Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises

Jitong Chen[a] and Yuxuan Wang
Department of Computer Science and Engineering, The Ohio State University, Columbus, Ohio 43210, USA

Sarah E. Yoho[b]
Department of Speech and Hearing Science, The Ohio State University, Columbus, Ohio 43210, USA

DeLiang Wang[b]
Department of Computer Science and Engineering, The Ohio State University, Columbus, Ohio 43210, USA

Eric W. Healy[b]
Department of Speech and Hearing Science, The Ohio State University, Columbus, Ohio 43210, USA

(Received 23 December 2015; revised 7 April 2016; accepted 13 April 2016; published online 11 May 2016)

Supervised speech segregation has been recently shown to improve human speech intelligibility in noise, when trained and tested on similar noises. However, a major challenge involves the ability to generalize to entirely novel noises. Such generalization would enable hearing aid and cochlear implant users to improve speech intelligibility in unknown noisy environments. This challenge is addressed in the current study through large-scale training. Specifically, a deep neural network (DNN) was trained on 10 000 noises to estimate the ideal ratio mask, and then employed to separate sentences from completely new noises (cafeteria and babble) at several signal-to-noise ratios (SNRs). Although the DNN was trained at the fixed SNR of − 2 dB, testing using hearing-impaired listeners demonstrated that speech intelligibility increased substantially following speech segregation using the novel noises and unmatched SNR conditions of 0 dB and 5 dB. Sentence intelligibility benefit was also observed for normal-hearing listeners in most noisy conditions. The results indicate that DNN-based supervised speech segregation with large-scale training is a very promising approach for generalization to new acoustic environments.
 [http://dx.doi.org/10.1121/1.4948445]

[JFL] Pages: 2604–2612

I. INTRODUCTION

A primary manifest of hearing loss, which affects roughly 10% of the population, is reduced speech intelligibility in background noises, particularly, nonstationary noises (Moore, 2007; Dillon, 2012). Compressive amplification implemented in modern hearing aids offers little help as both speech and noise are amplified. The lack of speech intelligibility improvement in noise is a main barrier to hearing aid adoption (Abrams and Kihm, 2015). As a result, noise reduction is considered one of the biggest challenges in hearing aid design.

Extensive effort has been made in speech and signal processing over the past several decades to improve speech intelligibility in background noise for hearing-impaired (HI) listeners. A main approach involves speech enhancement, which is a class of monaural (single-microphone) speech segregation algorithms, including spectral subtraction and mean-square error estimation (Loizou, 2013). Speech enhancement algorithms are capable of improving signal-to-noise ratio (SNR) and speech quality, but they fail to deliver speech intelligibility benefit (Luts *et al.*, 2010; Loizou, 2013).

Recently, supervised speech segregation has received increasing attention. In its simplest form, supervised segregation estimates an ideal time-frequency (T-F) mask of a noisy mixture using a trained classifier, typically, a deep neural network (DNN). An ideal T-F mask indicates whether, or to what extent, each T-F unit is dominated by target speech. A binary decision leads to the ideal binary mask (IBM; Hu and Wang, 2001; Wang, 2005), whereas a ratio decision leads to the ideal ratio mask (IRM; Srinivasan *et al.*, 2006; Narayanan and Wang, 2013; Hummersone *et al.*, 2014; Wang *et al.*, 2014). Unlike traditional speech enhancement, supervised segregation does not make explicit statistical assumptions about the underlying speech or noise signal, but rather learns data distributions from a training set. DNN-based IBM and IRM estimators have been demonstrated to improve intelligibility of noisy speech by HI listeners (Healy *et al.*, 2013; Healy *et al.*, 2015). A critical issue associated with this work involves the ability to generalize to unseen noisy conditions—those not employed during training. In the context of supervised speech segregation, generalization to unseen noisy environments is key. In Kim *et al.* (2009), a Gaussian mixture model-based IBM classifier was trained and tested on the same brief noise segments, with very limited generalizability (May and Dau, 2014). Healy *et al.* (2013) used random cuts from longer-duration noise segments for training and testing in order to reduce

[a]Electronic mail: chenjit@cse.ohio-state.edu
[b]Also at: Center for Cognitive and Brain Sciences, The Ohio State University, Columbus, OH 43210, USA.

dependency on the specific characteristics of the training conditions. However, both training and test segments were drawn from the same overall noise segments, and generalizability was still limited.

A more recent study (Healy *et al.*, 2015) took this issue a step further by dividing 10-min nonstationary noises into two non-overlapping time portions with the first part used for training and the second part for testing. Using different portions of a noise for training and testing is considered an important requirement for evaluating supervised segregation algorithms (May and Dau, 2014). With relatively long noise segments for training and a noise perturbation technique (Chen *et al.*, 2016) to further expand the set of training noise samples, this DNN-based IRM estimator improved speech intelligibility for HI listeners in novel noise segments. However, the mask-estimation algorithm was trained and tested using the same noise type. In addition, the SNR was the same for both training and testing, which necessitated training to be repeated at each SNR tested.

The aim of the current study was to develop and test a speech segregation algorithm that can generalize to completely new noises, as well as to untrained SNRs. As the performance of supervised learning is predicated upon the information contained in a training set, the approach employed here for broad generalization was to enlarge the training set by including various acoustic conditions (see Wang and Wang, 2013). This conceptually simple approach, often referred to as multi-condition training, is widely used in automatic speech recognition (ASR) and robust ASR. In the current study, large-scale multi-condition training was employed for DNN-based IRM estimation. The training set included 10 000 noises, which exposed the IRM estimator to a large variety of noisy environments. The trained DNN was then used to segregate speech from two noises not included in those used for training: multi-talker babble and cafeteria noise. Further, training was completed at a single SNR, whereas testing was completed at multiple SNRs. Finally, the performance of the algorithm was evaluated using HI and normal-hearing (NH) listeners.

II. METHOD

A. Stimuli

The stimuli included Institute of Electrical and Electronics Engineers (IEEE) sentences (IEEE, 1969). They were spoken by one male talker and digitized at 44.1 kHz with 16-bit resolution. Each sentence in this corpus contained five scoring keywords. The background noises, also employed by Healy *et al.* (2015), were employed here to test algorithm performance. These included 20-talker babble (both male and female voices) and cafeteria noise, both from an Auditec CD (St. Louis, MO, www.auditec.com). The cafeteria noise consisted of three overdubbed recordings made in a hospital employee cafeteria. SNRs employed to test algorithm performance were selected to obtain scores for unprocessed sentences in noise below and above 50%. These were 0 and + 5 dB for the HI subjects and − 2 and − 5 dB for the NH subjects. Stimuli were downsampled to 16 kHz prior to processing.

Of the total of 720 IEEE sentences, 160 were arbitrarily selected to test algorithm performance. The remaining 560 IEEE sentences were employed for algorithm training, as described in Sec. II B. Thus, as in previous works (Healy *et al.*, 2013; Healy *et al.*, 2015), sentences employed for algorithm testing were not employed for training. Test stimuli were created by mixing each test sentence with a segment of noise randomly selected from the final 2 min of the babble or cafeteria noise recording. This method follows that of Healy *et al.* (2015), hence, facilitating detailed comparison. An unprocessed speech-in-noise condition consisted of test sentences mixed with randomly selected segments of babble or cafeteria noise at the appropriate SNR. The algorithm-processed condition employed these same test sentences, each mixed with the same randomly selected noise segment used for the unprocessed condition. Thus, the only difference between the unprocessed and segregated conditions was algorithm processing.

B. Algorithm description

1. IRM estimation using DNN

The IRM was employed as the training target for supervised speech segregation (Srinivasan *et al.*, 2006; Narayanan and Wang, 2013; Hummersone *et al.*, 2014; Wang *et al.*, 2014). The IRM is defined as

$$\mathrm{IRM}(t,f) = \sqrt{\frac{S(t,f)}{S(t,f)+N(t,f)}},$$

where $S(t,f)$ and $N(t,f)$ denote speech and noise energy within a T-F unit at time t and frequency f, respectively. The IRM is computed from the cochleagram (Wang and Brown, 2006) of the premixed speech and noise. The cochleagram has 64 frequency channels centered from 50 to 8000 Hz and equally spaced on the equivalent rectangular bandwidth scale. Figure 1 shows a diagram of DNN-based IRM estimation for speech segregation. IRM estimation starts with extraction of acoustic features from noisy speech. The DNN is then trained using these features from each speech-plus-noise mixture, along with the IRM for that mixture. After training, the DNN is used to estimate the IRM when provided only the speech-plus-noise mixture, which is then applied to the noisy speech to resynthesize a segregated speech signal. It was chosen to estimate the IRM instead of the IBM because ratio masking leads to better speech quality without compromising intelligibility (Wang *et al.*, 2014; see also Healy *et al.*, 2015).

Specifically, the IRM was computed with a 20-ms frame length and 10-ms frame shift. The power (1/15) compressed cochleagram of noisy speech was used as the only acoustic feature for IRM estimation. To incorporate temporal context, 23 frames of acoustic features were concatenated as the input to a 5-hidden-layer DNN, which simultaneously predicted 5 frames of the IRM. Since each frame of the IRM was predicted five times, the average was taken as the final estimate. Predicting multiple frames of training targets in this way encodes a measure of ensemble learning and yields a

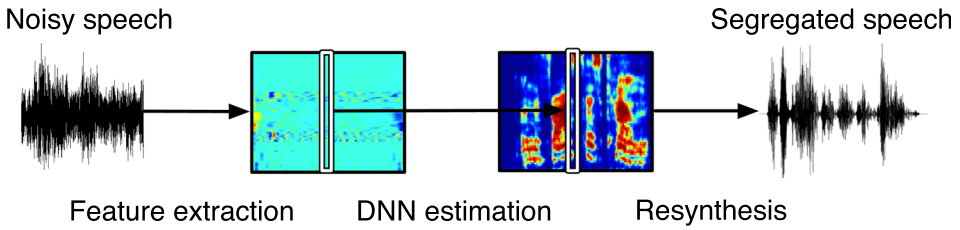

FIG. 1. (Color online) Diagram of the proposed DNN-based supervised speech segregation system.

consistent improvement in speech segregation performance (Wang *et al.*, 2014). The DNN had 23 × 64 units in the input layer, 2048 rectified linear units (Nair and Hinton, 2010) in each of the 5 hidden layers, and 5 × 64 sigmoidal units in the output layer. Dropout with a ratio of 0.2 was used for all hidden layers. Stochastic gradient descent with a mini-batch size of 256 and mean square error loss function was employed to train the DNN.

2. Large-scale training

As discussed in the Introduction, the approach employed currently for better generalization was to perform large-scale training to expose the DNN to a broad variety of noisy conditions. A large training set was created by mixing the 560 IEEE sentences with 10 000 non-speech sounds from a sound-effect library (Richmond Hill, Ontario, Canada, www.sound-ideas.com). The total duration of the noises was ∼125 h. The training set consisted of 640 000 mixtures, each of which was created by mixing a randomly selected IEEE sentence with a random segment of a randomly selected noise at the fixed SNR of −2 dB. Both random selections (sentence and noise) were done with replacement. The total duration of the training mixtures was ∼380 h. It is worth emphasizing that the 160 IEEE sentences and the 2 noises used to create test stimuli (described in Sec. II A) for speech intelligibility evaluation were not employed (seen) during training. To facilitate discussion, the model trained with 10 000 noises is called the 10 K-noise model.

In order to demonstrate the effect of the number of noises on generalization, a 100-noise model was trained using the same settings described above except that 100, rather than 10 000, non-speech environmental sounds (Columbus, OH, www.cse.ohiostate.edu/pnl/corpus/HuCorpus.html) were used, as in Wang and Wang (2013). Again, 640 000 mixtures were prepared using the 560 training sentences randomly paired with these 100 noises, so that total duration of the training set was the same as that for the 10 K-noise model.

To put the performance of the noise-independent models (i.e., 10 K-noise and 100-noise models) in perspective, the same DNN-based IRM estimator was trained and tested on the same noise type, denoted as the noise-dependent model. This model was trained on one time portion of a noise and tested on another portion of the same noise, with no overlap between noise segments used for training and those used for testing. Specifically, the two Auditec (St. Louis, MO) noises (20-talker babble and cafeteria noise) were each 10 min long, and the noise-dependent model was trained on the first 8 min of each noise and tested on the remaining 2 min of the same noise. In addition to these Auditec noises, two other noises from the NOISEX corpus (Varga and Steeneken, 1993) were used for evaluating the noise-dependent model. These noises were factory noise and 100-talker babble noise (denoted as "babble2"). The NOISEX noises are each 4 min long, and the noise-dependent model was trained on the first 2 min of each noise and tested on the remaining 2 min of the same noise. As for the other models tested currently, the 560 IEEE training sentences and an SNR of −2 dB were employed. For each of the four noises, the training set for the noise-dependent model consisted of 560 × 50 mixtures, with half of the noise samples created using frequency perturbation (Chen *et al.*, 2016; also see Healy *et al.*, 2015).

C. Subjects

A first group of subjects consisted of ten bilateral hearing-aid wearers having a sensorineural hearing loss. These HI listeners were representative of typical audiology patients seen at The Ohio State University Speech-Language-Hearing Clinic. Ages ranged from 24 to 73 yr (mean = 54.8 yr), and seven were female. Hearing status was evaluated on day of test (or within one week prior to test, for two of ten subjects) through otoscopy, tympanometry (ANSI, 1987), and pure-tone audiometry (ANSI, 2004, 2010). Pure-tone averages (PTAs, average of audiometric thresholds at 500, 1000, and 2000 Hz) ranged from 33 to 69 dB hearing level (HL; average 42.2). Hearing losses therefore ranged from mild to severe and were moderate on average. Audiograms are presented in Fig. 2, where subjects are numbered and plotted in order of increasing PTA. Also provided are subject numbers, ages, and genders.

A second group of subjects was composed of ten listeners (nine female) having NH, as defined by audiometric thresholds on day of test at or below 20 dB HL at octave frequencies from 250 to 8000 Hz (ANSI, 2004, 2010). They were recruited from undergraduate courses at The Ohio State University and had ages ranging from 19 to 41 yr (mean = 22.9 yr). All subjects received a monetary incentive or course credit for participating. As in our previous work on this topic (Healy *et al.*, 2013; Healy *et al.*, 2015), age matching between HI and NH subjects was not performed because the goal was to assess the abilities of typical (often older) HI listeners relative to the ideal performance of young NH listeners. However, it is noteworthy that the HI and NH age groups ranged considerably and overlapped. Further, the mean age of the HI listeners tested currently was only 55 yr.

D. Procedure

Each subject heard 20 sentences in each of 8 conditions (2 noise types × 2 SNRs × 2 processing conditions). Care

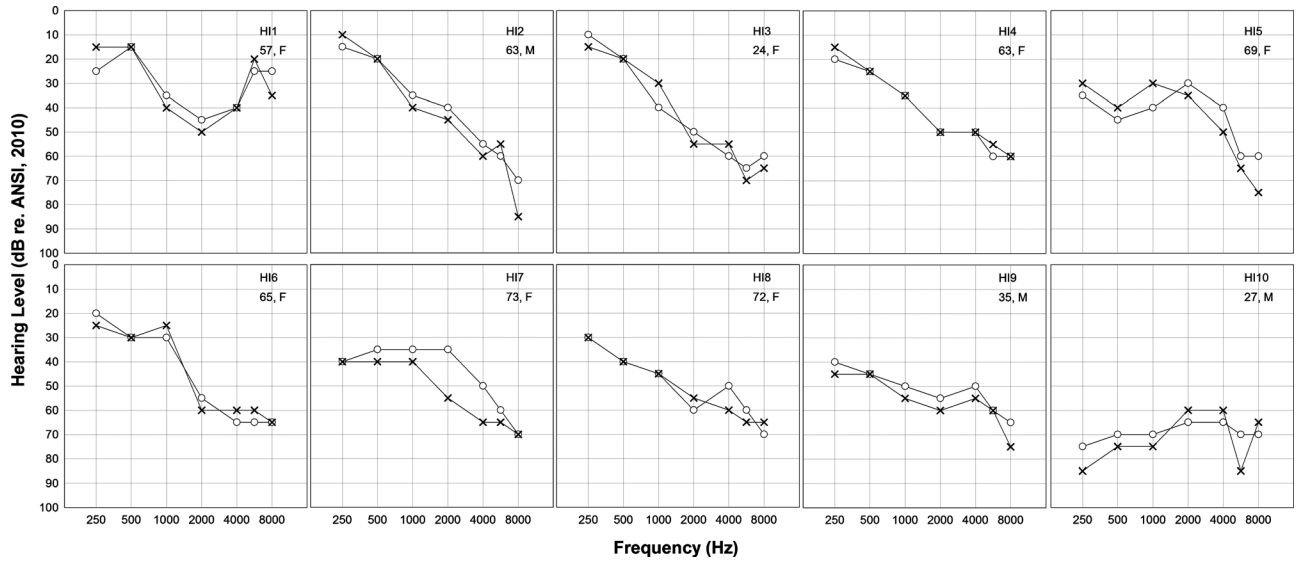

FIG. 2. Pure-tone air-conduction audiometric thresholds for the listeners with sensorineural hearing impairment. Right ears are represented by circles and left ears are represented by ×'s. Also displayed are subject number, listener age in years, and gender.

was taken to ensure that no subjects had prior exposure to the sentence materials and no sentence was repeated in any condition for any listener. Noise type and SNR were blocked so that unprocessed and algorithm conditions appeared juxtaposed in presentation order for each noise type and SNR. The order of conditions was balanced such that half the listeners heard unprocessed prior to algorithm for each noise type and SNR (and the other half heard the opposite order), and half of the subjects heard the babble conditions followed by the cafeteria-noise conditions (and the other half heard the opposite order). Sentence list-to-condition correspondence was pseudo-randomized for each subject.

The total RMS level of each stimulus in each condition was set to 65 dBA for NH listeners and 65 dBA plus frequency-specific gains as prescribed by the NAL-R hearing-aid fitting formula (Byrne and Dillon, 1986) for each individual HI listener. The fitting procedure employed in Healy *et al.* (2015) was employed, including the use of a RANE DEQ 60 L digital equalizer (Mukilteo, WA), to provide frequency-specific gains. Echo Digital Audio Gina 3 G digital-to-analog converters (Santa Barbara, CA) were employed, as was a Mackie 1202-VLZ mixer (Woodinville, WA) to adjust overall gain, and Sennheiser HD 280 Pro headphones (Wedemark, Germany) for diotic presentation. Calibration was performed using a Larson Davis sound-level meter and flat-plate headphone coupler (models 824 and AEC 101; Depew, NY). As subject-specific hearing-aid gains were provided by the experimental apparatus, HI listeners were tested with hearing aids removed.

Familiarization at the start of testing involved five IEEE sentences not employed for formal testing, first in quiet, followed by five sentences in the unprocessed noisy condition, then five in the algorithm condition. Babble or cafeteria noise was used, corresponding to whichever noise the subject was to receive first, and the SNR matched the least favorable employed during testing. This familiarization was repeated half way through the experiment using the other noise type, prior to switching noise types. The HI subjects were asked after presentation of the initial sentences if the stimuli were comfortable in level. The overall presentation level was reduced by 5 dB for the one subject who indicated that the stimuli sounded loud. This individual judged this reduced level to be comfortable. The overall presentation level was 96 dBA or below for all subjects. The experimenter was seated with the subject in a double-walled audiometric booth, and instructed the listeners to repeat back as much of each sentence as possible, controlled the presentation of each sentence, and scored responses.

III. RESULTS AND DISCUSSION

A. Predicted intelligibility results

Before presenting intelligibility results from HI and NH listeners, predicted intelligibility scores using an acoustic metric are provided. Specifically, the short-time objective intelligibility (STOI) metric (Taal *et al.*, 2011) was employed, as it is a standard speech intelligibility predictor involving a comparison between the envelopes of segregated speech and clean speech. STOI evaluation provides an opportunity to compare predicted and actual intelligibility scores and an objective benchmark for future algorithm comparisons.

Table I shows the STOI results for the unprocessed mixtures, the two noise-independent models, and the noise-dependent model. The mean STOI scores were computed for the 160 test sentences in each test-noise condition. Values are shown for each of the test noises and for the average

TABLE I. Speech segregation results for four test noises and their average at −2 dB SNR measured in STOI values.

	Babble	Cafeteria	Factory	Babble2	Average
Unprocessed	0.612	0.596	0.611	0.611	0.608
100-noise model	0.683	0.704	0.750	0.688	0.706
10 K-noise model	0.792	0.783	0.807	0.786	0.792
Noise-dependent model	0.833	0.770	0.802	0.762	0.792

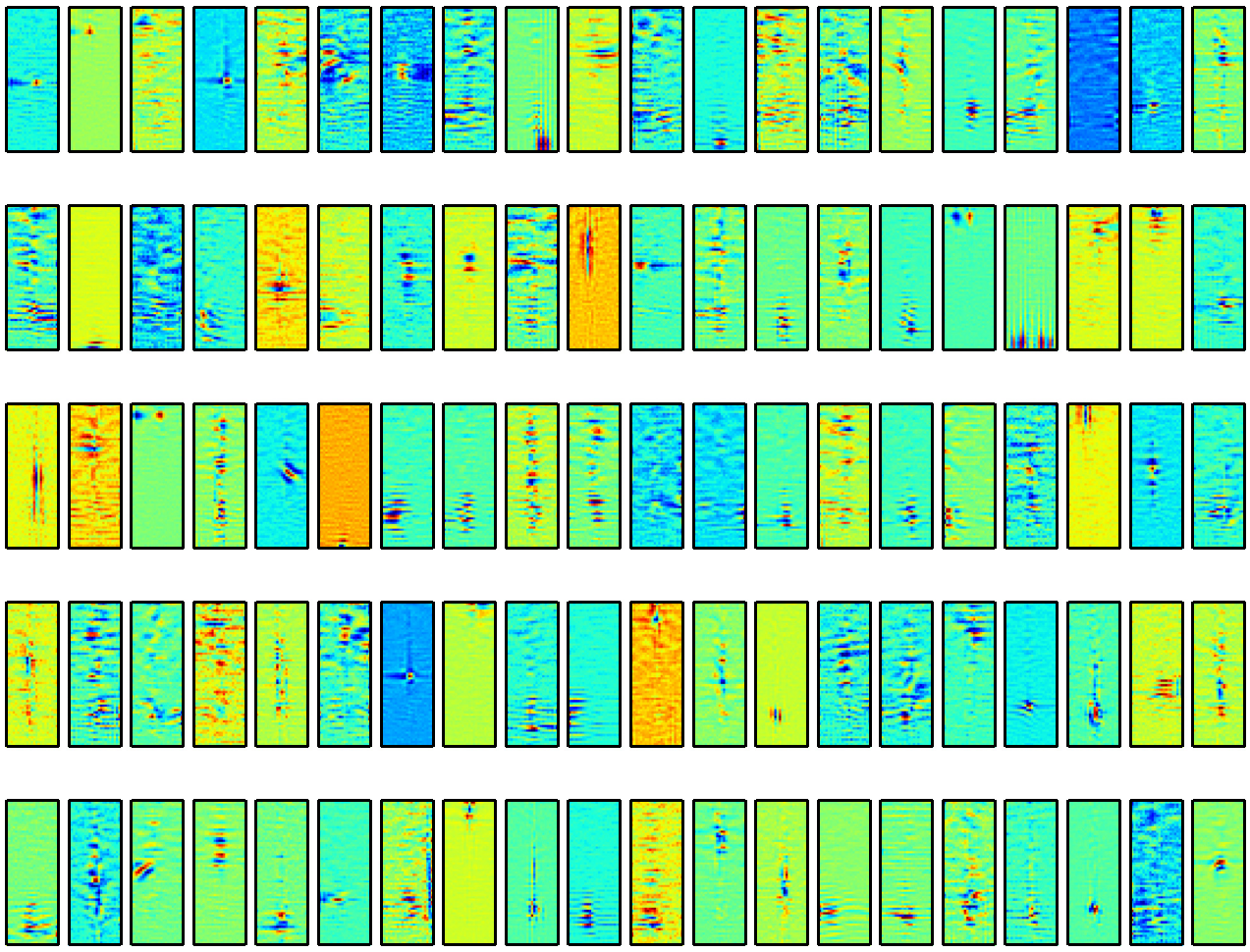

FIG. 3. (Color online) Visualization of 100 filters learned by the bottom hidden layer of a DNN trained on mixtures created using 10 000 noises. Each filter is shown in two dimensions: The abscissa represents time (23 frames) and the ordinate represents frequency (64 channels).

across noises. Apparent is that all models improved STOI scores relative to unprocessed speech in noise. The noise-independent model, trained with 100 noises, performed substantially poorer than that trained with 10 000 noises, even though the two were trained using the same number of mixtures (640 000). Therefore, it is the increase in the amount of distinct noise samples rather than the size of the training set that determines generalization ability. On the other hand, the 10 K-noise model provided identical performance on average to the noise-dependent model. This indicates that, with 10 000 noises, the noise-independent model has been exposed to an adequate variety of noisy environments. It is highly encouraging that the STOI scores for the noise-independent model match those for the noise-dependent model (see Wang *et al.*, 2015, for additional STOI results).

Figure 3 visualizes the first 100 learned filters taken from the first hidden layer of the 10 K-noise model. Each panel in Fig. 3 corresponds to a hidden unit, showing the weights coming from the input layer in two dimensions: The abscissa represents time (23 frames) and the ordinate represents frequency (64 channels). Apparent is that the network learns what appear to be speech-specific feature detectors. For example, some filters resemble harmonic detectors (e.g., the tenth filter in the last row), while some others seem to capture feature transitions (e.g., the fifth filter in the third row). These speech-specific feature detectors appear to encode fundamental characteristics of the speech signal, enabling the model to be noise independent.

Although the 10 K-noise model was trained on 640 000 mixtures created at −2 dB SNR, it is able to generalize to different SNRs. To demonstrate this, a second 10 K-noise model was trained on 640 000 new random mixtures created at −5 dB, and both models were evaluated on both the

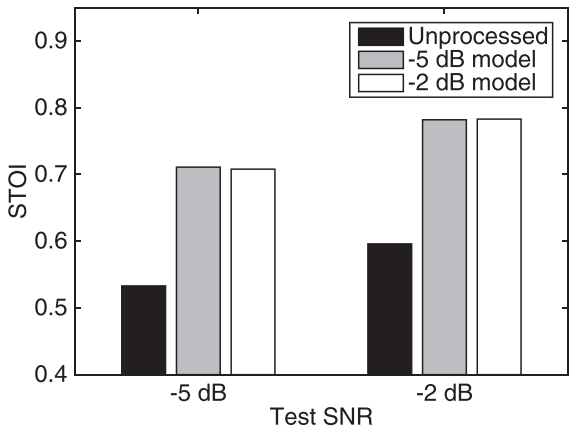

FIG. 4. STOI predictions for a noise-independent model trained and tested in matched and mismatched SNR conditions.

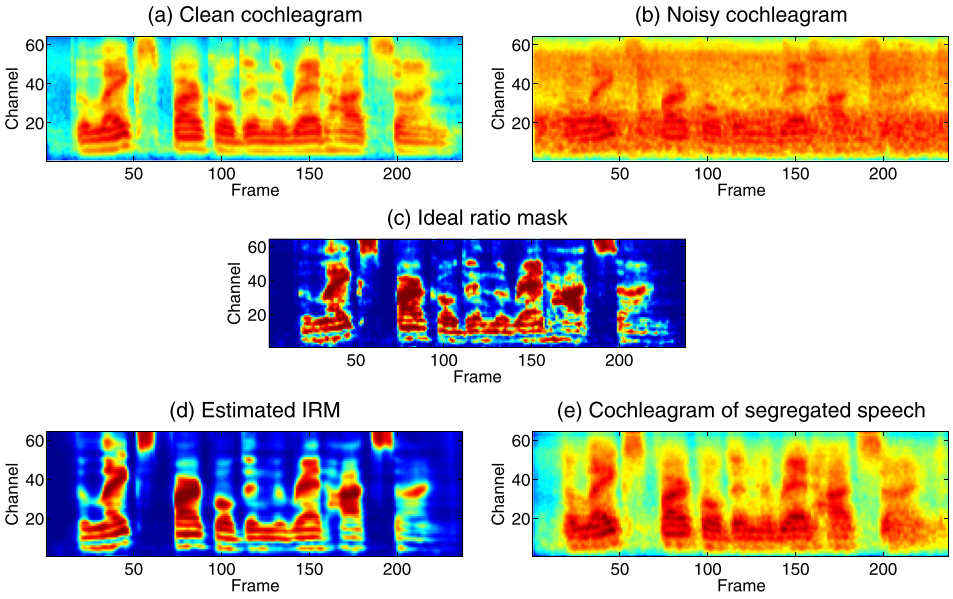

FIG. 5. (Color online) Segregation of an IEEE sentence ("The lake sparkled in the red hot sun") from cafeteria noise at 0 dB SNR; (a) cochleagram of the utterance in quiet; (b) cochleagram of the utterance in noise; (c) IRM for this mixture; (d) estimated IRM for this mixture; and (e) cochleagram of the segregated utterance by applying the estimated IRM to the noisy utterance.

−5 dB and −2 dB test sets. Cafeteria noise was employed. As shown in Fig. 4, the STOI difference between the matched and mismatched SNR conditions is negligible at both test SNR levels. This is likely because the model had seen sufficient local (i.e., frame level) SNR variations even with a fixed, overall utterance-level SNR in training. Therefore, the 10 K-noise model trained at −2 dB was used to produce the algorithm-processed stimuli for all SNR conditions employed for human-subject testing.

Figure 5 illustrates the results of using the 10 K-noise model trained at −2 dB to perform speech segregation on a mixture of an IEEE sentence and cafeteria noise at 0 dB SNR. The cochleagrams of clean speech, speech-plus-noise, and segregated speech are shown in Figs. 5(a), 5(b), and 5(e), respectively. The IRM is given in Fig. 5(c) and the estimated IRM in Fig. 5(d). It is clear that the target speech is well separated from the cafeteria noise despite that the test noise and test SNR were not used during the training stage.

Table II lists the STOI scores for the same test conditions used in the human-subjects listening tests presented in Sec. III B. Again, the mean STOI scores were computed for the 160 test sentences in each test-noise condition. As shown in Table II, the 10K-noise model substantially improves STOI values over unprocessed mixtures at all SNRs. For each SNR, similar STOI improvement was observed for the two noises, which was to be expected as the DNN was trained using a large number of noises, decreasing the likelihood of overfitting one specific noise.

TABLE II. STOI values for speech mixed with (unprocessed), and segregated from (processed), babble and cafeteria noise at the SNRs indicated.

	Babble noise		Cafeteria noise	
	Unprocessed	Processed	Unprocessed	Processed
5 dB	0.784	0.904	0.760	0.893
0 dB	0.663	0.834	0.642	0.823
−2 dB	0.612	0.792	0.596	0.783
−5 dB	0.541	0.707	0.533	0.708

B. Actual (human-subject) intelligibility results

Figure 6 shows intelligibility based on percentage of keywords reported by individual human listeners in each condition. Individual HI listeners are represented by filled symbols and NH listeners by open symbols. Scores on unprocessed speech in noise are represented by circles and those on algorithm-processed speech are represented by triangles. Algorithm benefit is therefore represented by the height of the line connecting these two symbols. As in Fig. 2, HI subjects are numbered and plotted in order of increasing PTA.

In the babble background, all but one HI subject received some benefit at the less favorable SNR. Benefit in this condition was 45 percentage points or greater for four of the ten HI listeners and was 20 points or greater for seven of the ten HI listeners. At the more favorable babble SNR, seven of ten HI subjects received some benefit. Benefit in this condition was reduced in magnitude compared to the less favorable SNR case, as most unprocessed scores were high. However, the HI listener with the lowest unprocessed score received a benefit of 42 percentage points. With regard to the NH listeners in babble noise, the majority also received some benefit (six of ten subjects at the less favorable SNR and seven of ten at the more favorable SNR). As in our previous work (Healy *et al.*, 2013; Healy *et al.*, 2015), the benefit for the NH listeners was smaller than that obtained for the HI listeners.

In the cafeteria-noise background, all but one HI listener received some benefit at the less favorable SNR. Benefit in this condition was 20 percentage points or greater for eight of the ten HI listeners. At the more favorable cafeteria-noise SNR, seven of ten HI subjects received some benefit. The HI listener with the lowest unprocessed score in this condition received a benefit of 41 percentage points. For the NH listeners in cafeteria noise, the majority also received some benefit (nine of ten subjects at the less favorable SNR and six of ten at the more favorable SNR).

Group-mean intelligibility scores in each condition are displayed in Fig. 7. In babble, the average benefit from

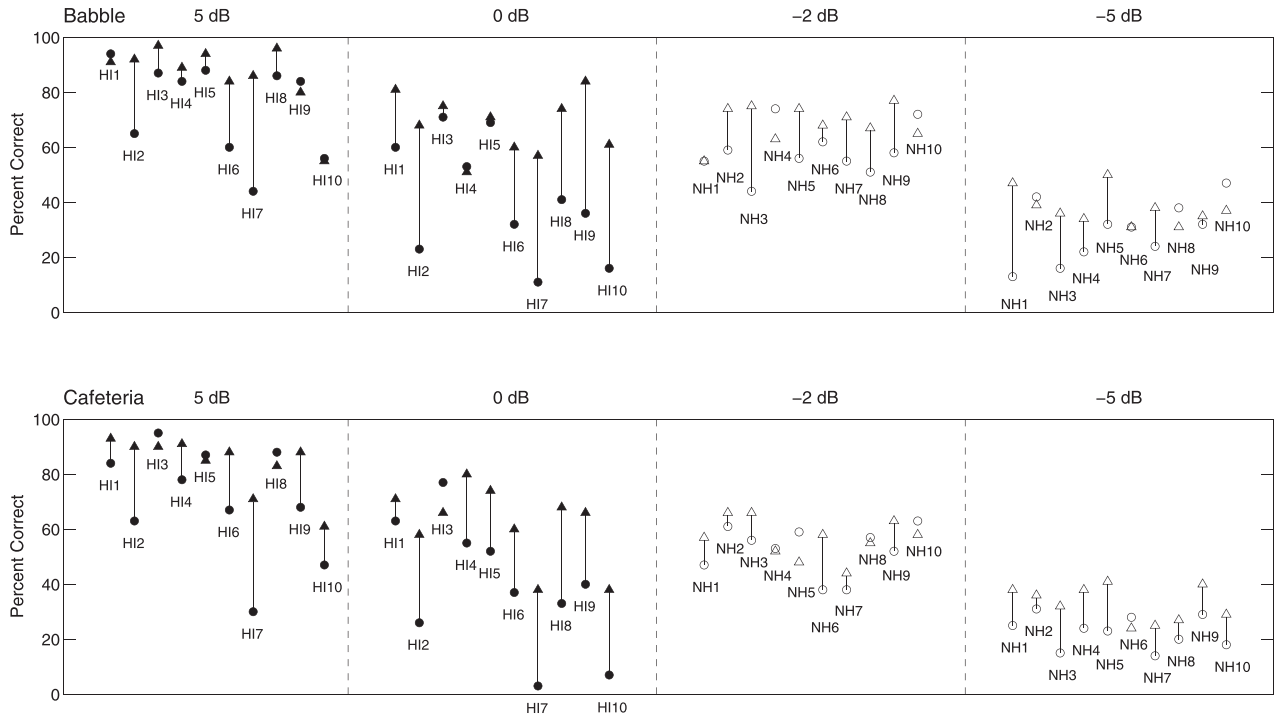

FIG. 6. Intelligibility of IEEE sentences based on percentage of keywords reported. The top panels represent scores in, or segregated from, babble noise, and the bottom panels represent scores in, or segregated from, cafeteria noise, all at the SNRs indicated. Individual HI listeners are represented by filled symbols and individual NH listeners are represented by open symbols. Scores for unprocessed speech in noise are represented by circles and scores for algorithm-processed noisy speech are represented by triangles. Algorithm benefit is represented by the height of the line connecting these symbols.

algorithm processing was 11.6 and 27.0 percentage points for the HI listeners at 5 and 0 dB SNR, and 10.3 and 8.1 percentage points for the NH listeners at −2 and −5 dB SNR, respectively. A series of planned comparisons (paired, uncorrected *t*-tests) between unprocessed and processed scores in each panel of Fig. 7 indicated that algorithm processing produced significant increases in intelligibility for both HI and NH listeners at all babble SNRs [$t(9) \geq 1.8$, $p < 0.05$].

In cafeteria noise, the average benefit from algorithm processing was 13.3 and 22.6 percentage points for the HI listeners at 5 and 0 dB SNR, and 4.3 and 10.3 percentage points for the NH listeners at −2 and −5 dB SNR, respectively. Planned comparisons indicated that algorithm

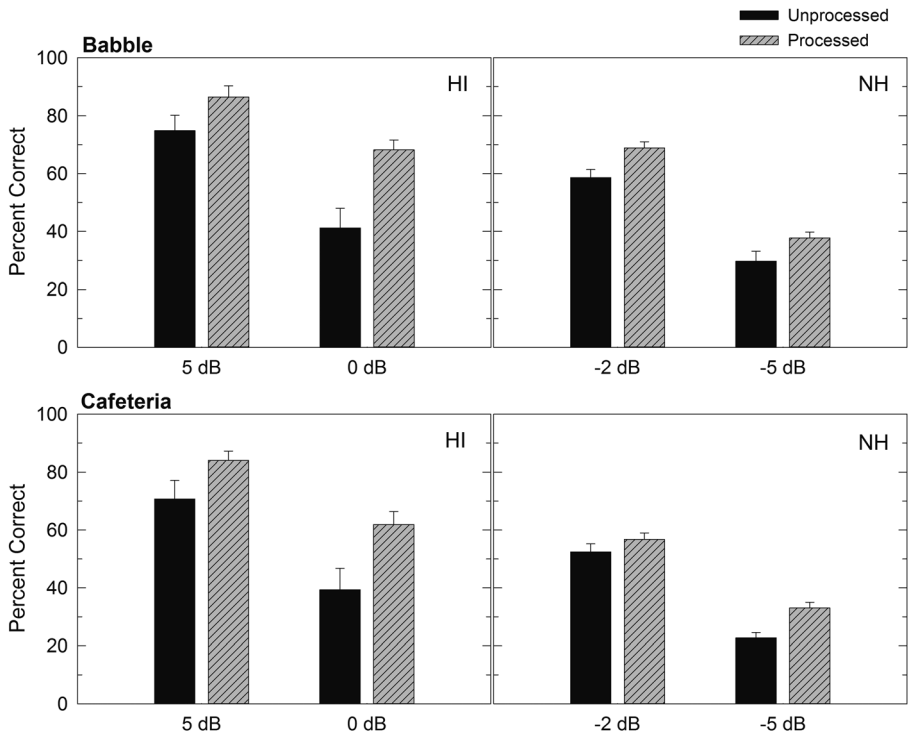

FIG. 7. Group-mean intelligibility scores and standard errors for HI and NH listeners hearing unprocessed IEEE sentences in noise and sentences following algorithm processing. The top panels show scores for a babble background and the bottom panels show scores for a cafeteria-noise background, at the SNRs indicated.

processing produced significant increases in intelligibility for the HI listeners at both cafeteria-noise SNRs [$t(9) \geq 2.8$, $p \leq 0.01$], and a significant increase in intelligibility for the NH listeners at the less favorable cafeteria-noise SNR [$t(9) = 5.1$, $p < 0.01$].

IV. GENERAL DISCUSSION

It is worth emphasizing that, in the current study, two aspects of generalization have been addressed. First, the noise-independent algorithm trained on a large set of noises that did not include the noises employed for testing and, therefore, it had to generalize to entirely novel noises. Second, the algorithm was trained at a single SNR, and it had to generalize to untrained SNRs during the test stage. These issues represent some of the most difficult challenges that must be overcome for an algorithm to have direct translational significance. Obviously, the ability to generalize to noisy environments unseen during training is a requirement for an algorithm to be useful in real-world hearing technology. Despite these challenging demands, the current model produced substantial improvements in human intelligibility.

A new group of NH listeners was tested currently in the unprocessed conditions that were identical to those in Healy *et al.* (2015), which provides an opportunity for comparison. These conditions involve the same speech materials, noise recordings, and SNRs, just different random selections of noise for each noisy sentence. Despite the use of different NH subjects, scores across the two studies were within 1 percentage point on average across the four unprocessed conditions, and no condition differed by more than 2 percentage points across studies.

The current benefits for HI subjects observed from algorithm processing are somewhat smaller than when the algorithm was tested on novel segments of the same noise type used in training (Healy *et al.*, 2015), rather than on new noises. However, much of this reduction in benefit can be attributed to the high unprocessed scores produced by the HI subjects employed in the current study. In accord with this generally better performance on unprocessed speech in noise, the PTAs of the current HI subjects are lower on average (reflecting less hearing loss) by 8.3 dB relative to their counterparts who participated in Healy *et al.* (2015). Despite the reduction in mean benefit due to high unprocessed scores, several aspects of the current results are particularly encouraging. First, those HI subjects having the lowest scores on unprocessed speech in noise received considerable benefit, in general accord with that observed when novel segments of the same noise recording were employed (Healy *et al.*, 2015). Second, the intelligibility scores following the current algorithm processing were higher than the corresponding results in Healy *et al.* (2015), particularly, for the cafeteria noise.

A third aspect that may be considered encouraging is that no decrement in performance was observed even for those subjects who displayed very high unprocessed speech-in-noise scores. This ability to avoid decrements in performance when unprocessed intelligibility is high is almost as important as the ability to improve performance when unprocessed intelligibility is low. Even when considering only the current conditions in which HI performance on unprocessed speech was 85% or above (see Fig. 6), a benefit of 1.6 percentage points was still observed. This result is consistent with our previous results on this topic (Healy *et al.*, 2013; Healy *et al.*, 2015), and suggests that the algorithm did not produce distortions that might potentially hinder better users.

A result that has been seen consistently (Healy *et al.*, 2013; Healy *et al.*, 2015) is that benefit demonstrated by HI listeners is greater than that displayed by NH listeners. This may be understood in terms of the different noise tolerances of the two groups. HI listeners are largely intolerant of noise and benefit considerably from algorithmic reduction of background noise. In contrast, NH listeners perceive speech in noise with considerable ease. Because they perform this task effectively in the unprocessed conditions, they benefit less from automatic speech segregation.

One result that differs from Healy *et al.* (2015) involves the benefit demonstrated by the NH listeners. In the 2015 report, the NH listeners demonstrated a sizeable benefit when listening to speech extracted from babble noise, but not for speech extracted from cafeteria noise. In the current study, the NH listeners received similar degrees of benefit in both noise types. Interestingly, the STOI improvements in Table II are also similar for both noise types. This similarity in benefit across the two noise types may be understood in terms of the current algorithm training procedure. Because neither of the test noises was employed during algorithm training, the generalization to them was similar and the algorithm produced similar degrees of STOI improvement. This differs from the 2015 training procedure in which generalization was tested on unseen segments of the same noise recording. In those 2015 conditions, it is apparent that generalization to a novel segment of cafeteria noise was more challenging algorithmically than generalization to a novel segment of babble, reflected by considerably larger STOI improvements for the babble noise (see Table I in Healy *et al.*, 2015). This is likely because the cafeteria noise is more dynamic, with more transient bursts than the babble noise.

Finally, human robustness to noise may have much to do with our extensive exposure to a wide range of noisy conditions. It is documented that children show elevated speech reception thresholds relative to adults when recognizing noisy speech (Hall *et al.*, 2002; Johnstone and Litovsky, 2006). Musicians exhibit higher intelligibility in speech-in-noise conditions compared to non-musicians (Parbery-Clark *et al.*, 2009), presumably, because of their more extensive experience in listening to polyphonic signals. Bilingual speakers have a deficit in speech perception in noisy conditions compared to monolingual speakers, even though these two groups show similar performance in quiet (Tabri *et al.*, 2011). All these effects are consistent with the idea that extensive training (or experience) is crucial for the remarkable noise robustness of the normal auditory system.

V. CONCLUSION

A DNN-based supervised speech segregation system with large-scale training was presented and shown to

generalize to untrained noises and SNR conditions. Speech intelligibility benefits were observed for HI listeners in both test noises and at both test SNRs. NH listeners displayed a benefit at both test SNRs for multi-talker babble noise, and at the less favorable SNR for the cafeteria noise. The current results represent a stride toward using supervised speech segregation in real-world environments.

ACKNOWLEDGMENTS

This work was supported in part by a grant from the National Institute on Deafness and other Communication Disorders (Grant No. R01 DC012048) and a grant from the Air Force Office of Scientific Research (Grant No. FA9550-12-1-0130). We gratefully acknowledge computing resources from the Ohio Supercomputer Center and contributions to the execution of the experiments by Jordan Vasko.

Abrams, H. B., and Kihm, J. (**2015**). "An introduction to MarkeTrak IX: A new baseline for the hearing aid market," Hear. Rev. **22**(6), 16.

ANSI (**1987**). S3.39 (R2012), *American National Standard Specifications for Instruments to Measure Aural Acoustic Impedance and Admittance (Aural Acoustic Immittance)* (Acoustical Society of America, New York).

ANSI (**2004**). S3.21 (R2009), *American National Standard Methods for Manual Pure-Tone Threshold Audiometry* (Acoustical Society of America, New York).

ANSI (**2010**). S3.6, *American National Standard Specification for Audiometers* (Acoustical Society of America, New York).

Byrne, D., and Dillon, H. (**1986**). "The National Acoustic Laboratories' (NAL) new procedure for selecting the gain and frequency response of a hearing aid," Ear Hear. **7**, 257–265.

Chen, J., Wang, Y., and Wang, D. L. (**2016**). "Noise perturbation for supervised speech separation," Speech Commun. **78**, 1–10.

Dillon, H. (**2012**). *Hearing Aids*, 2nd ed. (Boomerang, Turramurra, Australia), pp. 1–19.

Hall, J. W., Grose, J. H., Buss, E., and Dev, M. B. (**2002**). "Spondee recognition in a two-talker and a speech-shaped noise masker in adults and children," Ear Hear. **23**, 159–165.

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," J. Acoust. Soc. Am. **138**, 1660–1669.

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," J. Acoust. Soc. Am. **134**, 3029–3038.

Hu, G., and Wang, D. L. (**2001**). "Speech segregation based on pitch tracking and amplitude modulation," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 79–82.

Hummersone, C., Stokes, T., and Brooks, T. (**2014**). "On the ideal ratio mask as the goal of computational auditory scene analysis," in *Blind Source Separation*, edited by G. R. Naik and W. Wang (Springer, Berlin), pp. 349–368.

IEEE (**1969**). "IEEE recommended practice for speech quality measurements," IEEE Trans. Audio Electroacoust. **17**, 225–246.

Johnstone, P. M., and Litovsky, R. Y. (**2006**). "Effect of masker type and age on speech intelligibility and spatial release from masking in children and adults," J. Acoust. Soc. Am. **120**, 2177–2189.

Kim, G., Lu, Y., Hu, Y., and Loizou, P. C. (**2009**). "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," J. Acoust. Soc. Am. **126**, 1486–1494.

Loizou, P. C. (**2013**). *Speech Enhancement: Theory and Practice*, 2nd ed. (CRC Press, Boca Raton, FL), Chaps. 5–8, 12.

Luts, H., Eneman, K., Wouters, J., Schulte, M., Vormann, M., Buechler, M., Dillier, N., Houben, R., Dreschler, W. A., Froehlich, M., Puder, H., Grimm, G., Hohmann, V., Leijon, A., Lombard, A., Mauler, D., and Spriet, A. (**2010**). "Multicenter evaluation of signal enhancement algorithms for hearing aids," J. Acoust. Soc. Am. **127**, 1491–1505.

May, T., and Dau, T. (**2014**). "Requirements for the evaluation of computational speech segregation systems," J. Acoust. Soc. Am. **136**, EL398–EL404.

Moore, B. C. J. (**2007**). *Cochlear Hearing Loss*, 2nd ed. (Wiley, Chichester, UK), pp. 201–232.

Nair, V., and Hinton, G. E. (**2010**). "Rectified linear units improve restricted Boltzmann machines," in *Proceedings of ICML*, pp. 807–814.

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

Parbery-Clark, A., Skoe, E., Lam, C., and Kraus, N. (**2009**). "Musician enhancement for speech-in-noise," Ear Hear. **30**, 653–661.

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

Taal, C. H., Hendriks, R. C., Heusdens. R., and Jensen, J. (**2011**). "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," IEEE Trans. Audio Speech Lang. Process. **19**, 2125–2136.

Tabri, D., Chacra, K. M., and Pring, T. (**2011**). "Speech perception in noise by monolingual, bilingual and trilingual listeners," Int. J. Lang. Commun. Disord. **46**, 411–422.

Varga, A., and Steeneken, H. (**1993**). "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," Speech Commun. **12**, 247–251.

Wang, D. L. (**2005**). "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, edited by P. Divenyi (Kluwer, Norwell, MA), pp. 181–197.

Wang, D. L., and Brown, G. J., Eds. (**2006**). *Computational Auditory Scene Analysis: Principles, Algorithms and Applications* (Wiley-IEEE, Hoboken, NJ), pp. 1–44.

Wang, Y., Chen, J., and Wang D. L. (**2015**). "Deep neural network based supervised speech segregation generalizes to novel noises through large-scale training," Technical Report OSU-CISRC-3/15-TR02, Department of Computer Science and Engineering, Ohio State University, Columbus, OH. Available at http://www.cse.ohio-state.edu/research/techReport.shtml (Last viewed December 9, 2015).

Wang, Y., Narayanan, A., and Wang, D. L. (**2014**). "On training targets for supervised speech separation," IEEE/ACM Trans. Audio Speech Lang. Process. **22**, 1849–1858.

Wang, Y., and Wang, D. L. (**2013**). "Towards scaling up classification-based speech separation," IEEE Trans. Audio. Speech Lang. Process. **21**, 1381–1390.