**Technical Report OSU-CISRC-3/15-TR02**
Department of Computer Science and Engineering
The Ohio State University
Columbus, OH 43210-1277

Ftpsite: **ftp.cse.ohio-state.edu**
Login: **anonymous**
Directory: **pub/tech-report/2015**
File: **TR02.pdf**
Website: **http://www.cse.ohio-state.edu/research/techReport.shtml**

# Deep neural network based supervised speech segregation generalizes to novel noises through large-scale training

**Yuxuan Wang**
Department of Computer Science and Engineering
The Ohio State University, Columbus, OH 43210, USA
*wangyuxu@cse.ohio-state.edu*

**Jitong Chen**
Department of Computer Science and Engineering
The Ohio State University, Columbus, OH 43210, USA
*chenjit@cse.ohio-state.edu*

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

*Abstract* – Deep neural network (DNN) based supervised speech segregation has been successful in improving human speech intelligibility in noise, especially when DNN is trained and tested on the same noise type. A simple and effective way for improving generalization is to train with multiple noises. This letter demonstrates that by training with a large number of different noises, the objective intelligibility results of DNN based supervised speech segregation on novel noises can match or even outperform those on trained noises. This demonstration has an important implication that improving human speech intelligibility in unknown noisy environments is potentially achievable.

# 1 Introduction

Removing background noise using a single microphone is an important and long-standing challenge with many real-world applications, including improving speech intelligibility in noise for hearing-impaired listeners. Many techniques, such as traditional speech enhancement [8], have been proposed to address the monaural speech segregation problem, but with only limited success. Recently, supervised speech segregation has been established as a successful new paradigm for dealing with the monaural segregation problem. In its simplest form, supervised speech segregation estimates an ideal time-frequency (T-F) mask of a noisy mixture using a trained classifier, typically a deep neural network (DNN). Unlike speech enhancement, supervised segregation does not make explicit assumptions (e.g. stationarity) of the underlying speech or noise signal, but rather relies on the data distributions conveyed by the training set, i.e. it is a data-driven segregation framework. Recently, it has been demonstrated that a DNN based ideal binary mask (IBM) estimator is capable of substantially improving speech intelligibility for hearing-impaired (as well as normal-hearing) listeners [3]. An earlier study based on Gaussian mixture model classifiers yielded significant intelligibility improvements for normal-hearing listeners [6]. It is notable that these successes in intelligibility improvement are achieved only in the supervised segregation paradigm.

Generalization to unseen conditions is critical to any supervised learning algorithms. This applies to supervised speech segregation, as a practical segregation system must be able to deal with novel noise segments and/or noise types. Our previous work [14] and a recent study [9] have exposed the generalization issue of supervised speech segregation, where it is shown that a system trained on a single or a small number of noises may not work well on novel noises. The performance of supervised learning depends on the information contained in the training set, therefore a simple and effective way for improving generalization is to enlarge the training set by including various acoustic conditions. This technique, often referred as multi-condition training, is widely used in robust ASR systems [10]. Our previous work [14] also shows that multi-condition training with 100 environmental noises can indeed improve the generalization of DNN based IBM estimators to novel noises. On the other hand, these 100 short noise samples (totaling about 5 minutes) and the resulting 20,000 mixtures (totaling about 17 hours) do not by any means amount to a large-scale dataset, compared with typical supervised learning systems deployed in practice such as ASR.

In this study, we employ large-scale multi-condition training on a more recent DNN based mask estimator. The training set includes about 10,000 noises. We demonstrate its performance in low signal-to-noise ratio (SNR) conditions, where speech intelligibility is a main concern, and show that the performance on novel noises can match or even outperform that on trained noises in terms of objective intelligibility measures.

## 2 System description

Supervised speech segregation learns a mapping from noisy mixtures to ideal T-F masks. In this study, we employ DNN as the discriminative learning machine to estimate the ideal ratio mask (IRM), which is suggested as an alternative training target to the previously used IBM [15]. The ideal ratio mask is defined as follows:

$$IRM(t,f) = \sqrt{\frac{S^2(t,f)}{S^2(t,f) + N^2(t,f)}} \tag{1}$$

where $S^2(t,f)$ and $N^2(t,f)$ denote the speech and noise energy in a particular T-F unit, respectively. The resulting IRM is closely related, but not identical, to the Wiener filter [**?**]. Following common practice, we use a 64-channel Gammatone filterbank followed by 20-ms windowing with 10-ms overlap to estimate the IRM.

The employed DNN has 5 hidden layers, each containing 2048 rectified linear units. Standard backpropagation coupled with dropout regularization [1] (20% dropout) is used to train the network. We use mini-batch stochastic gradient descent with momentum (kept as 0.9) as the optimizer. The input features to the network are root compressed (with an exponent of 1/15) Gammatone filterbank energies. In this study, we splice a window of 23 frames of features as inputs, resulting in a 1472 dimensional ($64 \times 23$) feature vector.

The DNN is trained to estimate the IRM across all frequency bands, and the mean squared error is used as the loss function. Instead of predicting the IRM only for the center frame, a useful trick we found is to predict a window of IRM and then average the multiple estimates for each frame [15]. In this study, we predict 5 frames of the IRM (2 to each side of the center frame), which is found to converge well and give consistent improvements.

## 3 Results

### 3.1 Experimental setup

We use the IEEE corpus [5] containing a single male speaker as the training and test corpus, where the first 600 utterances are used for training and the remaining 120 for testing. The training utterances are mixed with about 10,000 non-speech sounds[1]. The total duration of the noises is about 125 hours. To create the training set, on average each noise gets mixed with 32 training utterances (randomly chosen each time), creating a total of 320,000 mixtures for training. The total duration of the training mixtures is about 250 hours. The test set is created by mixing test utterances with a variety of broadband non-stationary noises from different noise corpora (see details in the next subsection). Test noises are never seen during training. To isolate the effect of SNR, both training and test mixtures are created at -2 dB,

[1]Obtained from a sound effect library available at http://www.sound-ideas.com

Table 1: Effect of extensive training with a large variety of noises (SNR=-2 dB)

| | Babble-20 | | | Cafeteria | | |
|---|---|---|---|---|---|---|
| | STOI | HIT | FA | STOI | HIT | FA |
| Unprocessed | 0.63 | n/a | n/a | 0.61 | n/a | n/a |
| Train 100 noises | 0.70 | 89% | 49% | 0.72 | 82% | 18% |
| Train 10,000 noises | 0.79 | 77% | 10% | 0.79 | 76% | 4% |

Table 2: Segregation results in terms of STOI on a variety of novel noises (SNR=-2 dB)

| | Babble-20 | Cafeteria | Factory | Babble-100 | Living Room | Cafe | Park |
|---|---|---|---|---|---|---|---|
| Unprocessed | 0.63 | 0.61 | 0.62 | 0.62 | 0.84 | 0.71 | 0.73 |
| Noise-dependent model | 0.85 | 0.76 | 0.78 | 0.75 | 0.91 | 0.81 | 0.85 |
| Noise-independent model | 0.79 | 0.79 | 0.80 | 0.76 | 0.92 | 0.84 | 0.87 |

which is a very challenging condition that degrades speech intelligibility [3]. We will also show SNR generalization results in the next subsection.

For evaluation, we mainly use the Short-Time Objective Intelligibility score (STOI) [11] to measure the objective intelligibility. STOI measures a correlation of short-time temporal envelopes between clean and processed speech, and has been shown to be highly correlated to human speech intelligibility score.

### 3.2 Evaluation results

We first demonstrate the effectiveness of large-scale training with a variety of noises. A 20-talker babble (called "Babble-20") and a cafeteria noise from the Auditec CD[2] are used to create the test mixtures at -2 dB in this experiment. These two noises are highly non-stationary and symbolize daily adverse environments. We compare two DNN based IRM estimators that are trained using the same number of mixtures (320,000). The key difference is that one training set is created using 100 noises [4], whereas the other using 10,000 noises mentioned before. The 100-noise training set is created in a similar fashion as described above (i.e. on average each noise gets mixed 3,200 times). STOI results are listed in Table 1. To have a closer look at estimation errors, we convert the estimated IRM to the corresponding binary mask (by first converting ratio masks to local SNR estimates and then thresholding them at -7 dB) and compute the hit rate and the false alarm rate. The hit rate is the percent of correctly classified target-dominant (1's) units, and the false alarm rate is the percent of wrongly classified interference-dominant (0's) units. Similar to STOI, the hit minus false alarm rate (HIT−FA) is often used as a speech intelligibility predictor [6]. From Table 1, it is clear that the model trained on 100 noises performs significantly worse than the model trained on 10,000 noises. Specifically, the false alarm rates for both noises (especially for babble) are high, which could be detrimental to speech intelligibility [7]. By training with more noises, the false alarm rates for both noises are considerably reduced without significantly compromising hit rates. Although trained using the same number of mixtures, training with more noises

[2]Available at http://www.auditec.com

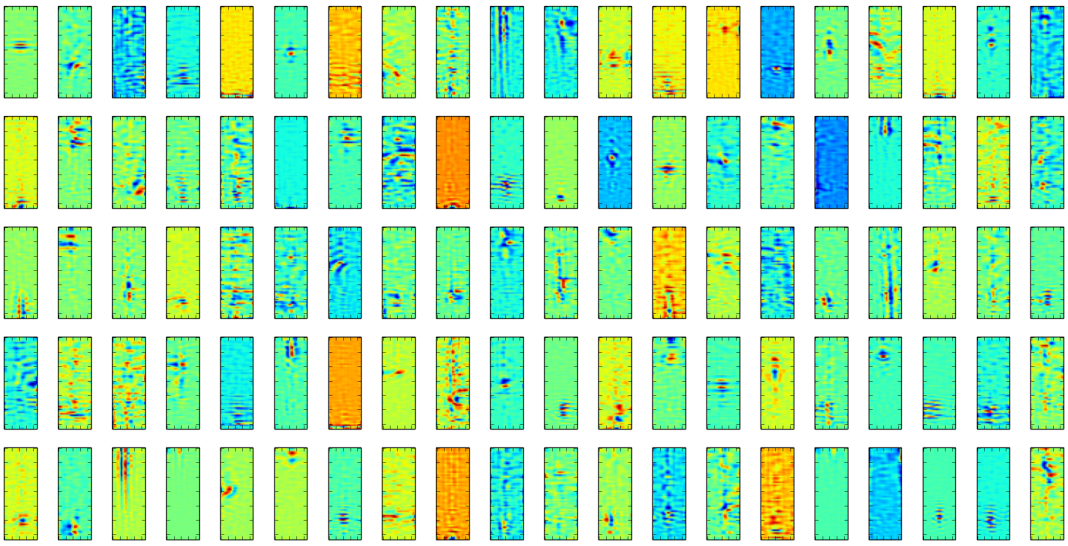
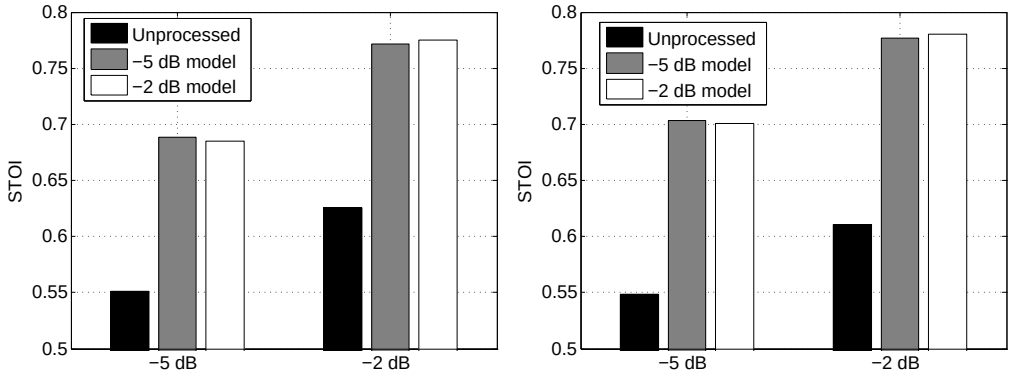


Figure 1: (Color online) Visualization of the first hidden layer filters from the noise-independent model.

clearly results in better STOI and HIT−FA rates. This indicates that it is the increase in the amount of distinct noise samples rather than the sheer size of the training set that improves generalization.

We refer to our model trained on 10,000 noises as the noise-independent model. Next, we present its performance on a variety of novel noises from different noise corpora. These noises include the 20-talker babble and the cafeteria noise used above, a factory and a 100-talker babble (called "Babble-100") from the NOISEX-92 corpus [13], and a living room, a cafe and a park noise from the DEMAND corpus [12]. To put the performance of the noise-independent model in perspective, in Table 2 we compare it with the DNN based IRM estimators that are trained and tested on the same noise. These noise-dependent models are trained on one part of the noise and tested on another, where the test noise segments do not overlap with the ones used in training. The two Auditec noises are 10 minutes long, and the noise-dependent models are trained on the first 8 minutes and tested on the remaining 2 minutes. The other noises are 4 minutes long, and the models are trained on the first 2 minutes and tested on the remaining 2 minutes. We emphasize that the noise-dependent DNN mask estimators represent very strong baselines, which are capable of improving speech intelligibility [2]. Specifically, the DNN based IRM estimators used in Healy *et al.* [2] are also trained on 8 minutes of the Auditec noises and tested on the remaining 2 minutes. Substantial intelligibility improvement is observed for hearing-impaired listeners in both noise types. Normal-hearing listeners also show substantial improvement in Babble-20. From Table 2, we can see that the noise-independent model consistently matches or even outperforms noise-dependent models except for Babble-20. The training set seems to adequately cover the test set distributions, such that the mismatch in




(a) Babble-20 (b) Cafeteria

Figure 2: STOI results of noise-independent models when trained and tested in matched and mismatched SNR conditions.

noise is well compensated for by training with a large number of noises. The performance degradation on the 20-talker babble noise is likely because this specific noise contains features similar to speech signal, which can not be sufficiently represented by the non-speech training noises. Nevertheless, we point out that 0.79 is already a strong STOI result at -2 dB input SNR.

Figure 1 visualizes the first 100 learned filters taken from the first hidden layer of the noise-independent model. Each panel in Fig. 1 represents a hidden unit, where the weights (coming from the input layer) are shown in two dimensions: x-axis represents time (23 frames) and y-axis represents frequency (64 frequency channels). We can see that the network learns what appears to be speech-specific feature detectors. For example, some filters resemble harmonic detectors (e.g. the 7th filter in the last row), while some others seem to capture feature transitions (e.g. the 7th filter in the third row). These speech-specific feature detectors encode fundamental characteristics of speech signal, enabling the model to be noise agnostic.

Previous experiments fix the SNR of the training and test mixtures to be -2 dB. In our last experiment, we show the performance with a mismatch between the training and test SNR. We train two noise-independent models as described before, where one model is trained exclusively on -2 dB mixtures and the other on -5 dB mixtures. We generate a -2 dB and -5 dB test set for each of the two Auditec noises and evaluate the two models on them. From Fig. 3.2, we can see that the STOI difference between the matched and mismatched SNR test conditions is negligible, which holds for both SNR levels and noises. This is likely because the model has seen enough local (e.g. frame level) SNR variations, even with a fixed, global (utterance level) SNR in training.

## 4 Concluding remarks

We have demonstrated that DNN based supervised speech segregation is capable of generalizing to novel noises if the network is trained on a large variety of noises. Specifically, a standard IRM estimator trained on about 10,000 noises is shown to match or even outperform the corresponding noise-dependent models that are already very strong baselines. As shown in Table 1, the current study represents a clear advance over our previous work [14] in segregation performance. Being able to generalize to novel noises makes a big stride towards a practical system for improving speech intelligibility or as a robust speech processing frontend.

We mainly consider the mismatch in background noise in this study. We believe that other generalization issues, such as the mismatch in speaker and room reverberation, can be tackled in a similar fashion.

**Acknowledgments**

This research was supported in part by an AFOSR grant (FA9550-12-1-0130), an NIDCD grant (R01 DC012048) and the Ohio Supercomputer Center.

## References

[1] G. E. Dahl, T. N. Sainath, and G. E. Hinton, "Improving deep neural networks for lvcsr using rectified linear units and dropout," in *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing*, 2013, pp. 8609–8613.

[2] E. Healy, S. Yoho, J. Chen, Y. Wang, and D. Wang, "An algorithm to improve speech intelligibility for hearing-impaired listeners in novel noise segments," *submitted*, 2015.

[3] E. Healy, S. Yoho, Y. Wang, and D. Wang, "An algorithm to improve speech recognition in noise for hearing-impaired listeners," *Journal of the Acoustical Society of America*, pp. 3029–3038, 2013.

[4] G. Hu, "100 nonspeech environmental sounds (http://www.cse.ohio-state.edu/pnl/corpus/HuCorpus.html)," 2004.

[5] IEEE, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.*, vol. 17, pp. 225–246, 1969.

[6] G. Kim, Y. Lu, Y. Hu, and P. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *Journal of the Acoustical Society of America*, pp. 1486–1494, 2009.

[7] N. Li and P. Loizou, "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," *Journal of the Acoustical Society of America*, vol. 123, no. 3, pp. 1673–1682, 2008.

[8] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2007.

[9] T. May and T. Dau, "Requirements for the evaluation of computational speech segregation systems," *Journal of the Acoustical Society of America*, vol. 136, no. 6, pp. EL398–EL404, 2014.

[10] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing*, 2013, pp. 7092–7096.

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

[12] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database: A database of multichannel environmental noise recordings," *Journal of the Acoustical Society of America*, vol. 133, no. 5, p. 3591, 2013.

[13] A. Varga and H. Steeneken, "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 Communication*, vol. 12, pp. 247–251, 1993.

[14] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, pp. 1381–1390, 2013.

[15] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, pp. 1849–1858, 2014.