INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Identifying Important Time-frequency Locations in Continuous Speech Utterances

Hassan Salami Kavaki[1], *Michael I Mandel*[1,2]

[1]The Graduate Center, CUNY, New York, USA
[2]Brooklyn College, CUNY, New York, USA

hsalami@gradcenter.cuny.edu, mim@sci.brooklyn.cuny.edu

Abstract

Human listeners use specific cues to recognize speech and recent experiments have shown that certain time-frequency regions of individual utterances are more important to their correct identification than others. A model that could identify such cues or regions from clean speech would facilitate speech recognition and speech enhancement by focusing on those important regions. Thus, in this paper we present a model that can predict the regions of individual utterances that are important to an automatic speech recognition (ASR) "listener" by learning to add as much noise as possible to these utterances while still permitting the ASR to correctly identify them. This work utilizes a continuous speech recognizer to recognize multi-word utterances and builds upon our previous work that performed the same process for an isolated word recognizer. Our experimental results indicate that our model can apply noise to obscure 90.5% of the spectrogram while leaving recognition performance nearly unchanged.

Index Terms: Speech importance, time-frequency regions, speech recognition in noise

1. Introduction

Automatic speech recognition (ASR) systems do not recognize speech as well as human listeners in noisy environments, despite a great deal of recent progress [1]. Recent findings [1] have shown that human listeners only require a fraction of time-frequency points to correctly recognize speech signals. In addition, the randomized "bubble noise" technique allows the direct comparison of the regions used by human listeners to those used by ASR systems, and has shown that recognizers utilize very different cues from human listeners [2, 3]. It seems likely that if speech recognizers were able to utilize the same cues as humans, they would become more noise robust. As the first step towards this goal, the current paper aims to build a model that can predict the cues that are important to a listener in recognizing a particular utterance, which would greatly accelerate the process of identifying them over the current randomized approach [1].

With the help of neural networks we propose a data-driven approach to determine time-frequency points that are useful to ASR systems. Thus, the aim of this paper is the prediction of a mask which indicates important regions, useful time-frequency points for listeners, and unimportant regions of spectrogram of individual speech signals. Unlike vision, where eye tracking can provide insight into the process of attention and salience, there is no way to directly observe the hearing process and it must be investigated indirectly. Our work builds upon that of Trinh et al. [4], which introduced the bubble cooperative network (BCN), a method for predicting important speech cues from clean speech, but only applied to a very limited ASR system, one predicting whether a noisy isolated word matches a clean reference word from another talker. In this work, we replace this simple recognizer with an end-to-end recognizer capable of processing continuous speech and scalable to large vocabularies using the ESPNet framework [5]. In both cases, there is a separate importance estimation network that aims to add as much noise as possible to a given utterance without hurting recognition performance and do to so, it is best served by adding noise to unimportant regions of the utterance, i.e., those that the recognizer is not utilizing.

2. Related work

Many studies has compared ASR systems with human listeners. Lippmann [6] performed parallel experiments on humans and ASRs with several datasets, showing that the gap between the performance of non-neural-networks ASR and human listeners increased dramatically as the speech signal is corrupted by noise. More recently, Spille et al. [7] showed that to achieve the same accuracy as human listeners, ASR systems required a 12 dB higher signal to noise ratio (SNR) in spatial scenes with diffuse noise and moving talkers. Juneja et al. [8] analyzed the performance of ASR systems on noisy utterances with a null grammar to remove the effect of language modeling, finding that even without some cues that might help human listeners to perceive better, there was a fairly large gap between humans and machines. Stolcke et al. [9] compared automatic conversational speech transcriptions to human performance. The ASR system confused filled pauses like "uh," while human listeners recognized them accurately. Su et al. [10] showed that prosodically emphasized words contain more information in speech signals, i.e., that all information in speech signals is not equally useful and important for information retrieval. Krug et al. [11] introduced an introspection method to predict patterns of letters from spectrograms of speech signals. They found that the predictable patterns are interpretable. Spille et al. [12] showed that ASR and human listeners have very close speech recognition threshold (SRT) in stationary and amplitude modulated speech-shaped noise. They applied a relevant propagation analysis an found that the ASR relied upon time-frequency glimpses of high local SNR to make correct identifications.

In pursuit of a characterization of important speech frequencies, Healy et al. [13] measured the importance of frequency bands averaged over time. Importance of target band is estimated by comparing the trial when target band is present along with four other bands with the trail when target band is absent and only four other bands are present. The interaction of target band and other bands leads to the estimation of band importance. Mandel et al. [1] built upon this method to measure an importance function that varied across both frequency and time. In this method each utterance is mixed with very loud noise with randomly placed "bubbles" of silence cut out of it, through which the speech could be glimpsed. Important time-frequency

http://dx.doi.org/10.21437/Interspeech.2020-2637

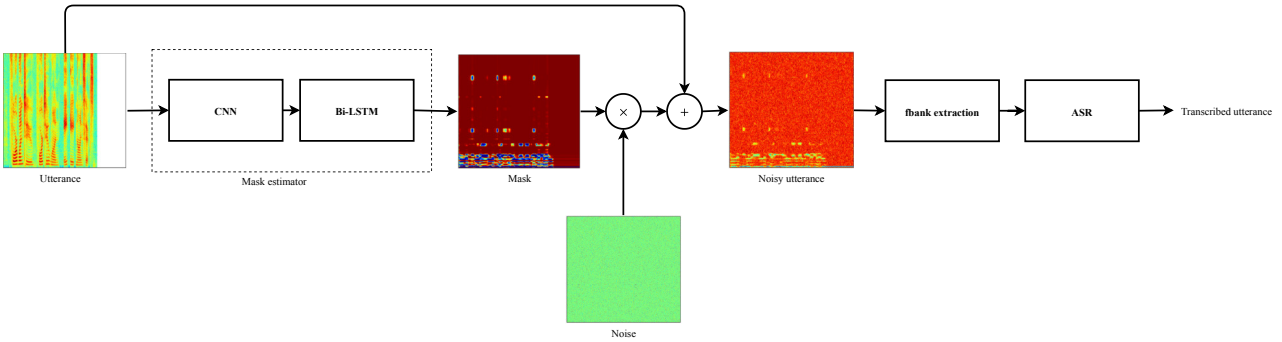


Figure 1: *Flow-chart of the proposed system, including the mask estimator that identifies important speech regions in order to add as much noise as possible.*

regions are identified as those where audibility correlates with a listener's correct identification of the word.

3. Method and model

In this section we introduce the proposed model, including the mask estimator that identifies important speech regions, along with discussing the design principles behind it. A flowchart of the entire model is shown in Fig. 1, it has two main components, the mask estimator and the ASR.

The mask estimator is a deep network that decides how much noise to add to each time-frequency point in an input spectrogram by producing a mask, which modulates the noise. Therefore, a mask value near zero adds little noise while a value near one adds maximum noise. The estimator model is a deep neural network (DNN) with parameters θ, taking the short time Fourier transform (STFT) of the clean speech, $X(f,t)$ as input, and producing the mask as:

$$M_\theta(f,t) = g\left(X(f,t);\theta\right) \tag{1}$$

We add a small amount of dither noise to each clean utterance to allow the model to generalize more effectively, but this noise is not shown in Fig. 1 or (1).

The masked noise is added to the clean speech, and the noisy speech is then fed to the ASR system. The ASR system transcribes it into a predicted word sequence,

$$\hat{\boldsymbol{y}} = h\left(X(f,t) + \alpha N(f,t) \odot M_\theta(f,t);\phi\right) \tag{2}$$

where $N(f,t)$ is a spectrogram of random white gaussian noise, α is a noise gain, ϕ is the set of parameters of the ASR network, and $\odot$ is a pointwise multiplication.

Both models are trained to minimize the combined loss:

$$\begin{aligned}\mathcal{L}(\theta,\phi) = \lambda_a L(\boldsymbol{y},\hat{\boldsymbol{y}};\theta,\phi) - \frac{\lambda_m}{TF}\sum_{t,f} M_\theta(f,t) \\ - \frac{\lambda_e}{TF}\sum_{t,f}\left(M_\theta(f,t)\log M_\theta(f,t)\right. \\ \left.+(1-M_\theta(f,t))\log(1-M_\theta(f,t))\right).\end{aligned} \tag{3}$$

The first term, $L(\boldsymbol{y},\hat{\boldsymbol{y}};\theta,\phi)$, is the ASR loss between the true word sequence, $\boldsymbol{y}$ and the predicted word sequence of the noisy speech, $\hat{\boldsymbol{y}}$. Because $h()$ is an end-to-end recognizer, this is a combination of the CTC and seq2seq loss. The other terms encourage specific mask properties. The second term encourages the mask to contain many 1's, i.e., to add as much noise as possible. The third term encourages the mask to be close to either 0 or 1, but not in-between, i.e., to have lower entropy. Each loss term has a corresponding coefficient, λ_a for the ASR loss, λ_m for the mask loss, and λ_e for the entropy loss. Our experiments explore different combinations of these coefficients to both add a large amount of noise while simultaneously preserving ASR recognition accuracy.

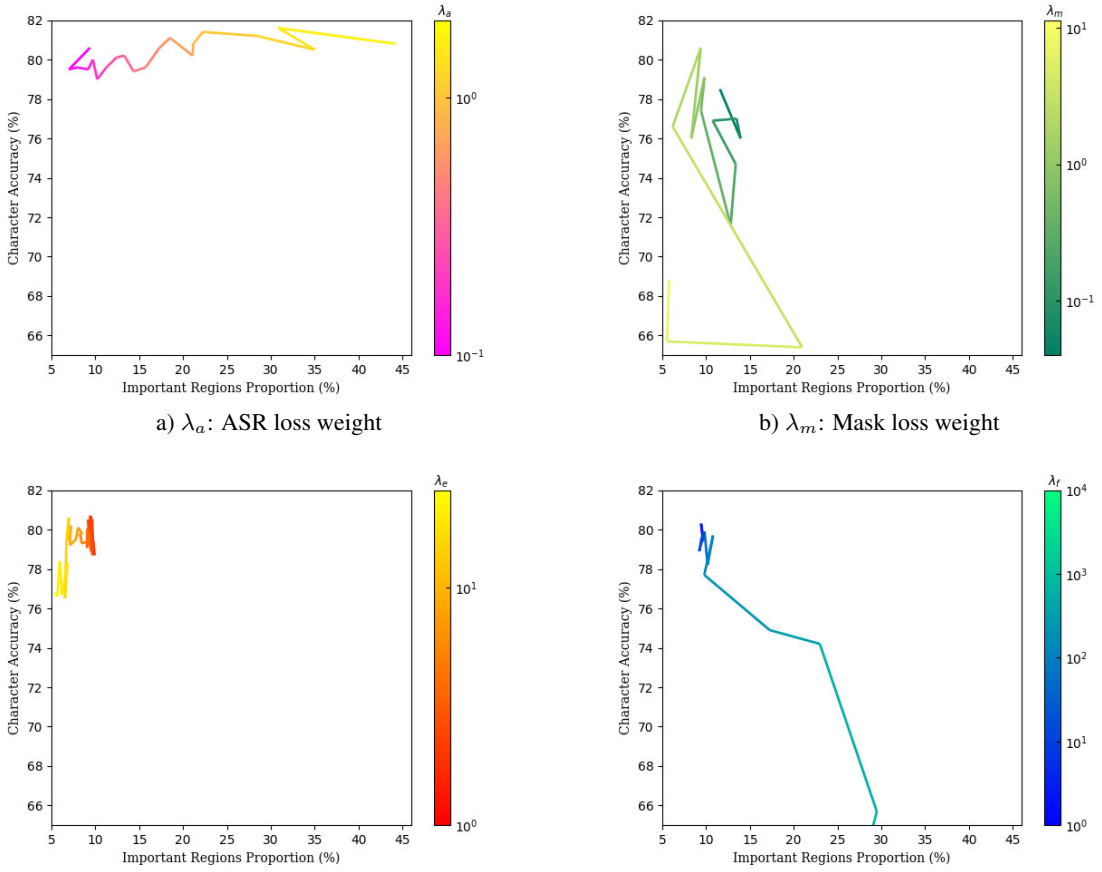
The mask estimator decides where to add noise, and thus decides which regions are important for the recognizer to observe cleanly. It is shown in the dotted rectangle in Fig. 1. It consists of three main components: a convolutional layer, a bidirectional recurrent layer, and a fully-connected layer that predicts the importance mask using the softmax function.

The last component in Fig. 1 is the ASR. There are two main approaches for end-to-end ASR. In Connectionist Temporal Classification (CTC), a recurrent neural network predicts symbols at every acoustic frame and these predictions are combined using dynamic programming. In attention-based sequence-to-sequence recognition, each symbol is predicted conditioned on the previous predicted symbols and encoded acoustic frames combined using an attention function, which implicitly performs the alignment. We use ESPnet which is a hybrid CTC/attention based ASR, providing the benefits of both models. The encoder network of the attention-based model is the CTC model. Thus this model is trained both by the forward-backward algorithm of CTC and the data-driven attention method which causes it to produce more accurate alignments by incorporating the CTC model in long sequences [14]. We used the ESPnet toolkit to implement our ASR model. ESPnet uses chainer and pytorch as a main deep learning engines [5]. In addition, we implemented "fbank" feature extraction in pytorch to allow gradients to be propagated through to our mask estimator model.

4. Experiments

4.1. Dataset

Our experiments are performed on the AN4 alphanumeric dataset [15], also known as the CMU census dataset. Each utterance is in the form of a phone number, birth date, spelled out address, etc. In addition, the speakers also produced random sequences of control words. All recordings were made with a close talking microphone. All data are sampled at 16 kHz, 16-bit linear sampling. The dataset contains 1078 utterances, of which we used 848 for training, 100 for validation, and 130 for test.



a) λ_a: ASR loss weight

b) λ_m: Mask loss weight

c) λ_e: Mask entropy weight

d) λ_f: Mask smoothness in frequency

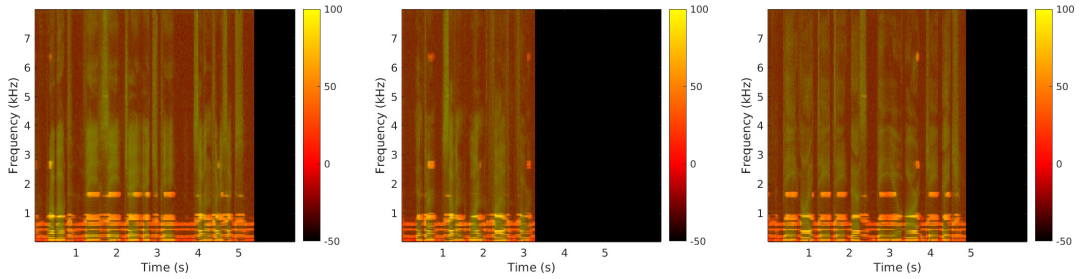
Figure 2: *Effect of changing λ weights on character accuracy of noisy speech and the proportion of spectrogram points where noise is added with a mask greater than 0.5. When varying one λ, the others were held at $\lambda_a = 0.1$, $\lambda_e = 2$, $\lambda_f = 0$, $\lambda_m = 10$.*

4.2. Training the model

We trained the ASR and mask estimator models separately. First we trained the ASR on the clean speech utterances. We used 80-dimensional fbank features with a frame size of 25 ms and a hop size of 10 ms. The longest utterance is 6.4 s, so we zero-pad the shorter utterance to that length. After training the ASR, we freeze its weights and train the mask estimator. The mask estimator uses an STFT with a 1024-point FFT (64 ms) as input. Hyperparameters were selected to balance high character accuracy with a low proportion of mask values below 0.5, indicating a concentrated prediction of importance on the validation set. This balance is best ahcieved by $\lambda_a = 0.1$, $\lambda_e = 2$, $\lambda_m = 10$ (see Fig. 2). Before applying the spectrogram features to the convolutional layer we normalize them to have zero mean and unit variance. We also perform batch normalization between layers [16]. The convolutional layer of the mask estimator has 256 kernels with size 11×32 and stride 1 in time and 16 in frequency. The bidirectional LSTM has 512 hidden units. The noise gain is 50000 in comparison to 16 bit integer speech waveforms (± 32768), which was chosen to make sure the noise is powerful enough to make the clean speech signal completely inaudible. The dither gain is 6.25. We initialized all the wights of the layers with a uniform distribution [17]. We trained the mask estimator model with mini-batch stochastic gradient descent with batch size 30, and we use AdaDelta [18] with learning rate 10^{-6} and running avareage 0.95. The hyperparameters of the ASR model are summarized in Table 1.

Table 1: *Hyperparameters of ASR*

Hyperparameter	**Value**
# of encoder layers	4
# of encoder BLSTM cells	300
# of encoder projection units	320
# of attention transformation dimensions	320
# of heads for multi head attention	4
# of attention convolution filters	100
# of decoder layers	1
# of decoder LSTM cells	320
Multitask learning coefficient	0.5
Optimization	AdaDelta
AdaDelta ϵ	10^{-8}
AdaDelta ϵ decaying factor	10^{-2}
Gradient norm clip threshold	5
Maximum epoch	30
Threshold to stop iteration	10^{-4}



Rubout V-Z-J-H-P seven thirty six | Enter seven one five four | P-I-T-T-S-B-U-R-G-H

Figure 3: *Important regions of three utterances. Important regions are set to full lightness in the HSV color space*

5. Results

As shown in Tab. 2, on clean speech, the ASR achieves a character error rate (CER) of 11.1% on the evaluation set. In addition, we added noise to the entire spectrogram of each signal to make sure the noise is powerful enough to contaminate the clean speech signal thoroughly. This full-noise condition led to a CER of 123.7%, confirming that the noise is disruptive. With the best setting of loss weights, the mask estimator is able to add noise to 90.5% of the spectrogram, while still achieving a CER of 13.9%, very close to that on clean speech. This indicates that the mask estimator is very effective at adding noise to unimportant regions of the spectrogram and is therefore very effective at identifying important regions.

It is important to point out that there is a trade off between the proportion of masked spectrogram points and the character accuracy of the resulting signal. Fig. 2 shows several experiments varying the λ coefficients to explore this tradeoff. The result of changing the ASR loss coefficient, λ_a, is shown in Fig. 2(a), increasing it from 0.1 to 2.0 equally spaced on a log scale. As can be seen, increasing λ_a increases the character accuracy and decreases the proportion of masked spectrogram points. This is as expected. Fig. 2(b) shows the effect of increasing λ_m from 0.04 to 11.37 equally space on log scale. It shows that increasing λ_m decreases the character accuracy and increases the proportion of masked spectrogram point. Fig. 2(c) shows the result of changing the entropy loss coefficient, λ_e, from 1.0 to 25.71 equally spaced on a log scale. As the entropy loss coefficient increases, the character accuracy decreases and the proportion of masked spectrogram points increases. In contrast to the λ_a curve, the λ_e curve shows a discontinuity in character accuracy from 80% to 77% at a revealed proportion of 7%.

Fig 3 shows the spectrogram of the clean speech signal with the predicted mask overlayed. Mask values that are low (meaning less noise is added and they are presumably important) are shown in full lightness, while mask values that are high (meaning more noise is added and they are presumably not important) are shown at half lightness. As can be seen, a great majority of the speech signal has a high mask value and is corrupted by loud noise. the generated masks indicate that most of the important regions are in low frequency regions below 1000 Hz, but with a secondary band around 1700 Hz. These could be detecting formants, but it is surprising that there are no important regions predicted above 1700 Hz for fricatives or plosives. This is in contrast to human importance maps measured using randomized bubble noise [1–3].

Table 2: *ASR results of the final system showing the CER and corresponding proportion of spectrogram points with a noise mask greater than 0.5 in comparison with two baselines.*

Noise	**CER**		**Noise proportion**	
	valid	eval	valid	eval
ASR with clean speech	17.6	11.1	0.0	0.0
ASR with mask estimator	19.7	13.9	90.6	90.5
ASR with noisy signal	131.9	123.7	100.0	100.0

As an additional experiment, we added a term to the loss function, equation (3), to penalize abrupt discontinuities in mask values across frequency. This term causes the important regions to become more smooth in the frequency dimension:

$$\mathcal{L}_f(\theta) = \frac{\lambda_f}{TF} \sum_{t,f} |\Delta_f M_\theta(f,t)| \tag{4}$$

where Δ_f is a first order difference across frequency. The result of this experiment is shown in Fig. 2(d) for the values of λ_f from 1 to 10000 equally spaced on a log scale. This plot shows that increasing λ_f decreases character accuracy and decreases the proportion of masked spectrogram points. Thus it does not help to add this term, and looking at the masks shown in Fig. 3, they appear to be relatively smooth in frequency already.

6. Conclusion and future work

In this paper we showed that our previously introduced Bubble Cooperative Network model can be successfully applied to a continuous speech recognition system. The model estimates the important regions of a speech utterance by predicting where noise can be added while still allowing the ASR to transcribe the new noisy speech signal approximately as well as the original clean signal. Parameter sweeps found settings of weights for the loss function that balance the two competing objectives of high character accuracy and a high amount of added noise.

Future work will utilize this predicted mask for data augmentation to improve the accuracy of the recognizer in noise, and will also explore ways in which this model can serve to initialize a model of importance for human listeners.

7. Acknowledgements

This material is based upon work supported by the National Science Foundation (NSF) under Grant 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] M. I. Mandel, S. E. Yoho, and E. W. Healy, "Measuring time-frequency importance functions of speech with bubble noise," *The Journal of the Acoustical Society of America*, vol. 140, no. 4, pp. 2542–2553, 2016.

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

[3] V. A. Trinh and M. I. Mandel, "Directly comparing the listening strategies of humans and machines," *IEEE Tr. Aud., Spch., & Lang. Proc.*, 2020, under review.

[4] V. A. Trinh, B. McFee, and M. I. Mandel, "Bubble cooperative networks for identifying important speech cues." in *Interspeech*, 2018, pp. 1616–1620.

[5] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "Espnet: End-to-end speech processing toolkit," in *Interspeech*, 2018, pp. 2207–2211.

[6] R. P. Lippmann, "Speech recognition by machines and humans," *Speech communication*, vol. 22, no. 1, pp. 1–15, 1997.

[7] C. Spille, B. Kollmeier, and B. T. Meyer, "Comparing human and automatic speech recognition in simple and complex acoustic scenes," *Computer Speech & Language*, vol. 52, pp. 123–140, 2018.

[8] A. Juneja, "A comparison of automatic and human speech recognition in null grammar," *The Journal of the Acoustical Society of America*, vol. 131, no. 3, pp. EL256–EL261, 2012.

[9] A. Stolcke and J. Droppo, "Comparing human and machine errors in conversational speech transcription," *arXiv preprint arXiv:1708.08615*, 2017.

[10] C.-Y. Su and C.-Y. Tseng, "How prosodic cues could lead to information center in speech-an alternative to asr," in *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*. IEEE, 2017, pp. 1–6.

[11] A. Krug and S. Stober, "Introspection for convolutional automatic speech recognition," in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 2018, pp. 187–199.

[12] C. Spille and B. T. Meyer, "Listening in the dips: Comparing relevant features for speech recognition in humans and machines." in *INTERSPEECH*, 2017, pp. 2968–2972.

[13] E. W. Healy, S. E. Yoho, and F. Apoux, "Band importance for sentences and words reexamined," *The Journal of the Acoustical Society of America*, vol. 133, no. 1, pp. 463–473, 2013.

[14] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid ctc/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[15] A. Acero, *Acoustical and environmental robustness in automatic speech recognition*. Kluwer Academic Publishers, 1993. [Online]. Available: http://www.speech.cs.cmu.edu/databases/an4

[16] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[17] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010, pp. 249–256.

[18] M. D. Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.