency bin. To train a DNN, each training sample is the feature vector extracted from the current time frame (plus its neighboring frames), and the target is an ($M$+1)-dimensional vector of the pitch states whose element is 1 if the ground-truth pitch falls into the corresponding frequency bin, and 0 otherwise. One more state is used to designate unvoicing.

An RNN is a natural extension of a feedforward DNN. In an RNN, the depth comes from not only multiple hidden layers employed in a DNN but also unfolding layers through time. As a result, RNN is capable of capturing the long-term dependencies through connections between hidden layers. These considerations motivated us to use RNN to model pitch dynamics. One of the key challenges for using RNN is that training with long-term dependencies can be quite difficult. In our study, we use a classic RNN and learn the model with truncated backpropagation through time.

With the posterior probability distribution at each time frame learned by a DNN or RNN, Viterbi decoding utilizes the likelihood and the transition probability to generate an optimal sequence of pitch states for a sentence. We convert the sequence of pitch states to a sequence of pitch frequencies and then use a 3-point moving average for smoothing to generate final pitch contours.

Figure 1 illustrates pitch tracking results using the proposed methods. The example is a female utterance from the TIMIT corpus: “Readiness exercises are almost continuous”, mixed with a factory noise at -5 dB SNR. Fig. 1(a) and (b) show the spectrograms of clean speech and noisy speech from 0 to 2000 Hz respectively. Comparing Fig. 1(b) to Fig. 1(a), the harmonics are severely corrupted by noise, leading to a major difficulty in pitch tracking. Fig. 1(c) shows the ground-truth pitch states extracted from the clean

speech using Praat, a standard pitch tracking algorithm for clean speech. As shown in the figure, Praat even makes a few pitch doubling or halving errors at around 160 ms and 280 ms. But since these errors are not serious, we do not correct them and still treat them as the ground-truth. The probabilistic outputs of the DNN and the RNN are shown in Figs. 1(d) and 1(e), respectively. Comparing to Fig. 1(c), the probabilities of the correct pitch states dominate in most time frames in both Figs. 1(d) and (e), demonstrating that the neural networks successfully predict pitch states from noisy speech. In some time frames (e.g., 100 ms to 120 ms), the RNN yields better probabilistic outputs than the DNN, because the RNN can better capture the temporal context and its outputs are smoother than those of the DNN. Figs. 1(f) and (g) show extracted pitch contours after Viterbi decoding. Both the DNN and the RNN produce accurate pitch contours. A few errors occur from 260 ms to 280 ms due to severe interference.

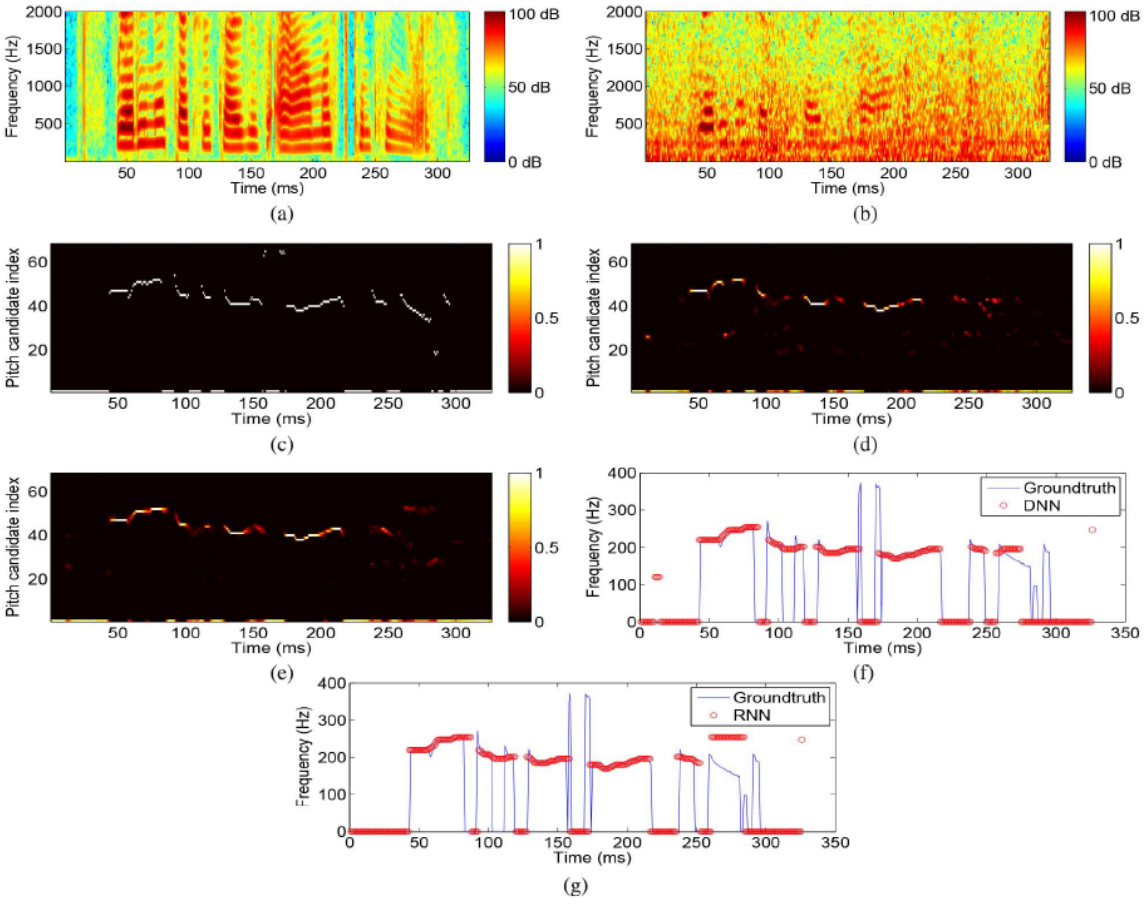


**Figure 1**. Neural network based pitch tracking. Noisy speech is a female utterance mixed with factory noise in -5 dB SNR. (a) Spectrogram of clean speech from 0 to 2000 Hz. (b) Spectrogram of noisy speech from 0 to 2000 Hz. (c) Ground-truth pitch states. In each time frame, the probability of a pitch state is 1 if it corresponds to the ground-truth pitch and 0 otherwise. (d) Probabilistic outputs from the DNN. (e) Probabilistic outputs from the RNN. (f) DNN based pitch contours. The circles denote the generated pitches, and solid lines denote the ground-truth. (g) RNN based pitch contours.

We have systematically evaluated our pitch tracking approach and compared it with four leading algorithms, two of them supervised and the other two unsupervised. The evaluation results demonstrate that the proposed pitch tracking algorithms are robust to different noise conditions and can even be applied to reverberant speech. The proposed approach significantly outperforms the four comparison algorithms. Furthermore, our system generates strong results across multiple unseen conditions, including different speakers, SNRs, noises, and room impulse responses. A paper describing our neural network based pitch tracking algorithms was published in a 2014 paper by K. Han and D.L. Wang, entitled "Neural network based pitch tracking in very noisy speech," in *IEEE/ACM Transactions on Audio, Speech and Language Processing*. More details about this work can be found in this publication.

### 1.2 Analysis of Training Targets for Supervised Speech Separation

In the simplest form of supervised speech separation, acoustic features are extracted from noisy mixtures to train a supervised learning algorithm, such as a DNN. Traditionally, the training target (or the desired signal) is set to the ideal binary mask (IBM), which turns speech separation into a binary classification problem – a well studied machine learning task. Although the IBM is the optimal binary mask, it may not necessarily be the best target for training and prediction. In addition, speech quality is a persistent issue for binary masking. The supervised separation framework, however, is not limited to binary targets. So what training targets are appropriate for supervised speech separation? This is clearly an important question with potentially major implications for separation performance. We have addressed this question systematically.

In our study, we have analyzed a number of different training targets, including the IBM, the target binary mask (TBM), the IRM, the short-time Fourier transform spectral magnitude (FFT-MAG) and its corresponding mask (FFT-MASK), and the gammatone-frequency power spectrum (GF-POW). An illustration of different training targets is shown in Figure 2. Among them, the IRM is defined as

$$IRM(t,f) = \sqrt{\frac{S^2(t,f)}{S^2(t,f) + I^2(t,f)}}$$

where $S^2(t,f)$ and $I^2(t,f)$ denote the target speech energy and interference energy, respectively, in a given time-frequency (T-F) unit. The IRM can be viewed as the square root of the Wiener filter, which is the optimal estimator of the power spectrum of the target speech signal. With the IRM as the training target, the learning task becomes a problem of regression (or function approximator), rather than classification associated with the IBM.

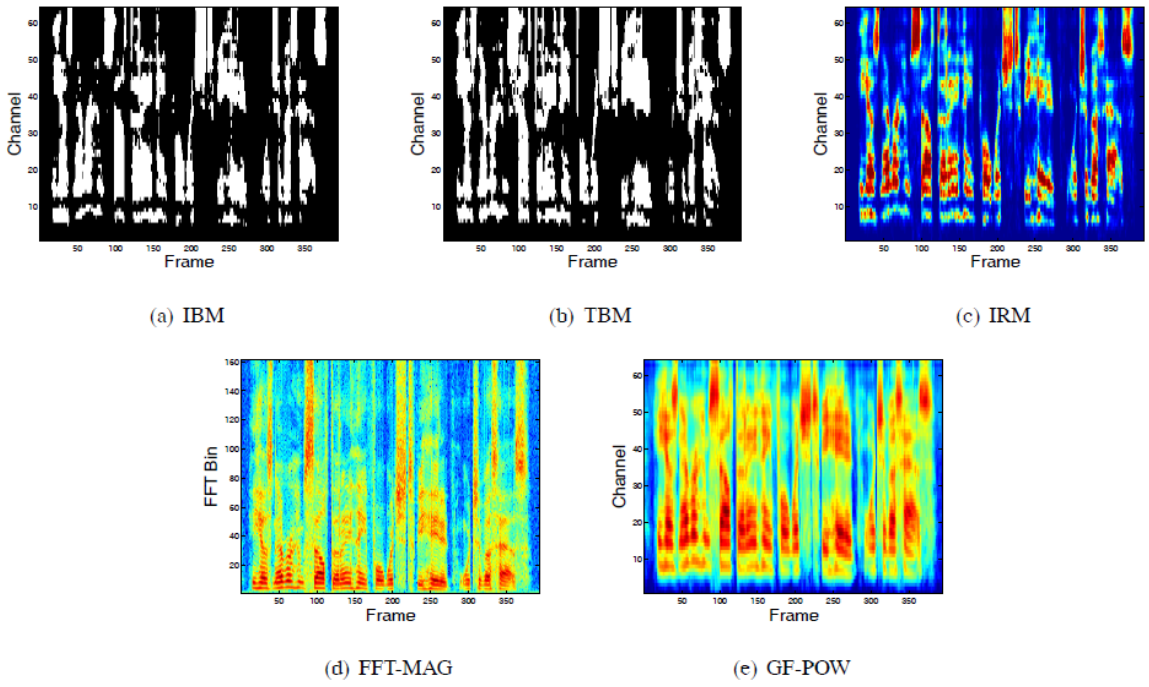


**Figure 2**. Various training targets for a TIMIT utterance mixed with a factory noise at -5 dB SNR.

Throughout our analysis, we use a fixed set of complementary features and a fixed DNN as the discriminative learning machine. For evaluation metrics, besides SNR, we use the Short-Time Objective Intelligibility (STOI) to measure predicted speech intelligibility and the Perceptual Evaluation of Speech Quality (PESQ) to measure objective speech quality. The STOI score ranges from 0 to 1, and the PESQ score from -0.5 to 4.5. Both STOI and PESQ are shown to be highly correlated to human speech perception.

Table 1 shows the comparisons between different targets when input SNR is -5 dB; the performance trends at other input SNR levels are similar. Regardless of the target of choice, the supervised speech separation framework provides substantial improvements compared to unprocessed mixtures. For the two binary masking targets, the IBM appears to be a better choice than the TBM, probably because the TBM is defined by completely ignoring the noise characteristics in the mixture.

Going from binary masking to ratio masking improves all objective metrics, as exemplified by the performance of the IRM. Although predicting the IRM achieves slightly better or equal STOI results than predicting the IBM, the IRM seems to be especially beneficial for improving objective speech quality. For example, the PESQ score improves by 0.64 and 0.76 in the engine noise compared to the IBM and unprocessed mixtures, respectively.

Interestingly, FFT-MASK produces comparable STOI and PESQ results to the IRM, but significantly better results than FFT-MAG. This contrast with FFT-MAG appears surprising, considering that the DNNs in both cases are essentially trained to estimate the

same underlying target: the clean magnitude. Our further analysis shows that the estimation of spectral magnitudes magnifies estimation errors, and the one-to-one mapping in T-F mask estimation is easier to learn than the many-to-one mapping in spectral magnitude estimation.

We have also compared with recent methods in supervised NMF (nonnegative matrix factorization) and speech enhancement, and the comparison shows clear performance advantages of DNN-based supervised speech separation. Our analysis of training targets was published in a 2014 paper by Y. Wang, A. Narayana and D.L. Wang, entitled "On training targets for supervised speech separation," in *IEEE/ACM Transactions on Audio, Speech and Language Processing*.

**Table 1**. Separation performance comparisons among various training targets for speech utterances mixed with five noises at -5 dB SNR. The best STOI and PESQ scores for each noise are highlighted by boldface.

| Target/System | Factory1 | | | Babble | | | SSN | | | Engine | | | Oproom | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | STOI | PESQ | SNR | STOI | PESQ | SNR | STOI | PESQ | SNR | STOI | PESQ | SNR | STOI | PESQ | SNR |
| Mixture | 0.54 | 1.29 | -5.00 | 0.55 | 1.42 | -5.00 | 0.57 | 1.48 | -5.00 | 0.57 | 1.41 | -5.00 | 0.59 | 1.40 | -5.00 |
| IBM | 0.66 | 1.49 | 6.63 | 0.63 | 1.50 | 3.98 | 0.72 | 1.45 | 8.71 | 0.78 | 1.53 | 13.24 | 0.77 | 1.81 | 12.24 |
| TBM | 0.65 | 1.33 | 5.19 | 0.62 | 1.32 | 3.08 | 0.72 | 1.45 | 8.71 | 0.77 | 1.52 | 6.16 | 0.76 | 1.60 | 6.38 |
| IRM | 0.67 | 1.75 | 8.27 | 0.63 | 1.64 | 4.39 | 0.73 | **1.87** | 10.81 | **0.80** | **2.17** | 15.66 | **0.79** | **2.19** | 15.33 |
| FFT-MAG | 0.66 | 1.73 | 5.45 | 0.62 | 1.50 | 3.80 | 0.72 | 1.76 | 5.18 | 0.76 | 2.02 | 6.09 | 0.74 | 2.01 | 5.84 |
| FFT-MASK | **0.68** | 1.77 | 7.59 | **0.65** | **1.65** | 5.52 | **0.74** | **1.87** | 7.58 | 0.78 | 2.16 | 9.73 | 0.77 | 2.15 | 9.89 |
| GF-POW | 0.67 | **1.80** | 8.23 | 0.62 | 1.63 | 5.98 | 0.72 | 1.85 | 8.62 | 0.76 | 2.06 | 9.83 | 0.74 | 2.14 | 9.31 |

A tangible benefit of this training target analysis is a recent successful intelligibility test conducted on both normal-hearing (NH) and hearing-impaired (HI) listeners. The speech segregation algorithm tested here employs a DNN to estimate the IRM, rather than the IBM as done in an earlier test (Healy et al., 2013; see Section 2.4). Besides different training targets, the current algorithm uses a noise perturbation technique to generate new noise samples to expand the training set. Unlike Healy et al.'s 2013 study where noise samples were drawn from the same short segments of two noises, in the current study, we evaluated the algorithm on noise samples drawn from novel segments of nonstationary noises.

The DNN-based IRM estimator was used to segregate IEEE sentences from two nonstationary noises: multi-talker babble and a cafeteria noise. Each noise is 10 minutes long, and the first 8 minutes were used in training and the last two minutes in testing to ensure no overlap between training and test noise segments. The speech utterances used in the test were also different from those used in training. Ten HI listeners and ten NH listeners participated in this experiment. The tested SNRs were 0 and 5 dB for the HI listeners, and -5 and -2 dB for the NH listeners.

For the babble noise, the average recognition improvement from algorithm processing was 27.8 and 44.4 percentage points for the HI listeners (at 5 and 0 dB SNR, respectively) and 21.5 and 26.8 percentage points for the NH listeners (at -2 and -5 dB SNR). For the cafeteria noise, the average improvement from algorithm processing was 18.2 and 26.9

percentage points for the HI listeners (at 5 and 0 dB SNR). Unlike the HI listeners, the NH listeners did not benefit from algorithm processing in the cafeteria noise, likely because NH listeners typically benefit less from algorithm processing due to their already remarkable segregation ability.

The current study demonstrates that a supervised segregation algorithm designed to estimate the IRM can successfully generalize to novel segments of nonstationary noises and produce substantial speech intelligibility improvements for HI listeners, as well as NH listeners. This test was published in a 2015 paper by E. Healy et al., entitled "An algorithm to increase speech intelligibility for hearing-impaired listeners in novel segments of the same noise type," in the *Journal of the Acoustical Society of America*.

### 1.3 Cochannel Speaker Identification in Reverberant Conditions

Speaker identification (SID) in cochannel speech, where two speakers are talking simultaneously over a single recording channel, is a well-recognized challenge. Previous studies address this problem in the anechoic environment in the Gaussian mixture model (GMM) framework. On the other hand, cochannel SID in reverberant conditions has not been addressed at all. Partly driven by the need for speech separation in multi-talker conditions, we have investigated cochannel SID in both anechoic and reverberant conditions. In fact, we are the first to employ DNN for cochannel SID.

We formulate cochannel SID as a discriminative learning problem, where we directly learn a mapping from cochannel observations to the corresponding speaker identities. Thus, cochannel SID is treated as a multi-class classification problem and DNN is employed as the learning machine. More specifically, our method trains a DNN using frame level features. The output layer has the same number of nodes as speakers; for cochannel SID, only the two nodes corresponding to the underlying speakers have non-zero training labels. During testing, the frame level outputs are aggregated across time to generate the final output.

To encode temporal context, we splice a window of 11 frames of log-spectral features to train the DNN. The training target of the DNN is the true speaker identities. We use soft training labels where the two underlying speakers each have a probability of generating the current frame. The sum of their probabilities equals one, whereas the other speakers have zero probabilities. We compare frame level energy of two speakers and use their ratio for the soft labels. More specifically, we construct the IBM during training, and the frame level energy of each speaker is calculated from the mixture cochleagram according to the IBM.

To evaluate cochannel SID performance, we randomly select 100 speakers from the 2008 NIST Speaker Recognition Evaluation (SRE) dataset (short2 part of the training set). The telephone conversation excerpt of each speaker is roughly 5 minutes long. Large chunks of silence in the excerpt are removed. Then we divide the recording into 5 second pieces. Two pieces with the highest energies are used for tests in order to provide sufficient speech information. The rest is used for training. Overall each speaker has about 20 training utterances.

Figure 3 shows the systematic cochannel SID results with respect to the size of the

speaker set. It should be noted that, in the accuracy results of Fig. 3, SID is considered correct only when both speakers are correctly identified. The cochannel GMM results are included in the figure as a strong baseline for comparison. For evaluation with reverberant cochannel speech, speech utterances from the NIST dataset are convolved with room impulse responses with the reverberation time of 600 ms.

There are a number of observations from Figure 3. GMM and DNN-based approaches both work very well with the small speaker set of 10, even in the reverberant conditions. Both approaches show a decline of performance with the increase of speaker set size. This is to be expected as SID is more prone to error with more speaker models to choose from. However, reverberation exacerbates the degradation. Overall, the DNN-based approach declines at a much slower pace than the GMM-based approach in the anechoic condition, indicating better scalability to speaker set size. However, none of them scale well in the reverberant conditions, although the DNN-based approach still holds a sizeable advantage over the GMM-based method.

With its excellent performance of cochannel SID, DNN represents a promising direction to pursue noise-robust and reverberation-robust SID, which should play an important role in speech separation, particularly in multi-talker conditions. This SID contribution was published in a 2015 paper by X. Zhao, Y. Wang, and D.L. Wang, entitled "Cochannel speaker identification in anechoic and reverberant conditions," in *IEEE/ACM Transactions on Audio, Speech and Language Processing*.

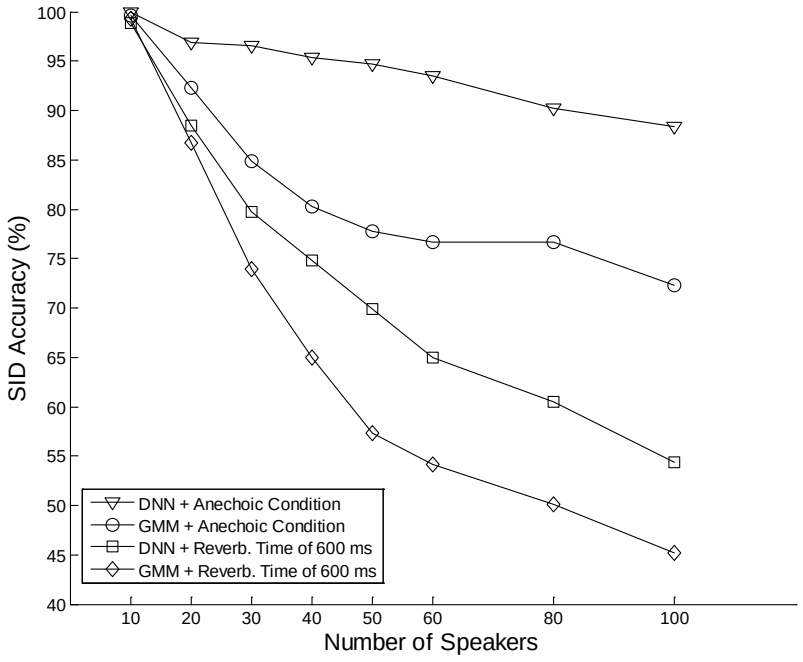


**Figure 3**. Cochannel SID accuracy of DNN- and GMM-based approaches with respect to the number of speakers in both anechoic and reverberant conditions.

## 1.4 Spectral Mapping for Speech dereverberation and Denoising

In real-world environments, speech is usually distorted by both reverberation and background noise, which have negative effects on speech intelligibility and speech quality. They also cause performance degradation in many speech technology applications, such as automatic speech recognition. Therefore, the dereverberation and denoising problems must be dealt with in daily listening environments. Reverberation corresponds to a convolution of the direct sound and a room impulse response (RIR), which distorts the spectrum of speech in both time and frequency domains. Thus, dereverberation may be treated as inverse filtering. The magnitude relationship between an anechoic signal and its reverberant version is relatively consistent in different reverberant conditions, especially within the same room. Even when reverberant speech is mixed with background noise, it is still possible to restore speech to some degree from the mixture, because speech is highly structured. These properties motivate us to utilize supervised learning to model the reverberation and mixing process.

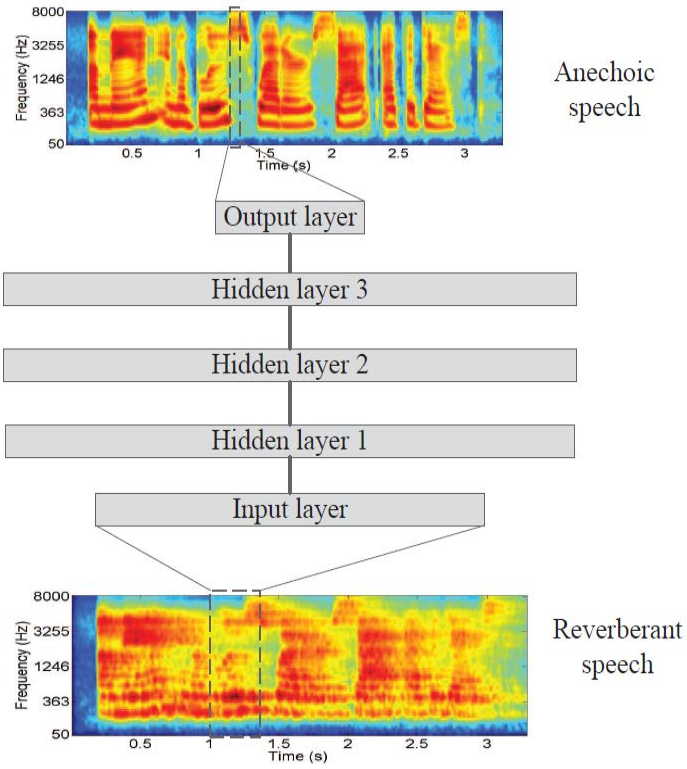


**Figure 4**. Diagram of DNN based spectral mapping for speech dereverberation. The inputs are the log spectrogram of the current frame and its neighboring frames of reverberant speech, and the outputs are the log spectrogram of the current frame of clean (anechoic) speech.

We have proposed to learn the spectral mapping from reverberant speech to its anechoic version. The mapper is trained where the input is the spectral representation of reverberant speech and the desired output is that of anechoic speech. We then extend the spectral mapping approach to perform both dereverberation and denoising. Specifically, we train a DNN to learn the spectral mapping from reverberant, or reverberant and noisy, signals to clean signals. The input for each training sample is the log magnitude spectrogram in a window of frames, and the number of input units is the same as the dimensionality of the feature vector. The output is the log magnitude spectrogram of clean speech in the current frame. Our approach for speech dereverberation is illustrated in Figure 4.

We use cross validation on a development set to choose the number of hidden layers and the number of units in each hidden layer. The objective function for optimization during training is based on mean square error. The activation function in the hidden layers is the rectified linear function and the output layer uses the sigmoid function. The optimization technique uses gradient descent along with adaptive learning rates and a momentum term. With the capacity of learning internal representations, DNN is able to encode the spectral transformation from corrupted speech to clean speech and help to restore the magnitude spectrogram of the clean signal.

Figure 5 shows an example of spectral mapping for dereverberation for a female sentence. Figures 5(a) and (b) show the log magnitude spectrogram of the clean speech and the reverberant speech with reverberation time T60 = 0.6 s. The corresponding DNN output is displayed in Fig. 5(c). As shown in Fig. 5(c), the smearing effect caused by reverberation is largely removed or attenuated, and the boundaries between voiced and unvoiced frames are considerably restored, showing that the DNN output is a very good estimate of the spectrogram of the clean speech.

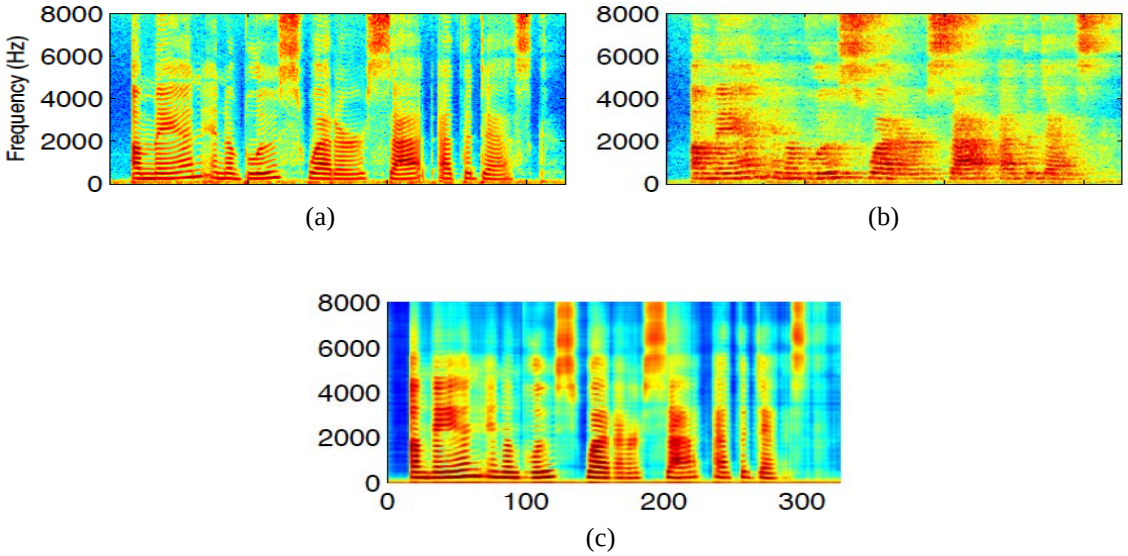


**Figure 5**. Spectral mapping for speech dereverberation. (a) Log magnitude spectrogram of clean speech. (b) Log magnitude spectrogram of reverberant speech with T60 = 0.6 s. (c) Log magnitude spectrogram of dereverberated signal.

When dealing with both dereverberation and denoising, the only change to the spectral mapping approach is that the input to the DNN is now the log magnitude spectrogram of reverberant and noisy speech, and the output is the log magnitude spectrogram of anechoic clean speech. Figure 6 illustrates the result of spectral mapping for both dereverberation and denoising. Fig. 6(a) copies Fig. 5(a) showing the same clean utterance. Fig. 6(b) shows the spectrogram of the reverberant version of Fig. 6(a) with T60 = 0.6 s that is further mixed with a factory noise at 0 dB SNR. The DNN output is shown in Fig. 6(c). It is clear from this figure that DNN-based spectral mapping does a very job at restoring the time-frequency structure of speech that is corrupted both room reverberation and background noise.

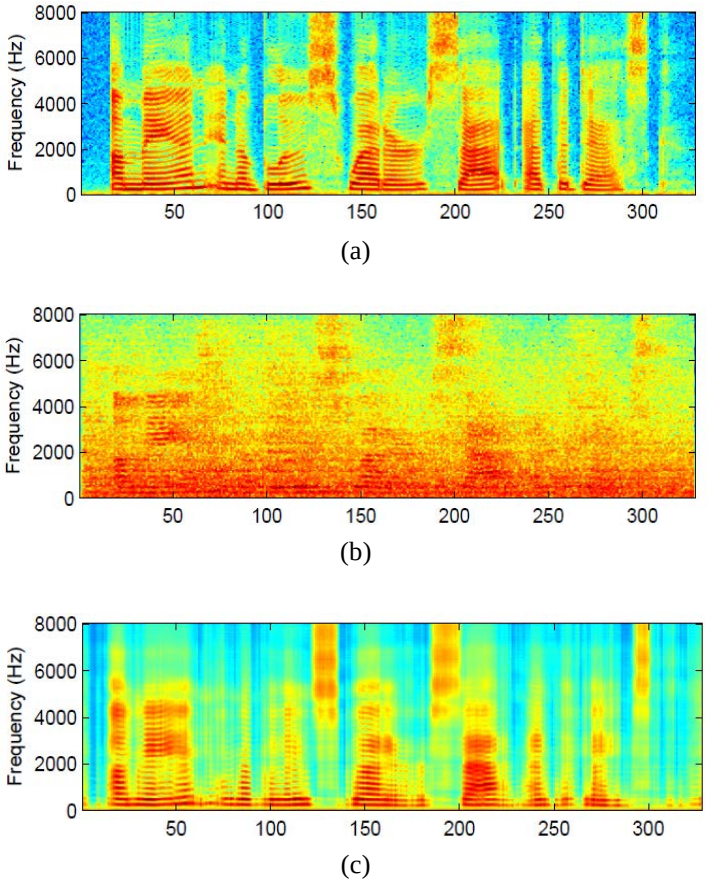


**Figure 6**. Spectral mapping for speech dereverberation and denoising. (a) Log magnitude spectrogram of clean speech. (b) Log magnitude spectrogram of reverberant and noisy speech. (c) Log magnitude spectrogram of dereverberated and denoised signal.

Extensive evaluations demonstrate that the spectral mapping approach leads to significant improvements of objective speech intelligibility and quality, as well as automatic speech recognition in reverberant noisy conditions. In addition, systematic

comparisons show that our approach substantially outperforms related methods. To our knowledge, this is the first study employing supervised learning to address the important problem of speech dereverberation. This work was published in a 2015 paper by K. Han, et al., entitled “Learning spectral mapping for speech dereverberation and denoising,” in *IEEE/ACM Transactions on Audio, Speech and Language Processing*.

## 2. OTHER INFORMATION

### 2.1 Development of Human Resources

The project in various stages has supported four doctoral students as graduate research assistants: Kun Han, Xiaojia Zhao, Yuxuan Wang, and Jitong Chen. The support enabled Han, Zhao and Wang to complete their doctoral studies. Chen is a Ph.D. candidate and is expected to finish in a year. Chen’s research addresses feature extraction, noise perturbation to expand training samples, and large-scale training to address generalization in supervised speech separation.

Kun Han’s dissertation work helped to establish the classification-based approach to speech separation. In addition, he addressed the generalization issue in the support vector machine framework. His later work employed DNN for pitch tracking in adverse conditions, and developed the spectral mapping approach to deal with dereverberation and denoising. Two pieces of his work are described in Sections 1.1 and 1.4. Han’s dissertation was completed in 2014. An executive summary of the dissertation is given in Appendix 1. His dissertation is available online at:

https://etd.ohiolink.edu/pg_10?0::NO:10:P10_ACCESSION_NUM:osu1407865723

Xiaojia Zhao’s doctoral research deals with robust speaker identification. His research has made several contributions. First, he proposed a robust SID algorithm based on computational auditory scene analysis. He then proposed an SID method that is robust to both additive noise and room reverberation. Finally, he put forward methods to deal with SID in cochannel speech in reverberant conditions. As described in Sect. 1.3, he was the first to introduce DNN to address robust SID. An executive summary of Zhao’s dissertation is given in Appendix 2. His dissertation is available online at:

https://etd.ohiolink.edu/pg_10?0::NO:10:P10_ACCESSION_NUM:osu1402620178

Yuxuan Wang’s dissertation work is the first to introduce deep neural networks to the domain of speech separation or enhancement. He has made influential contributions to a number of important topics, including feature design, training targets, generalization via extensive training, and time-domain signal construction to improve speech quality. His DNN-based separation algorithm was the first monaural method to substantially improve intelligibility of noisy speech for hearing-impaired listeners. In short, his dissertation has played a pivotal role in establishing DNN-based supervised speech separation. His training target work is described in Sect. 1.2. An executive summary of Wang’s dissertation is given in Appendix 3. His dissertation is available online at:

https://etd.ohiolink.edu/pg_10?0::NO:10:P10_ACCESSION_NUM:osu1426366690

This grant has helped the PI to update a graduate-level course entitled "Computational audition", and enhance the existing graduate-level courses “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 Awards/Honors

The PI received the 2014 Distinguished Scholar Award from the Ohio State University (OSU). This award s annually recognizes and honors up to six faculty members who have made exceptional achievements in their fields.

Yuxuan Wang, the PI's doctoral student, received the 2015 Starkey Signal Processing Research Award. This award honors the student(s) of an outstanding paper in the areas of assisted listening technologies, speech enhancement, noise suppression and low power real-time embedded design for hearing instruments, accepted for publication in the International Conference on Acoustics, Speech, and Signal Processing (ICASSP) sponsored by the IEEE Signal Processing Society. He also received the OSU Presidential Fellowship in Autumn 2013, the highest honor bestowed to about a dozen students by the OSU Graduate School in each competition. In addition, he received the Chandrasekaran and Mamrak Graduate Research Award in Spring 2015 from the OSU Department of Computer Science and Engineering.

### 2.3 Transition or Collaborative Activities

The PI was awarded in 2015 a 2-year contract from AFRL/IF in Rome to apply the results of speech segregation to automatic speech recognition (ASR) in noisy conditions. This contract aims to achieve robust ASR in the deep neural network framework through integrated acoustic modeling and separation. The performance of the proposed system will be systematically evaluated using the series of recently constructed CHIME corpora.

Kuzer, a small-business company located in Seattle, Washington, partnered with the PI in winning a Phase II STTR project funded by AFOSR. This two-year project started in late 2011, hence partly overlapping with this AFOSR project. The project successfully developed a prototype speech separation system that is computationally efficient and can operate with a processing delay close to real time. The project led to a patent application filed in 2014.

Our proposed spectral mapping method (see Sect. 1.4) was transitioned to Starkey Hearing Technologies, the largest hearing aid manufacturer in the U.S. Starkey sponsored a graduate student in the PI's laboratory to analyze the computational complexity of the algorithm and simplify its computations for potential incorporation in hearing devices. At Starkey's invitation, the PI spent 3 months in 2014 to evaluate the potential of this method in improving the speech intelligibility of hearing-impaired listeners in reverberant and noisy environments. The Starkey contact is Dr. Tao Zhang.

The PI currently serves as a technical advisor to Audience, a Knowles company. Audience is a provider of audio and noise suppression processors for mobile equipment, including Android phones. The PI advises the company on speech separation, pitch tracking, and deep learning.

## 2.4 Publications

### Journal articles


Han K. and Wang D.L. (2012): "A classification based approach to speech segregation," *Journal of the Acoustical Society of America*, vol. 132, pp. 3475-3483.

Hu K. and Wang D.L. (2013): "An unsupervised approach to cochannel speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, pp. 120-129.

Han K. and Wang D.L. (2013): "Towards generalizing classification based speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, pp. 166-175.

Wang Y., Han K., and Wang D.L. (2013): "Exploring monaural features for classification-based speech segregation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, pp. 270-279.

Woodruff J. and Wang D.L. (2013): "Binaural detection, localization, and segregation in reverberant environments based on joint pitch and azimuth cues," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, pp. 806-815.

Wang Y. and Wang D.L. (2013): "Towards scaling up classification-based speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, pp. 1381-1390.

Narayanan A. and Wang D.L. (2013): "The role of binary mask pattern in automatic speech recognition in background noise," *Journal of the Acoustical Society of America*, vol. 133, pp. 3083-3093.

Hu K. and Wang D.L. (2013): "An iterative model-based approach to cochannel speech separation," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2013, Article ID 2013-14, 11 pages.

Healy E.W., Yoho S.E., Wang Y., and Wang D.L. (2013): "An algorithm to improve speech recognition in noise for hearing-impaired listeners," *Journal of the Acoustical Society of America*, vol. 134, pp. 3029-3038.

Hartmann W., Narayanan A., Fosler-Lussier E., and Wang D.L. (2013): "A direct masking approach to robust ASR," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, pp. 1993-2005.

Narayanan A. and Wang D.L. (2014): "Investigation of speech separation as a front-end for noise robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 826-835.

Zhao X., Wang Y., and Wang D.L. (2014): "Robust speaker identification in noisy and reverberant conditions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 836-845.

Williamson D.S., Wang Y., and Wang D.L. (2014): "Reconstruction techniques for improving the perceptual quality of binary masked speech," *Journal of the Acoustical Society of America*, vol. 136, pp. 892-902.

Wang Y., Narayanan A. and Wang D.L. (2014): "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 1849-1858.

Chen J., Wang Y., and Wang D.L. (2014): "A feature study for classification-based speech separation at low signal-to-noise ratios," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 1993-2002.

Han K. and Wang D.L. (2014): "Neural network based pitch tracking in very noisy speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 2158-2168.

Healy E.W., Yoho S.E., Wang Y., Apoux F., and Wang D.L. (2014): "Speech cue transmission by an algorithm to increase consonant recognition in noise for hearing-impaired listeners," *Journal of the Acoustical Society of America*, vol. 136, pp. 3325-3336.

Narayanan A. and Wang D.L. (2015): "Improving robustness of deep neural network acoustic models via speech separation and joint adaptive training," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, pp. 92-101.

Han K., Wang Y., Wang D.L., Woods W.S., Merks I., and Zhang T. (2015): "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, pp. 982-992.

Zhao X., Wang Y., and Wang D.L. (2015): "Cochannel speaker identification in anechoic and reverberant conditions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, pp. 1727-1736.

Williamson D.S., Wang Y., and Wang D.L. (2015): "Estimating nonnegative matrix model activations with deep neural networks to increase perceptual speech quality," *Journal of the Acoustical Society of America*, vol. 138, pp. 1399-1407.

Healy E.W., Yoho S.E., Chen J., Wang Y., and Wang D.L. (2015): "An algorithm to increase speech intelligibility for hearing-impaired listeners in novel segments of the same noise type," *Journal of the Acoustical Society of America*, vol. 138, pp. 1660-1669.

Zhang X.-L. and Wang D.L. (2016): "Boosting contextual information for deep neural network based voice activity detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, pp. 252-264.

Chen J., Wang Y., and Wang D.L. (2016): "Noise perturbation for supervised speech separation," *Speech Communication*, vol. 78, pp. 1-10.

Williamson D.S., Wang Y., and Wang D.L. (2016): "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, pp. 483-492.

Wang Z.-Q. and Wang D.L. (2016): "A joint training framework for robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, pp. 796-806.

Zhang X.-L. and Wang D.L. (2016): "A deep ensemble learning method for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, pp. 967-977.

Chen J., Wang Y., Yoho S.E., Wang D.L., and Healy E.W. (2016): "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises," *Journal of the Acoustical Society of America*, vol. 139, pp. 2604-2612.

**Book chapters**

Narayanan A. and Wang D.L. (2013): "Computational auditory scene analysis and automatic speech recognition," In: Virtanen T., Raj B. and Singh R. (ed.), *Techniques for Noise Robustness in Automatic Speech Recognition*, Wiley, Chichester U.K., pp. 433-462.

**Conference papers**

Narayanan A. and Wang D.L. (2012): "On the role of binary mask pattern in automatic speech recognition," *Proceedings of INTERSPEECH-12*, pp. 1239-1242.

Wang Y., Han K., and Wang D.L. (2012): "Acoustic features for classification based speech separation," *Proceedings of INTERSPEECH-12*, pp. 1532-1535.

Wang Y. and Wang D.L. (2012): "Boosting classification based speech separation using temporal dynamics," *Proceedings of INTERSPEECH-12*, pp. 1528-1531.

Wang Y. and Wang D.L. (2012): "Cocktail party processing via structured prediction," *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS-12)*, pp. 224-232.

Han K. and Wang D.L. (2013): "Learning invariant features for speech separation," *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP-13)*, pp. 7492-7496.

Wang Y. and Wang D.L. (2013): "Feature denoising for speech separation in unknown noisy environments," *Proceedings of ICASSP-13*, pp. 7472-7476.

Zhao X. and Wang D.L. (2013): "Analyzing noise robustness of MFCC and GFCC features in speaker identification," *Proceedings of ICASSP-13*, pp. 7204-7208.

Williamson D.S., Wang Y. and Wang D.L. (2013): "A sparse representation approach for perceptual quality improvement of separated speech," *Proceedings of ICASSP-13*, pp. 7015-7019.

Narayanan A. and Wang D.L. (2013): "Ideal ratio mask estimation using deep neural networks for robust speech recognition," *Proceedings of ICASSP-13*, pp. 7092-7096.

Narayanan A. and Wang D.L. (2013): "Coupling binary masking and robust ASR," *Proceedings of ICASSP-13*, pp. 6817-6821.

Wang Y. and Wang D.L. (2014): "A structure-preserving training target for supervised speech separation," *Proceedings of ICASSP-14*, pp. 6148-6152.

Chen J., Wang Y. and Wang D.L. (2014): "A feature study for classification-based speech separation at very low signal-to-noise ratio," *Proceedings of ICASSP-14*, pp. 7089-7093.

Narayanan A. and Wang D.L. (2014): "Joint noise adaptive training for robust automatic speech recognition," *Proceedings of ICASSP-14*, pp. 2523-2527.

Zhao X., Wang Y., and Wang D.L. (2014): "Robust speaker identification in noisy and reverberant conditions," *Proceedings of ICASSP-14*, pp. 4025-4029.

Williamson D.S., Wang Y., and Wang D.L. (2014): "A two-stage approach for improving the perceptual quality of separated speech," *Proceedings of ICASSP-14*, pp. 7084-7088.

Han K., Wang Y., and Wang D.L. (2014): "Learning spectral mapping for speech dereverberation," *Proceedings of ICASSP-14*, pp. 4661-4665.

Han K., Wang Y., and Wang D.L. (2014): "Neural networks for supervised pitch tracking in noise," *Proceedings of ICASSP-14*, pp. 1502-1506.

Zhang X.-L. and Wang D.L. (2014): "Boosted deep neural networks and multi-resolution cochleagram features for voice activity detection," *Proceedings of INTERSPEECH-14*, pp. 1534-1538.

Jiang Y., Wang D.L., and Liu R.S. (2014): "Binaural deep neural network classification for reverberant speech segregation," *Proceedings of INTERSPEECH-14*, pp. 2400-2404.

Wang Y. and Wang D.L. (2015): "A deep neural network for time-domain signal reconstruction," *Proceedings of ICASSP-15*, pp. 4390-4394.

Zhao X., Wang Y., and Wang D.L. (2015): "Deep neural networks for cochannel speaker identification," *Proceedings of ICASSP-15*, pp. 4824-4828.

Williamson D.S., Wang Y., and Wang D.L. (2015): "Deep neural networks for estimating speech model activations," *Proceedings of ICASSP-15*, pp. 5113-5117.

Chen J., Wang, Y., and Wang D.L. (2015): "Noise perturbation improves supervised speech separation," *Proceedings of LVA/ICA-15*, pp. 83-90.

Han K., He Y., Bagchi D., Fosler-Lussier E., and Wang D.L. (2015): "Deep neural network based spectral feature mapping for robust speech recognition," *Proceedings of INTERSPEECH-15*, pp. 2484-2488.

Zhang X.-L. and Wang D.L. (2015): "Multi-resolution stacking for speech separation based on boosted DNN," *Proceedings of INTERSPEECH-15*, pp. 1745-1749.

Liu Y. and Wang D.L. (2015): "Speaker-dependent multipitch tracking using deep neural networks," *Proceedings of INTERSPEECH-15*, pp. 3279-3283.

Wang Z.-Q. and Wang D.L. (2015): "Joint training of speech separation, filterbank and acoustic model for robust automatic speech recognition," *Proceedings of INTERSPEECH-15*, pp. 2839-2843.

Zhao Y., Wang D.L., Merks I., and Zhang T. (2016): "DNN-based enhancement of noisy and reverberant speech," *Proceedings of ICASSP-16*, pp. 6525-6529.

Wang Z.-Q. and Wang D.L. (2016): "Robust speech recognition from ratio masks," *Proceedings of ICASSP-16*, pp. 5720-5724.

Wang Z.-Q., Zhao Y., and Wang D.L. (2016): "Phoneme-specific speech separation," *Proceedings of ICASSP-16*, pp. 146-150.

Liu Y. and Wang D.L. (2016): "Robust pitch tracking in noisy speech using speaker-dependent deep neural networks," *Proceedings of ICASSP-16*, pp. 5255-5259.

Williamson D.S., Wang Y., and Wang D.L. (2016): "Complex ratio masking for joint enhancement of magnitude and phase," *Proceedings of ICASSP-16*, pp. 5220-5224.

## Appendix 1. Executive Summary of Kun Han's Ph.D. Dissertation

In real-world environments, speech often occurs simultaneously with acoustic interference, such as background noise or reverberation. The interference usually leads to adverse effects on speech perception, and results in performance degradation in many speech applications, including automatic speech recognition and speaker identification. Monaural speech separation and processing aim to separate or analyze speech from interference based on only one recording. Although significant progress has been made on this problem, it is a widely regarded challenge.

Unlike traditional signal processing, this dissertation addresses the speech separation and processing problems using machine learning techniques. This doctoral research first proposes a classification approach to estimate the ideal binary mask (IBM) which is considered as a main goal of sound separation in computational auditory scene analysis (CASA). The dissertation employs support vector machines (SVMs) to classify time-frequency (T-F) units as either target-dominant or interference-dominant. A rethresholding method is incorporated to improve classification results and maximize hit minus false alarm rates. Systematic evaluations show that the proposed approach produces accurate estimated IBMs.

In a supervised learning framework, the issue of generalization to conditions different from those in training is very important. The generalization issue is addressed through methods that require only a small training corpus and can generalize to unseen conditions. The system utilizes SVMs to learn classification cues and then employs a rethresholding technique to estimate the IBM. A distribution fitting method is introduced to generalize to unseen signal-to-noise ratio conditions and voice activity detection based adaptation is used to generalize to unseen noise conditions. In addition, the dissertation proposes to use a novel metric learning method to learn invariant speech features in the kernel space. The learned features encode speech-related information and can generalize to unseen noise conditions. Experiments show that the proposed approaches produce high quality IBM estimates under unseen conditions.

Besides background noise, room reverberation is another major source of signal degradation in real environments. Reverberation when combined with background noise is particularly disruptive for speech perception and many applications. The work described in the dissertation performs dereverberation and denoising using supervised learning. A deep neural network (DNN) is trained to directly learn a spectral mapping from the spectrogram of corrupted speech to that of clean speech. The spectral mapping approach substantially attenuates the distortion caused by reverberation and background noise, leading to improvement of predicted speech intelligibility and quality scores, as well as speech recognition rates.

Pitch is one of the most important characteristics of speech signals. Although pitch tracking has been studied for decades, it is still challenging to estimate pitch from speech in the presence of strong noise. A pitch estimation method is proposed, and it uses supervised learning where probabilistic pitch states are directly learned from noisy speech data. Two alternative neural networks are investigated in order to model pitch state distribution given observations, i.e., a feedforward DNN and a recurrent deep neural network (RNN). Both DNNs and RNNs produce accurate probabilistic outputs of pitch states, which are then connected into pitch contours by Viterbi decoding. Experiments show that the proposed algorithms are robust to different noise conditions.

## Appendix 2. Executive Summary of Xiaojia Zhao's Ph.D. Dissertation

As a primary topic in speaker recognition, speaker identification (SID) aims to identify the underlying speaker(s) given a speech utterance. SID systems perform well under matched training and test conditions. In real-world environments, mismatch caused by background noise, room reverberation or competing voice significantly degrades the performance of such systems. Achieving robustness to the SID systems becomes an important research problem. Existing approaches address this problem from different perspectives such as proposing robust speaker features, introducing noise to clean speaker models, and using speech enhancement methods to restore clean speech characteristics. Inspired by auditory perception, computational auditory scene analysis (CASA) typically segregates speech from interference by producing a time-frequency mask. This dissertation aims to address the SID robustness problem in the CASA framework.

The doctoral research first deals with the noise robustness of SID systems. The dissertation employs an auditory feature, gammatone frequency cepstral coefficient (GFCC), and shows that this feature captures speaker characteristics and performs substantially better than conventional speaker features under noisy conditions. To deal with noisy speech, CASA separation is applied, followed by either reconstruction or marginalization of corrupted components indicated by a CASA mask. Both reconstruction and marginalization are found to be effective. These two methods are further combined into a single system based on their complementary advantages, and this system achieves significant performance improvements over related systems under a wide range of signal-to-noise ratios (SNR). In addition, systematic investigation is conducted on why GFCC shows superior noise robustness with the conclusion that nonlinear log rectification is likely the reason.

Speech is often corrupted by both noise and reverberation. There have been studies to address each of them, but the combined effects of noise and reverberation have been rarely studied. This issue is addressed in two phases. First, background noise is removed through binary masking using a deep neural network (DNN) classifier. Then, robust SID is performed with speaker models trained in selected reverberant conditions, on the basis of bounded marginalization and direct masking. Evaluation results show that the proposed method substantially improves SID performance compared to related systems in a wide range of reverberation time and SNRs.

The aforementioned studies handle mixtures of target speech and non-speech intrusions by taking advantage of their different characteristics. Such methods may not apply if the intrusion is a competing voice, which is of similar characteristics as the target. SID in cochannel speech, where two speakers are talking simultaneously over a single recording channel, is a well-known challenge. Previous studies address this problem in the anechoic environment under the Gaussian mixture model (GMM) framework. On the other hand, cochannel SID in reverberant conditions has not been addressed. This dissertation studies cochannel SID in both anechoic and reverberant conditions. This dissertation first investigates GMM-based approaches and proposes a combined system that integrates two cochannel SID methods. Secondly, DNNs are explored for cochannel SID, resulting in a DNN-based recognition system. Evaluation results demonstrate that the proposed systems significantly improve SID performance over recent approaches in both anechoic and reverberant conditions and various target-to-interferer ratios.

## Appendix 3. Executive Summary of Yuxuan Wang's Ph.D. Dissertation

Speech is crucial for human communication. However, speech communication for both humans and automatic devices can be negatively impacted by background noise, which is common in real environments. Due to numerous applications, such as hearing prostheses and automatic speech recognition, separation of target speech from sound mixtures is of great importance. Among many techniques, speech separation using a single microphone is most desirable from an application standpoint. The resulting monaural speech separation problem has been a central problem in speech processing for several decades. However, its success has been limited thus far.

Time-frequency (T-F) masking is a proven way to suppress background noise. With T-F masking as the computational goal, speech separation reduces to a mask estimation problem, which can be cast as a supervised learning problem. This opens speech separation to a plethora of machine learning techniques. Deep neural networks (DNN) are particularly suitable to this problem due to their strong representational capacity. This dissertation presents a systematic effort to develop monaural speech separation systems using DNNs.

The dissertation starts by presenting a comparative study on acoustic features for supervised separation. In this relatively early work, support vector machine is used as the classifier to predict the ideal binary mask (IBM), which is a primary goal in computational auditory scene analysis. It is found that traditional speech and speaker recognition features can actually outperform previously used separation features. Furthermore, a feature selection method is presented to systematically select complementary features. The resulting feature set is used throughout the dissertation.

DNN has shown success across a range of tasks. The dissertation then studies IBM estimation using DNN, and shows that it is significantly better than previous systems. Once properly trained, the system generalizes reasonably well to unseen conditions. It is demonstrated that the proposed system can improve speech intelligibility for hearing-impaired listeners. Furthermore, by considering the structure in the IBM, the work described in this dissertation shows how to improve IBM estimation by employing sequence training and optimizing a speech intelligibility predictor.

The IBM is used as the training target in previous work due to its simplicity. DNN based separation is not limited to binary masking, and choosing a suitable training target is obviously important. The performance of a number of targets is investigated and it is found that ratio masking can be preferable, and T-F masking in general outperforms spectral mapping. In addition, a new target is proposed that encodes structure into ratio masks.

Generalization to noises not seen during training is key to supervised separation. A simple and effective way to improve generalization is to train on multiple noisy conditions. Along this line, it is demonstrated that the noise mismatch problem can be well remedied by large-scale training. This important result substantiates the practicability of DNN based supervised separation.

Aside from speech intelligibility, perceptual quality is also important. In the last part of the dissertation, a new DNN architecture is proposed that directly reconstructs time-domain clean speech signal. The resulting system significantly improves objective speech quality over standard mask estimators.

1.

**1. Report Type**

Final Report

**Primary Contact E-mail**

**Contact email if there is a problem with the report.**

dwang@cse.ohio-state.edu

**Primary Contact Phone Number**

**Contact phone number if there is a problem with the report**

614-292-6827

**Organization / Institution name**

Ohio State University

**Grant/Contract Title**

**The full title of the funded effort.**

Speech segregation based on binary classification

**Grant/Contract Number**

**AFOSR assigned control number. It must begin with "FA9550" or "F49620" or "FA2386".**

FA9550-12-1-0130

**Principal Investigator Name**

**The full name of the principal investigator on the grant or contract.**

DeLiang Wang

**Program Manager**

**The AFOSR Program Manager currently assigned to the award**

Patrick Bradshaw

**Reporting Period Start Date**

05/01/2012

**Reporting Period End Date**

04/30/2016

**Abstract**

Speech segregation is a fundamental challenge in speech and audio processing. This AFOSR project aimed to develop a speech segregation system that can potentially improve speech intelligibility in noise for human listeners. Motivated by the perceptual principles of auditory scene analysis and the speech intelligibility studies of ideal time-frequency masking, the project sought to develop a classification-based approach to address the speech segregation challenge. The supervised approach is in sharp contrast to traditional speech segregation approaches. There are four major accomplishments made in this project. First, a supervised approach based on neural networks was developed to perform pitch tracking in very noisy conditions. Second, different training targets were examined for supervised speech segregation, leading to the adoption of the ideal ratio mask (IRM). A subsequent listening evaluation shows increased intelligibility in noise for human listeners following IRM estimation. Third, an algorithm was proposed to recognize speakers in cochannel (two-talker) conditions. This algorithm uses deep neural networks for cochannel speaker identification, and achieves the state-of-the-art results in both anechoic and reverberant conditions. Fourth, a spectral mapping method was developed to address the issue of robustness to room reverberation. This supervised method learns a mapping from the magnitude spectrogram of reverberant speech to that of anechoic speech, as well as from the spectrogram of reverberant-noisy speech to that of anechoic-clean speech. Besides these accomplishments, this project has contributed to the development

of an unsupervised approach to cochannel speech separation, analysis of generalization of the binary classification approach, and a study of acoustic-phonetic features.

**Distribution Statement**

**This is block 12 on the SF298 form.**

Distribution A - Approved for Public Release

**Explanation for Distribution Statement**

**If this is not approved for public release, please provide a short explanation. E.g., contains proprietary information.**

**SF298 Form**

**Please attach your SF298 form. A blank SF298 can be found here. Please do not password protect or secure the PDF The maximum file size for an SF298 is 50MB.**

sf298_16.pdf

**Upload the Report Document. File must be a PDF. Please do not password protect or secure the PDF . The maximum file size for the Report Document is 50MB.**

Final16.pdf

**Upload a Report Document, if any. The maximum file size for the Report Document is 50MB.**

**Archival Publications (published) during reporting period:**

See uploaded report.

**2. New discoveries, inventions, or patent disclosures:**
**Do you have any discoveries, inventions, or patent disclosures to report for this period?**

Yes

**Please describe and include any notable dates**

A US patent application was filed on September 29, 2014. The title of the invention is "Monaural speech filter", and application number is 14/388,260.

**Do you plan to pursue a claim for personal or organizational intellectual property?**

No

**Changes in research objectives (if any):**

**Change in AFOSR Program Manager, if any:**

From Dr. Willard Larkin to Dr. Patrick Bradshaw

**Extensions granted or milestones slipped, if any:**

**AFOSR LRIR Number**

**LRIR Title**

**Reporting Period**

**Laboratory Task Manager**

**Program Officer**

**Research Objectives**

**Technical Summary**

**Funding Summary by Cost Category (by FY, $K)**

| | Starting FY | FY+1 | FY+2 |
|---|---|---|---|
| Salary | | | |
| Equipment/Facilities | | | |
| Supplies | | | |
| Total | | | |

**Report Document**

**Report Document - Text Analysis**

**Report Document - Text Analysis**

**Appendix Documents**

## 2. Thank You

**E-mail user**

Jul 15, 2016 12:27:42 Success: Email Sent to: dwang@cse.ohio-state.edu