Long Short-Term Memory for Speaker Generalization in Supervised Speech Separation

Jitong Chen[1], *DeLiang Wang*[1,2]

[1]Department of Computer Science and Engineering, The Ohio State University, USA
[2]Center for Cognitive and Brain Sciences, The Ohio State University, USA

chenjit@cse.ohio-state.edu, dwang@cse.ohio-state.edu

Abstract

Speech separation can be formulated as a supervised learning problem where a time-frequency mask is estimated by a learning machine from acoustic features of noisy speech. Deep neural networks (DNNs) have been successful for noise generalization in supervised separation. However, real world applications desire a trained model to perform well with both unseen speakers and unseen noises. In this study we investigate speaker generalization for noise-independent models and propose a separation model based on long short-term memory to account for the temporal dynamics of speech. Our experiments show that the proposed model significantly outperforms a DNN in terms of objective speech intelligibility for both seen and unseen speakers. Compared to feedforward networks, the proposed model is more capable of modeling a large number of speakers, and represents an effective approach for speaker- and noise-independent speech separation.

Index Terms: speech separation, speaker generalization, long short-term memory

1. Introduction

Monaural speech separation is a challenging problem with many applications such as hearing aid design and robust automatic speech recognition (ASR). One way to deal with this problem is to apply speech enhancement [1] [2] [3], which has limited success in highly nonstationary noises. Another way is to apply masking to a time-frequency (T-F) representation of noisy speech, where an ideal mask keeps speech-dominant T-F units and discards noise-dominant T-F units. Supervised learning can be employed to estimate a T-F mask from acoustic features extracted from noisy speech. This data-driven approach, which does not make any assumption about the statistical distribution of underlying speech or noise signals, represents a new strategy to deal with highly nonstationary noises.

In supervised speech separation, the ideal binary mask (IBM) [4], which classifies T-F units as speech-dominant or noise-dominant, is typically used as the learning target. A recent study has shown that speech separation by estimating the IBM using deep neural networks (DNNs) leads to substantial speech intelligibility improvement [5]. Alternatively, one can perform speech separation by estimating the ideal ratio mask (IRM) [6] [7]:

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

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

where $S(t,f)^2$ and $N(t,f)^2$ denote speech energy and noise energy of a T-F unit at time t and frequency f, respectively. It has been shown that ratio masking leads to better speech quality than binary masking [6]. A recent study has shown that IRM estimation improves speech intelligibility of hearing-impaired listeners [8]. In this study, we use the IRM as the learning target of supervised speech separation.

For supervised learning tasks, generalizing to unseen conditions is a critical issue. Noise generalization and speaker generalization are two important problems in supervised speech separation. With enough training noises but a fixed speaker, a DNN is able to generalize to unseen noises [9]. However, it remains unknown how well such a model generalizes to an unseen speaker and an unseen noise at the same time.

In this study, we first train a DNN based speech separation system with multiple speakers and test it on both seen and unseen speakers. The experimental results show that a DNN does not perform well with multi-speaker training. Unlike a single-speaker-dependent model, a DNN trained on multiple speakers tries to detect many distinctive speech patterns of different speakers. This poses a challenge of separating a target speaker from interference especially when the background noise (such as babble noise) includes speech components. With more training speakers, we observe performance degradation of a DNN on seen speakers. The input to a DNN is typically a limited temporal window of acoustic features, which are not sufficient to decide the target speaker to focus on since the energy of target speech and noise fluctuates over time and the local SNR varies. In comparison, a single-speaker-dependent DNN attends to a specific speaker and hence better resolves the ambiguity of target speech and noise.

To deal with multiple speakers, one can train many single-speaker-dependent separation models and use a speaker classification module for model selection. This approach introduces three challenges. First, for seen speakers, the overall performance of this approach depends on speaker classification accuracy in noisy environments, which is known to be challenging

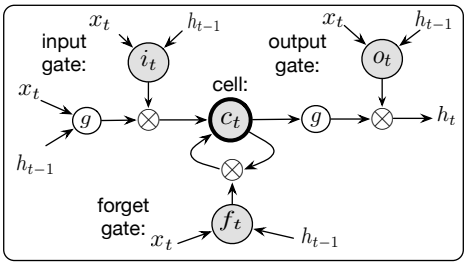


Figure 1: Diagram of LSTM block.

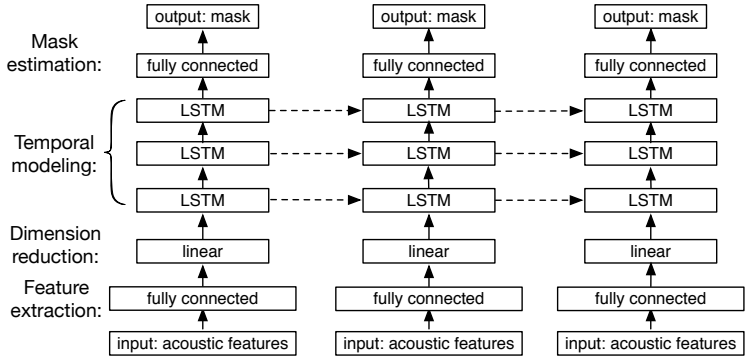


Figure 2: Diagram of the proposed system.

in low SNRs. Second, for unseen speakers, one has to consider the best strategy of assigning an unseen speaker to a trained speaker. Third, it is impractical to fit a large number of single-speaker-dependent models in portable devices such as hearing aids.

In this study, we propose a unified model that separates speech of an unseen speaker from an unseen noise. Our model is built with a recurrent neural network (RNN) which models the temporal dynamics of speech. Given an incoming stream of noisy speech, the model performs speech separation and analyzes speaker patterns at the same time. The model is able to utilize long-term context to better focus on a target speaker. We describe the proposed model in Section 2 and present experimental results in Section 3. Section 4 concludes the paper.

2. System description

In speaker-independent speech separation, to stay focused on one of many potential target speakers, a model need to take into account long-term context. RNNs are designed to model temporal dependencies and typically trained with back propagation through time (BPTT). A vanillar RNN suffers from the exploding and vanishing gradient problem during BPTT [10]. The long short-term memory (LSTM) [11], a specific type of RNN, mitigates this problem by introducing a memory cell, which facilitates information flow over time. LSTMs have been successful in modeling long temporal dependencies in many applications such as language modeling [12] [13], acoustic modeling [14] [15] and video classification [16]. As shown in Fig. 1, an LSTM block has a memory cell and three gates where the forget gate controls how much previous information should be erased from the cell and the input gate controls how much information should be added to the cell. In this study, we use the LSTM defined by the following equations:

$$i_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1} + b_i) \quad (2)$$

$$f_t = \sigma(W_{fx}x_t + W_{fh}h_{t-1} + b_f) \quad (3)$$

$$o_t = \sigma(W_{ox}x_t + W_{oh}h_{t-1} + b_o) \quad (4)$$

$$z_t = g(W_{zx}x_t + W_{zh}h_{t-1} + b_z) \quad (5)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot z_t \quad (6)$$

$$h_t = o_t \odot g(c_t) \quad (7)$$

where x_t, z_t, c_t, h_t represent input, block input, memory cell and hidden activation at time t, respectively. Input gate, forget gate and output gate are denoted as i_t, f_t and o_t. σ represents $sigmoid$ function and g represents $tahn$ function. W's and b's denote linear transforms and biases, respectively. $\odot$ denotes element-wise multiplication.

The proposed system is shown in Fig. 2. We use a fully connected layer for feature extraction, three LSTM layers for temporal modeling, and two fully connected layers and an output layer for IRM estimation. A linear layer is inserted as a dimension reduction layer to speed up training. We use rectified linear units (ReLUs) for all fully connected hidden layers. The output layer uses $sigmoid$ activation since a ratio mask has the value range of $[0, 1]$. A feature window of 23 (11 on the left, 11 on the right) are fed to the network to estimate one frame of the mask. The input features are 64-dimension gammatone filterbank energies [9] extracted from noisy speech and the IRM is defined on a 64-channel cochleagram [17] with a 20-ms window and a 10-ms shift. From the input layer to the output layer, the proposed network has 64×23, 1024, 512, 512, 512, 512, 512, 512 and 64 units, respectively. We compare our proposed system with a DNN baseline, which has five hidden layers with ReLU activation. From the input layer to the output layer, the DNN has 64×23, 2048, 2048, 2048, 2048, 2048 and 64 units, respectively.

3. Experimental results

3.1. Experimental setup

3.1.1. Data preparation

In our experiments, we use 7138 utterances (83 speakers) from the WSJ0 SI-84 training set [18], two highly-nonstationary test noises (babble and cafeteria) from the Auditec CD (available at http://www.auditec.com), and 10,000 training noises from a sound effect library (available at http://www.sound-ideas.com). Of the 83 speakers, 6 speakers are treated as unseen speakers for multi-speaker models. In other words, all multi-speaker models are trained with speakers drawn from the 77 remaining speakers (about 14 hours of speech). Since we study speaker generalization of noise-independent models, the two test noises are not used for training.

We create two test sets using 12 speakers and the babble noise. Each test mixture is created by mixing an utterance and a random cut from a 2-minute segment of the babble noise at -5 dB SNR. The two test sets are:

- Test Set 1: 150 mixtures are created from 25×6 utterances of 6 seen speakers (3 males and 3 females).
- Test Set 2: 150 mixtures are created from 25×6 utterances of 6 unseen speakers (3 males and 3 females).

To evaluate speaker generalization, we train four types of models using various numbers of speakers and the 10,000 sounds.

Each training mixture is created by mixing a randomly selected utterance and a random segment from the 10,000 sounds at a random SNR drawn from $\{-5, -4, -3, -2, -1, 0\}$ dB. All test utterances are excluded from the training sets. We categorize the trained models as follows.

- Single-speaker-dependent models:
  For each speaker in Test Set 1 and Test Set 2, we train and test on the same speaker. Each training set has 320,000 mixtures.
- Multi-speaker-dependent model:
  One model is trained and tested with the 6 speakers of Test Set 1. Each training set has 3,200,000 mixtures.
- Expanded multi-speaker-dependent models:
  Four models are trained with $\{10, 20, 40, 77\}$ speakers including the 6 speakers of Test Set 1 and evaluated with Test Set 1. Each training set has 3,200,000 mixtures.
- Speaker-independent models:
  Five models are trained with $\{6, 10, 20, 40, 77\}$ speakers and tested on the 6 unseen speakers of Test Set 2. Each training set includes 3,200,000 mixtures.

3.1.2. Training details

We train the DNN and LSTM with the mean square error (MSE) cost function and the Adam optimizer [19]. The initial global learning rate is set to 0.001 and reduced by half every epoch. The best model is selected by cross validation. We use a mini-batch size of 256 for single-speaker-dependent DNNs. A mini-batch size of 4096 is used for multi-speaker DNNs as we find a larger batch size slightly improves optimization for multi-speaker DNNs. All LSTMs are trained with a mini-batch size of 256 and a step size of 250.

3.2. Experimental results and analysis

We evaluate the impact of the number of training speakers on the DNN and LSTM. The MSE of the estimated mask and the short-time speech intelligibility (STOI) [20] are used for analysis. The STOI, which compares the envelopes of separated speech and clean speech, has been shown to well correlate with human speech intelligibility [8].

3.2.1. Performance trend on seen test speakers

We first analyze the performance of the DNN and LSTM on 6 seen speakers with an increasing number of training speakers. As shown in Fig. 3 (a) and Fig. 5 (a), starting from 10 training speakers, adding more training speakers hurts the performance of the DNN on the 6 seen speakers, indicating that the DNN fails to model a large number of speakers. As the DNN is exposed to more training speakers, it becomes more challenging to separate a target speaker from the babble noise, whose local spectral-temporal patterns resemble the ones of some training speakers. Hence the DNN is more likely to mistake background noise as the target speech. In contrast, Fig. 3 (b) and Fig. 5 (a) show that although the LSTM is prone to overfitting when the number of training speakers is small, it keeps improving as more training speakers are added. The performance of the LSTM on the 6 seen speakers are boosted by adding more and more training speakers. This indicates that the LSTM has the capacity of modeling a large number of speakers.

3.2.2. Performance trend on unseen test speakers

In terms of the performance on the 6 unseen speakers, Fig. 3 (c), Fig. 3 (d) and Fig. 5 (b) show that both the DNN and LSTM improves as the number of training speakers increases. However, the LSTM significantly outperforms the DNN in terms of the mask MSE and the STOI improvement. The STOI improvement of the LSTM is 4% higher than the DNN with babble noise at -5 dB SNR. This indicates that the LSTM achieves better speaker generalization than the DNN. In particular, visualization of estimated masks in Fig. 4 shows that the LSTM reduces the error of mistaking the background babble noise for the target speech (e.g. between time frame 800 and time frame 900).

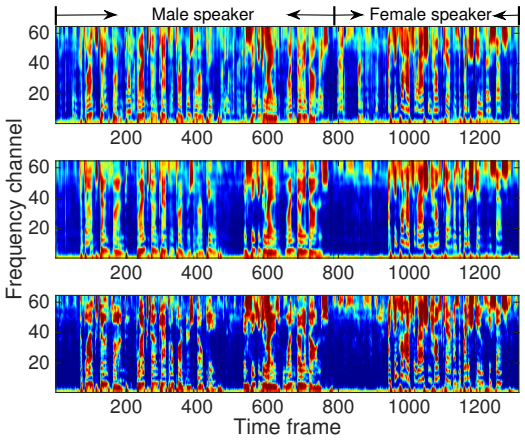


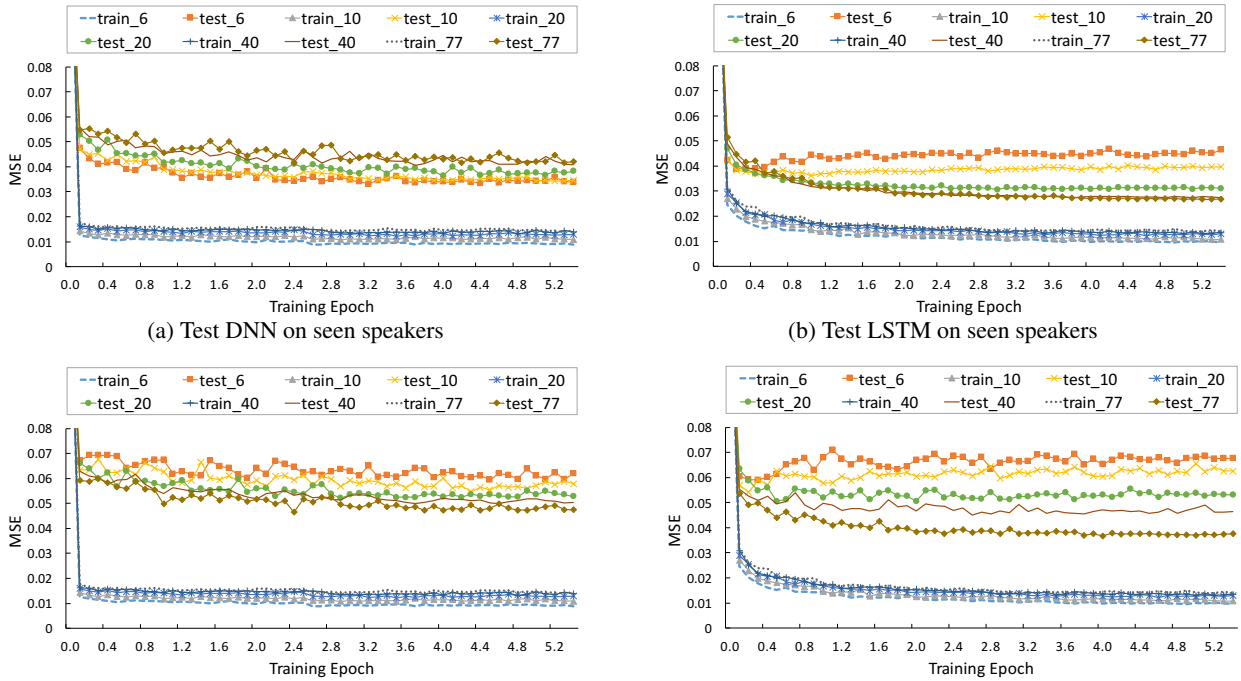
Figure 4: Visualization of the predicted masks by DNN (top) and LSTM (middle) and the ground truth mask (bottom). The first and second part of the mask are estimated from noisy speech of a male and a female with babble noise at -5 dB, respectively.

3.2.3. Model comparisons

The models are evaluated with the babble and cafeteria noise at -2 dB and -5 dB SNR. One the one hand, Table 1 shows that the expanded muti-speaker-dependent LSTM with 77 training speakers outperforms both single-speaker-dependent DNNs and expanded multi-speaker-dependent DNNs, indicating that the LSTM learns from other speakers to improve the performance on seen speakers. On the other hand, as shown in Table 2, the speaker-independent LSTM with 77 training speakers generalizes better to unseen speakers than DNNs, especially at the very low SNR of -5 dB. Besides, the performance gap between the speaker-independent LSTM and single-speaker-dependent models is small. We project that the performacne of the speaker-independent LSTM will further improve with more training speakers, following the trend shown in Fig. 5 (b).

4. Discussion

In this study, we have demonstrated that the proposed LSTM based speech separation model generalizes better to unseen speakers than a DNN in the context of noise-independent speech separation. Interestingly, as the number of training speakers increases, the performance of the DNN on trained speakers degrades, while the performance of the LSTM improves. This reveals the capacity of the LSTM in modeling a large number of speakers. With more training speakers, the LSTM improves with unseen speakers and significantly outperforms a DNN. The proposed model, which incorporates the temporal dynamics of speech, represents a major step towards speaker- and noise-independent speech separation.



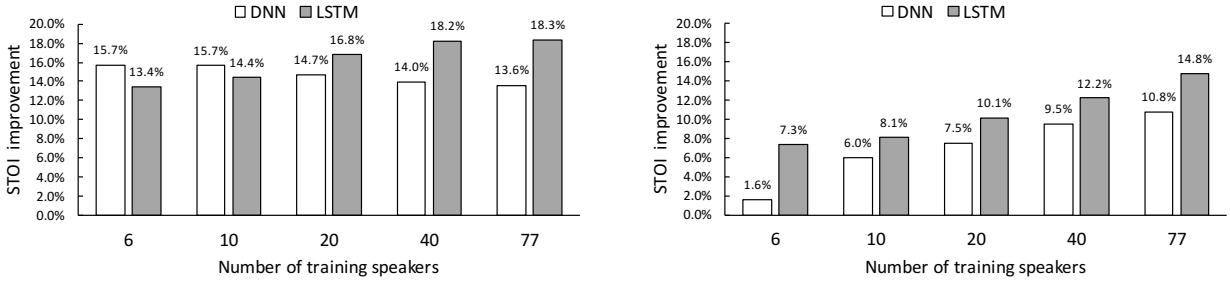
(a) Test DNN on seen speakers

(b) Test LSTM on seen speakers

(c) Test DNN on unseen speakers

(d) Test LSTM on unseen speakers

Figure 3: Training and test errors of DNN and LSTM. All models are evaluated with a test set of 6 seen speakers and a test set of 6 unseen speakers. Training mixtures are created with $\{6, 10, 20, 40, 77\}$ speakers. The two test sets are created with the babble noise at -5 dB SNR. All models are noise independent. (a) shows that the performance of DNN on the 6 seen speakers drops as the number of speakers increases. In contrast, (b) shows that LSTM is able to improve the results on the 6 seen speakers when more speakers are added. (c) and (d) show that both DNN and LSTM achieve better performance on the 6 unseen speakers as the number of speakers increases, and that LSTM outperforms DNN when a large number of speakers are used for training.



(a) Seen speakers with -5 dB SNR

(b) Unseen speakers with -5 dB SNR

Figure 5: Comparison of the DNN and LSTM in terms of STOI improvement (processed STOI $-$ unprocessed STOI). The LSTM outperforms the DNN when a large number of speakers are used for training.

Table 1: Comparison of expanded multi-speaker-dependent models (trained on 77 speakers and tested on 6 seen speakers) and single-speaker-dependent models in terms of processed STOI (in %). The unprocessed STOI is shown in parenthesis.

Model	-5 dB babble	-5 dB cafeteria	-2 dB babble	-2 dB cafeteria
Expanded multi-speaker-dependent DNN	72.2 (58.6)	71.7 (57.4)	80.2 (66.2)	79.6 (65.2)
Expanded multi-speaker-dependent LSTM	76.9 (58.6)	74.7 (57.4)	83.3 (66.2)	81.6 (65.2)
Single-speaker-dependent DNN	74.3 (58.6)	72.9 (57.4)	81.1 (66.2)	80.0 (65.2)

Table 2: Comparison of speaker-independent models (trained on 77 speakers and tested on 6 unseen speakers) and single-speaker-dependent models in terms of processed STOI (in %). The unprocessed STOI is shown in parenthesis.

Model	-5 dB babble	-5 dB cafeteria	-2 dB babble	-2 dB cafeteria
Speaker-independent DNN	69.1 (58.3)	70.9 (57.2)	77.7 (65.3)	78.8 (64.6)
Speaker-independent LSTM	73.1 (58.3)	72.5 (57.2)	80.4 (65.3)	79.8 (64.6)
Single-speaker-dependent DNN	74.5 (58.3)	73.5 (57.2)	80.9 (65.3)	80.2 (64.6)

5. References

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Sig. Process.*, vol. 32, pp. 1109–1121, 1984.

[2] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete fourier coefficients with generalized gamma priors," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, pp. 1741–1752, 2007.

[3] P. C. Loizou, *Speech enhancement: theory and practice*. Boca Raton FL: CRC press, 2007.

[4] D. L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech separation by humans and machines*, P. Divenyi, Ed. Boston MA: Kluwer Academic Pub., 2005, pp. 181–197.

[5] E. W. Healy, S. E. Yoho, Y. Wang, and D. L. Wang, "An algorithm to improve speech recognition in noise for hearing-impaired listeners," *J. Acoust. Soc. Am.*, vol. 134, pp. 3029–3038, 2013.

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

[7] J. Chen, Y. Wang, and D. L. Wang, "Noise perturbation for supervised speech separation," *Speech Communication*, vol. 78, pp. 1–10, 2016.

[8] E. W. Healy, S. E. Yoho, J. Chen, Y. Wang, and D. L. Wang, "An algorithm to increase speech intelligibility for hearing-impaired listeners in novel segments of the same noise type," *J. Acoust. Soc. Am.*, vol. 138, pp. 1660–1669, 2015.

[9] J. Chen, Y. Wang, S. E. Yoho, D. L. Wang, and E. W. Healy, "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises," *J. Acoust. Soc. Am.*, vol. 139, pp. 2604–2612, 2016.

[10] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *Proc. ICML*, 2013, pp. 1310–1318.

[11] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, pp. 1735–1780, 1997.

[12] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. NIPS*, 2014, pp. 3104–3112.

[13] M. Sundermeyer, H. Ney, and R. Schlüter, "From feedforward to recurrent lstm neural networks for language modeling," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, pp. 517–529, 2015.

[14] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. ICASSP*, 2013, pp. 6645–6649.

[15] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling." in *Proc. INTERSPEECH*, 2014, pp. 338–342.

[16] J. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *Proc. CVPR*, 2015, pp. 4694–4702.

[17] D. L. Wang and G. J. Brown, Eds., *Computational auditory scene analysis: Principles, algorithms and applications*. Hoboken NJ: Wiley-IEEE Press, 2006.

[18] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Proc. of the workshop on Speech and Natural Language*, 1992, pp. 357–362.

[19] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.

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