Interspeech 2018
2-6 September 2018, Hyderabad

Bubble Cooperative Networks for identifying important speech cues

Viet Anh Trinh[1], *Brian McFee*[2,3] *and Michael I Mandel*[1,4]

[1] The Graduate Center, CUNY, New York, USA
[2] Center for Data Science, New York University, USA
[3] Music and Audio Research Laboratory, New York University, USA
[4] Brooklyn College, CUNY, New York, USA

vtrinh@gradcenter.cuny.edu, brian.mcfee@nyu.edu, mim@sci.brooklyn.cuny.edu

Abstract

Predicting the intelligibility of noisy recordings is difficult and most current algorithms treat all speech energy as equally important to intelligibility. Our previous work on human perception used a listening test paradigm and correlational analysis to show that some energy is more important to intelligibility than other energy. In this paper, we propose a system called the Bubble Cooperative Network (BCN), which aims to predict important areas of individual utterances directly from clean speech. Given such a prediction, noise is added to the utterance in unimportant regions and then presented to a recognizer. The BCN is trained with a loss that encourages it to add as much noise as possible while preserving recognition performance, encouraging it to identify important regions precisely and place the noise everywhere else. Empirical evaluation shows that the BCN can obscure 97.7% of the spectrogram with noise while maintaining recognition accuracy for a simple speech recognizer that compares a noisy test utterance with a clean reference utterance. The masks predicted by a single BCN on several utterances show patterns that are similar to analyses derived from human listening tests that analyze each utterance separately, while exhibiting better generalization and less context-dependence than previous approaches.

Index Terms: Auditory importance, neural network interpretation, noise robustness, speech cues, deep learning.

1. Introduction

Noise and reverberation are among the biggest problems in automatic speech recognition (ASR) [1], hearing aids [2, 3], and other speech communication technologies [4]. These systems are in fact still much less noise robust than normal human listeners [5, 6]. One theory of the remarkable noise robustness of human speech perception is that listeners are able to identify glimpses of relatively clean speech in a noisy mixture through bottom-up processes and then use top-down knowledge of speech and language to fill in the missing information between these glimpses [7]. Our paper aims to train models to create glimpses through a noise field that are maximally useful to a listener, in this case a simple automatic speech recognizer. In order to do this, the model must predict the importance of individual time-frequency regions of an utterance to its being correctly identified by the recognizer. As a byproduct, such a model provides insight into the cues used by the recognizer to identify speech in noise, allowing direct comparisons of them for different systems. With certain modifications, they could also be compared with those used by human listeners.

The combination of noise generator and discriminative recognizer leads to a network that is related to a generative adversarial network (GAN) [8], but differs in several important respects. First, instead of generating entirely new signals, the generator component of the model creates masks that are applied to noise and then added to the speech. Second, instead of the generator and discriminator competing against one another, they are cooperating to correctly identify the speech in the presence of a maximal amount of noise. Thus we call this combination of components the bubble cooperative network.

2. Relation to prior work

The proposed technique for identifying important speech cues builds on our previous work to do so using randomized "bubble noise" stimuli [9]. In that system, an individual utterance was mixed with many different instances of "bubble noise", very loud speech-shaped noise with bubbles of silence placed at random times and frequencies. The intelligibility of each mixture was measured by presenting it to a listener and the importance of individual time-frequency points was characterized by the correlation across mixtures between the audibility of the speech at each point with the intelligibility of the mixture. We used this technique to directly compare the cues used by human listeners with those used by an ASR based on MFCCs and a GMM-HMM acoustic model [10], and found them to be quite different. Due to the need for many mixtures of each clean utterance, the technique of [9] requires approximately 10 minutes of listening time to analyze a single utterance.

Classifiers trained on the data from [9] to predict whether a given mixture would be intelligible to a listener were able to generalize to new productions of the same words, new talkers, and to some extent new words. Those results utilized a different classifier trained for each word, making it somewhat cumbersome to generalize these predictions to new contexts. The proposed BCN, in contrast, provides a single model predicting importance for all words, making it much more straightforward to generalize to new words and new contexts. In addition, the task performed by listeners in [9] is a forced choice between a small, closed set of options, causing the importance for one utterance to potentially be influenced by the options it is contrasted against. In the proposed work, the BCN predicts a single mask for a given utterance that must maintain its intelligibility in all contexts, making it more informative about the utterance itself. See Figures 2 and 3 for a comparison of predicted machine importance functions and measured human importance functions on the same utterances.

The BCN provides insight into why the recognizer makes a particular decision and there is a great deal of interest in techniques of this nature in the field of machine learning to aid in model development and to provide explanations that could build trust with consumers of model predictions and decisions. [11] searched for data points that maximally activated particular neurons. [12] proposed an approach that approximates the partial derivative of a particular network output with respect to input

10.21437/Interspeech.2018-2377

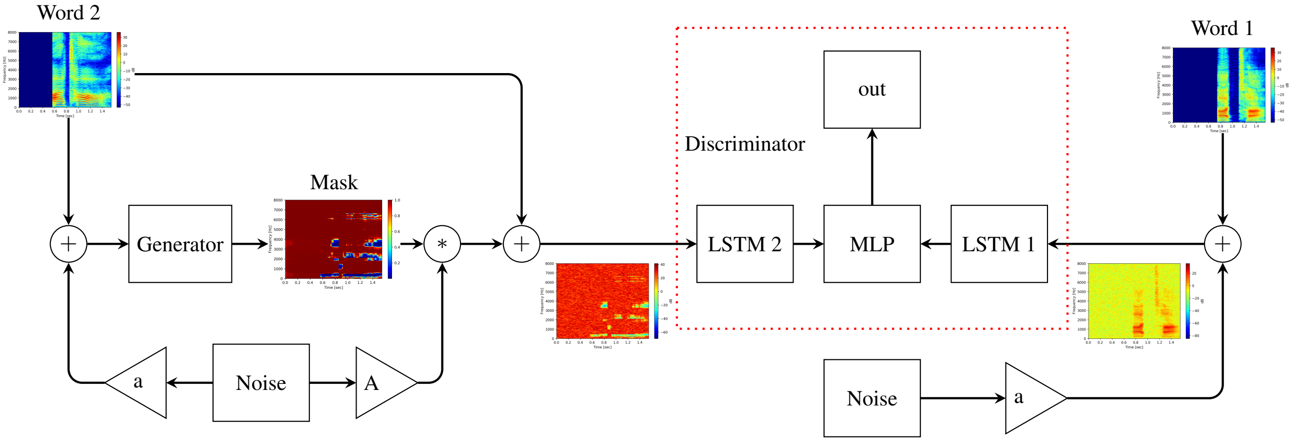

Figure 1: *Network diagram. In this example, word 1 is 'aTa' from speaker M5 and word 2 is 'aTa' from speaker W5. The network successfully identified that the two words are the same with prediction confidence 0.84.*

pixels in a convolutional neural network [13], similar to [14]. Layer-wise relevance propagation [15] proposes instead examining the geometry of the decision boundary close to a given observation to characterize the importance of each input dimension (e.g., pixel). [16] describe Linear Interpretable Model-agnostic Explanations, which trains an interpretable classifier based on local predictions of a more complicated classifier under analysis.

Our analyses provide a time-frequency representation of importance. Several approaches are popular for utilizing similar time-frequency masks for speech enhancement [3, 17] and noise robustness in ASR [18]. These approaches typically optimize criteria related to the proportion of speech energy that is correctly classified, treating all energy as equally important. Focusing instead on importance should more directly solve the problem of improving intelligibility. Models that touch on this to some extent are those of perceptual salience and attention [19], although they tend to focus on the perception of environmental soundscapes and longer-term sounds, and do not provide much detail at the level of individual phonemes. A similar trend of one-dimensional attention in deep learning models has been popular recently in the sequence-to-sequence framework [20], which has been successfully applied to direct audio-to-character ASR [21]. These systems have not been extensively evaluated on noisy tasks, for which the BCN could potentially provide benefit as a two-dimensional time-frequency attention.

3. Method

3.1. Network structure

Figure 1 shows a schematic of the BCN system. It consists of a noise mask generator and a discriminative recognizer. The **mask generator** takes clean speech as input and produces a mask that aims to reveal as little of the speech as possible while still allowing the recognizer to correctly identify the words spoken. Its input includes a small amount of dither noise so that its output is not deterministic. We have found this to facilitate generalization of the model. This mask is point-wise multiplied by the spectrogram of a sample of white noise generated in the time domain and added to the speech. The **recognizer** (or discriminator in the terminology of GANs) takes as input this noisy utterance along with a clean utterance from another talker (plus again a small amount of dither noise). The recognizer predicts whether the two utterances contain the same words or not and the dataset is designed so that the words match in only half of the training and testing instances. In this work, each utterance contains a single isolated word for simplicity. The discriminator consists of two Long Short Term Memory (LSTM) networks and a multilayer perceptron (MLP). Each LSTM processes one of the utterances into a fixed-size hidden representation, which the MLP uses to predict whether they contain the same word or not. Note that this very simple recognizer is used so that the entire system can be implemented easily in TensorFlow. It could be any speech recognition system that can be trained via gradient descent.

Denote the two clean utterances and the white noise in the time domain as $x_1(t), x_2(t)$ and $n(t)$, respectively. The corresponding short time Fourier transformations (STFTs) of these signals are $X_1(f,t), X_2(f,t)$ and $N(f,t)$, with frequency index f and time index t. When used without indices, these variables represent entire matrices. The output of the mask generator, with parameters θ, is

$$M_\theta = G(X_2 + aN) \in [0,1]^{F\times T} \tag{1}$$

where a is a small constant. This mask is multiplied by the noise and provided to the discriminator, D, along with the reference utterance, $X_1(t,f)$

$$\hat{y} = D(X_1 + aN, X_2 + AN \odot M_\theta) \in [0,1] \tag{2}$$

where A is a large constant and $\odot$ is the point-wise multiplication operator. The discriminator should output 1 if it predicts the words in these two utterances are the same and 0 otherwise.

In order to train the system, we minimize a loss function with several terms.

$$\begin{aligned}\mathcal{L}(\theta) = {} & \lambda_d \mathcal{L}_\mathrm{D}(y,\hat{y}) - \frac{\lambda_n}{TF}\sum_{f,t} M_\theta \\ & - \frac{\lambda_e}{TF}\sum_{f,t}(M_\theta \log M_\theta + (1-M_\theta)\log(1-M_\theta)) \\ & + \frac{\lambda_f}{TF}\sum_{f,t}|\Delta_f M_\theta| + \frac{\lambda_t}{TF}\sum_{f,t}|\Delta_t M_\theta|. \end{aligned} \tag{3}$$

The term weighted by λ_d is the recognition loss of the prediction $\hat{y}$ from the discriminator. In the case of this simple recognizer, this is the cross-entropy between the binary target and the prediction. This term encourages the discriminator to maintain its correct identification of the speech. The term weighted by λ_n encourages the mask to contain as many 1's as possible, due to its preceding negative sign, maximizing the amount of noise. The term weighted by λ_e encourages lower entropy of the mask entries, so that they are closer to either 0 or 1. The terms weighted by λ_f and λ_t together comprise a total variation penalty, but with different weighting in the time and frequency directions, encouraging the mask to be piece-wise constant. The Δ_f and Δ_t operators represent the first difference along frequency and time, respectively. Note that the continuity encouraged by the total variation penalty leads to masks that are more interpretable, but lower resolution.

4. Experiments

4.1. Datasets

We perform experiments on the speech material from Shannon et al [22]. This dataset includes all combinations of three vowels and 20 consonants in consonant-vowel (CV) and vowel-consonant-vowel (VCV) syllables. The vowels are /ɑ/, /i/, and /u/, so the words are of the form "aCa", "eeCee", and "ooCoo" for medial consonants and "Ca", "Cee", and "Coo" for initial consonants. The 20 consonants are /b, d, g, p, t, k, m, n, l, r, w, j, f, v, s, z, ʃ, ð, tʃ, dʒ/. We used recordings of these words from eight talkers, four men (M1, M3, M4, M5) and four women (W1, W3, W4, W5). The dataset recommends avoiding M2 and W2. This gives a total of 960 utterances made up of 8 speakers, 2 forms, 3 vowels, and 20 consonants. The signals were sampled at 44.1 kHz. This simple stimulus set allows us to focus on developing the BCN technique.

4.2. Training the network

We use the Librosa library [23] to transform the time domain signal into log-magnitude spectrograms by short-time Fourier transform (STFT) with a frame size of 64 milliseconds (ms), a hop size of 16 ms. Because the words have different durations, all utterances are padded to be the same length by inserting zeros before the speech. All operations on audio signals occur in the complex STFT domain. Before the complex STFTs are input to any neural network layer, the square magnitude is derived from the sum of square of real and imaginary matrices, and then the magnitude is converted to dB and normalized across time to have zero mean and unit variance. After this processing, the spectrograms have 1412 frequency bands and 94 time steps. Finally, we represent complex STFTs by stacking the real and imaginary matrices on top of one another in Tensorflow.

The generator is an LSTM network. The discriminator includes two LSTMs and an MLP. The hidden representations of the two LSTMs at the last time step of each utterance are concatenated together and fed into the MLP, which has two hidden layers and a single sigmoid output. The MLP uses the ReLU activation function. The LSTM weights are initialized using the Glorot method [24] and the MLP weights are initialized using the method of [25]. All the biases are initialized to 0. The network is trained using back propagation with Adam stochastic gradient descent [26].

4.3. Experiments

We performed two experiments to evaluate the BCN. The first uses just the discriminator to show that it can successfully recognize words from different speakers without additional noise. The second trains the generator to add noise to the input of the pre-trained discriminator.

Table 1: *Recognition accuracy (%) for the BCN (exps 1 and 2)*

Model	**Training**	**Development**	**Test**
Discriminator	89.1	85.4	84.0
Discriminator + Generator	88.5	84.8	83.8

Experiment 1: In this experiment, our network only includes the discriminator without the generator. Its purpose is to train the discriminator and find the best hyperparameters for it based on classification accuracy on the development set using a randomized hyperparameter search. Speakers M1, M3, W1, and W3 were used for the training set; M4 and W4 were used for the development set; and M5 and W5 were used for the test set. To train the model, we form pairs of words, with equal numbers of matching pairs and non-matching pairs. For each word in the training set, we generate three positive pairs by matching it with the same word spoken by the three other talkers, and three negative pairs by matching it with three randomly selected non-matching words, also from other talkers. Similarly, for the development and test sets, there is one positive pair and one negative pair, since a production is never paired with itself. Thus there are 2880, 480, and 480 pairs in the training, development, and test sets, respectively. The dither noise reduces over-fitting on this relatively small dataset by introducing variability in both the masks and the speech references between epochs. All models were trained using early stopping on the development set, using the weights from the epoch with the highest development set accuracy. The discriminator with the best development set accuracy has the following hyperparameters: both LSTMs have 200 hidden units, the MLP has two hidden layers consisting of 100 units each, the learning rate is 6×10^{-5}, batch size is 24, and the dither noise (a in Figure 1) has the value 0.05.

Experiment 2: In this experiment, we add the generator to the discriminator trained in experiment 1, the discriminator parameters are frozen. It uses the same input pairs as experiment 1. Its purpose is to find the mask that reveals as little speech as possible while allowing the discriminator to correctly identify the speech. The generator that minimizes the loss (3) on the development set has the following parameters: the LSTM has 100 hidden units, the gain $A = 4.0$, $\lambda_d = 2.0$, $\lambda_n = 1.1$, $\lambda_e = 0.05$ $\lambda_f = 20$, and $\lambda_t = 0.03753$.

5. Results

Table 1 shows the results of experiments 1 and 2. As expected, accuracy for both models is slightly higher on the training set than on the development set. Early stopping prevents this gap from growing too large. Additionally, both models generalize well to the test set of utterances from completely different talkers, achieving classification accuracies of 84.0% and 83.8%. This is well above the chance level of 50%, showing that this simple speech recognizer can accurately generalize across productions of the same word from different talkers. In addition, we compute that on average, 97.7% of spectrogram points on the 480 test utterances are obscured by noise (a mask value of at least 0.95), yielding an average signal to noise ratio (SNR) of -27.29 dB. The fact that the discriminator can achieve almost the same word identification accuracy when the generator obscures almost all of the test word with noise shows that the generator can accurately predict important regions from the clean speech.

Figure 2 shows example masks created by the generator on

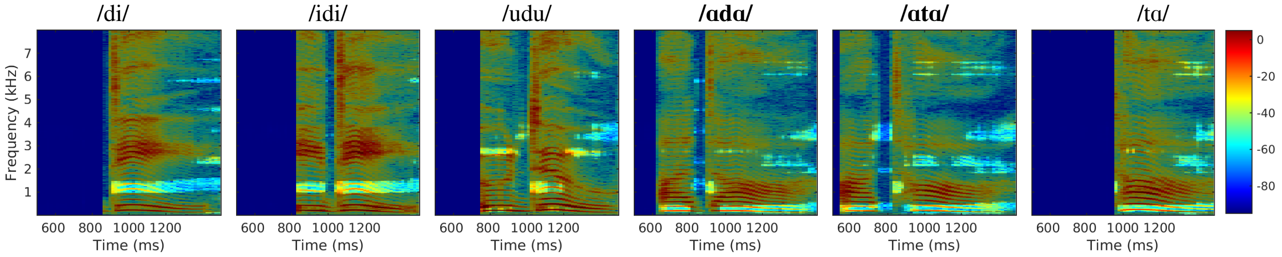

Figure 2: *Important regions for six utterances from talker W5 in the test set predicted by the BCN mask generator. Important regions are set to full lightness in the HSV color space, transitioning to half lightness for completely unimportant regions.*

the test set words /di/, /idi/, /udu/, /ɑdɑ/, /ɑtɑ/, and /tɑ/. These were selected so that they could be compared to the results for /ɑdɑ/ and /ɑtɑ/ from [9] (shown in Figure 3), and to provide an additional variety of vowels and word forms. First, focusing on /ɑdɑ/ and /ɑtɑ/, it can be seen that both human-derived and generator-predicted importance is high during and around the stop and burst of the consonant in the areas between the first two formants. Additionally, both show importance for the lowest frequencies around the fundamental, although for humans this only appears in /ɑdɑ/.

The largest difference between the human and machine importance is in the high frequencies around the stop burst. While the generator produces several smaller noise-free areas in this region, the human importance spans a large frequency range. This could be an artifact of the bubble measurement process for humans [9], because the bubbles are scaled to the ERB scale [27], which makes them "taller" at higher frequencies, reducing their resolving power. While the λ_f term in (3) attempts to create consistency in mask values at adjacent frequencies, it does not do so in a frequency-dependent way, as the ERB would. Formulating the mask prediction task in ERB frequency might make it more consistent with the human results.

Comparing the masks predicted for /idi/, /udu/, and /ɑdɑ/ in Figure 2 shows interesting differences between importance in vowels, which we did not previously investigate with humans because of the time required to perform the corresponding listening tests. It seems that for all of these words, the mask generator reveals specific regions during the vowels that differ between them, and a region at the beginning of the second syllable around 1300 Hz for all three vowels. They additionally all include some revealed regions during the stop burst. The masks for the consonant-initial words appear to be very similar to the second half of the consonant-medial counterparts. Surprisingly, the mask generator reveals a relatively large proportion of the ends of the utterances, where there is little speech energy in certain cases. It is possible that it is in fact this lack of speech energy that is informative. For example, /di/ and /idi/ show an important region between 1 and 2 kHz after 1200 ms, even though this area does not contain speech energy. Speech energy is present, however, in the same region in the words /udu/, /ɑdɑ/, /ɑtɑ/, and /tɑ/. It appears that a lack of energy helps to distinguish the vowel /i/ from /u/ and /ɑ/. Alternatively, it is possible that these trailing importance regions are caused by the time assymmetry of the uni-directional LSTM we are using. In future work, we will compare this with a bidirectional LSTM for mask prediction.

6. Conclusions and future work

In this paper, we introduce a deep neural network structure to identify the important regions of speech in noisy conditions. We show that a simple paired-input speech recognizer with a clean speech reference can produce accurate classifications of whether two utterances from different talkers contain the same word or not. Furthermore, we show that it is possible to train an LSTM neural network to identify from clean speech, large regions of the spectrogram where noise can be added without disrupting this recognition performance. These masks show patterns that are similar to analyses derived from much more expensive human listening tests [9], but the mask generator model provides a single predictor for all words and produces more general predictions of speech importance that are not dependent on the specific context of the choices offered to the listener.

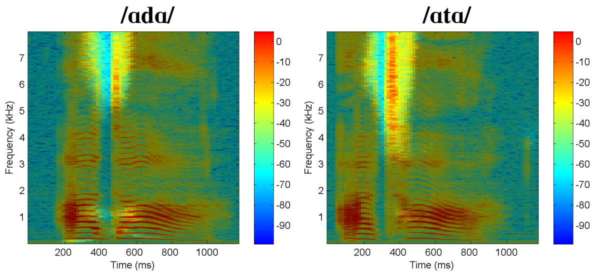

Figure 3: *Importance derived from human responses to random bubble noise (from [9]). Important regions are set to full lightness in the HSV color space, transitioning to half lightness for completely unimportant regions.*

Going forward, we will train mask generators for more sophisticated automatic speech recognizers to be able to compare more directly their performance and the cues that they use to identify specific utterances. The mask generator provides a data augmentation method that could help improve the noise robustness of these systems. Ultimately, we hope that it will be possible to use this system with human listeners to identify the cues that they use to recognize speech in noise and then to make ASR systems focus on these cues directly, hopefully improving the noise robustness of the ASR systems by doing so.

7. Acknowledgements

This material is based upon work supported by the Alfred P Sloan foundation and the National Science Foundation (NSF) under Grants IIS-1618061 and IIS-1750383. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

8. References

[1] T. Virtanen, B. Raj, and R. Singh, Eds., *Techniques for Noise Robustness in Automatic Speech Recognition*. John Wiley & Sons, Ltd, 2012.

[2] J. I. Alcántara, B. C. J. Moore, V. Kühnel, and S. Launer, "Evaluation of the noise reduction system in a commercial digital hearing aid," *Int J Audiol*, vol. 42, no. 1, pp. 34–42, 2003.

[3] E. W. Healy, S. E. Yoho, and F. Apoux, "Band-importance for sentences and words re-examined," *J. Acous. Soc. Am.*, vol. 133, no. 1, 2013.

[4] J. Hecht, "Why mobile voice quality still stinks—and how to fix it," *IEEE Spectrum*, Sep. 2014.

[5] O. Scharenborg, "Reaching over the gap: A review of efforts to link human and automatic speech recognition research," *Sp. Comm.*, vol. 49, no. 5, pp. 336–347, 2007.

[6] A. Juneja, "A comparison of automatic and human speech recognition in null grammar," *J. Acous. Soc. Am.*, vol. 131, no. 3, pp. EL256–EL261, 2012.

[7] M. P. Cooke, "A glimpsing model of speech perception in noise," *J. Acous. Soc. Am.*, vol. 119, pp. 1562–1573, 2006.

[8] I. Goodfellow, J. Pouget-Abadie, and M. Mirza, "Generative Adversarial Networks," *arXiv*, 2014. [Online]. Available: http://arxiv.org/abs/1406.2661

[9] M. I. Mandel, S. E. Yoho, and E. W. Healy, "Measuring time-frequency importance functions of speech with bubble noise," *J. Acous. Soc. Am.*, vol. 140, pp. 2542–2553, 2016.

[10] M. I. Mandel, "Directly comparing the listening strategies of humans and machines," in *Proc. Interspeech*, 2016, pp. 660–664.

[11] D. Erhan, Y. Bengio, A. Courville, and P. Vincent, "Visualizing higher-layer features of a deep network," University of Montreal Departement of computer science and operations research, Tech. Rep. 1341, 2009.

[12] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. ECCV*, 2014, pp. 818–833.

[13] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in *Proc. ICLR*, 2014.

[14] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. Müller, "How to Explain Individual Classification Decisions," *JMLR*, vol. 11, no. Jun, pp. 1803–1831, 2010.

[15] S. Bach, A. Binder, G. Montavon, F. Klauschen, K. R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS ONE*, vol. 10, no. 7, p. e0130140, 2015.

[16] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? explaining the predictions of any classifier," *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 39, no. 2011, p. 117831, 2016.

[17] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed. Boston: Springer US, 2005, ch. 12, pp. 181–197.

[18] X. Zhang, Z.-q. Wang, and D. Wang, "A speech enhancement algorithm by iterating single- and multi-microphone processing and its application to robust ASR," in *Proc. IEEE ICASSP*. IEEE, 2017, pp. 276–280.

[19] E. M. Kaya and M. Elhilali, "Modelling auditory attention," *Phil. Trans. R. Soc. B*, vol. 372, no. 1714, p. 20160101, 2017.

[20] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," in *Proc EMNLP*, 2014, pp. 1724–1734.

[21] J. Chorowski, D. Bahdanau, K. Cho, and Y. Bengio, "End-to-end Continuous Speech Recognition using Attention-based Recurrent NN: First Results," *Deep Learning and Representation Learning Workshop, NIPS 2014*, pp. 1–10, 2014.

[22] R. V. Shannon, A. Jensvold, M. Padilla, M. E. Robert, and X. Wang, "Consonant recordings for speech testing," *J. Acous. Soc. Am.*, vol. 106, no. 6, pp. L71+, 1999.

[23] B. McFee, M. McVicar, O. Nieto, S. Balke, C. Thome, D. Liang, E. Battenberg, J. Moore, R. Bittner, R. Yamamoto, D. Ellis, F.-R. Stoter, D. Repetto, S. Waloschek, C. Carr, S. Kranzler, K. Choi, P. Viktorin, J. F. Santos, A. Holovaty, W. Pimenta, H. Lee, and P. Brossier, "librosa 0.5.1," May 2017. [Online]. Available: https://doi.org/10.5281/zenodo.1022770

[24] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. AISTATS*, 2010, pp. 249–256.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE ICCV*, 2015, pp. 1026–1034.

[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[27] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, no. 1-2, pp. 103–138, 1990.