Fully Convolutional Network (FCN) Model to Extract Clear Speech Signals on Non-stationary Noises of Human Conversations for Cochlear Implants

Tsai, Yi-Ting, Liao, Lauren Diana

Tsai, Yi-Ting
Department of Computer Science
The University of Hong Kong
Hong Kong, China
u3512452@connect.hku.hk

Liao, Lauren Diana
Department of Mathematics
University of California, San Diego
La Jolla, California, USA
ldliao@ucsd.edu

Abstract—Cochlear implant (CI) electronically stimulates the nerve to help those with severe hearing lost. However, under noisy backgrounds, speech perception tasks have remained difficult for CI users. Therefore, speech enhancement (SE) is a critical component to improve speech perception examining through different noise scenarios. In this study, we developed the fully convolutional network (FCN) model to extract clear speech signals on non-stationary noises of human conversations in the background, and further compare the model's performance with previously developed log power spectrum (LPS) based Deep neural network (DNN) model's performance by conducting hearing test of enhanced speech which simulated in CI.

Keywords—speech enhancement; convolutional neural network; cochlear implant; denoising autoencoder

I. INTRODUCTION

Although with the advance in heading aids, Cochlear implant (CI) users continues to suffer from low perception of speech under noisy scenario. Limited studies had been done focusing on speech enhancement (SE) specifically on CI because of the processed sound signals contain a certain amount of distortions. In the past, SE has been conducted in traditional minimum-mean-square-error (MMSE) -based spectral amplitude estimator [1]. More recently, deep denoising autoencoder (DDAE), DNN-based and Convolutional neural network (CNN) model SE models are proposed to process SE. [2-3] However, most of these models are based on the magnitude spectrogram (i.e. Log power spectrum (LPS)), resulting in noisy results as the main phase component in speech is kept under its original noisy form. The problem is based on possible lack of structure for phase spectrogram. Therefore, difficult for the deep learning model to predict the clean format of the noisy speech. To address this problem, Fu et al. [4] has proposed a raw waveform-based SE model to keep the feature of phase and reduce the weights parameters. However, under the characteristic of raw-waveform, a feature point alone in the time domain does not have useful information; so, it must focus on the neighbor information as well. Therefore, adopting only the CNN model is not effective. The fully connected layer of CNN model makes learning the more difficult to generate high and low frequency parts of a waveform simultaneously. [4] As a result, we discount for the last layer of fully connected layer and adopt the fully convolutional network (FCN) model.

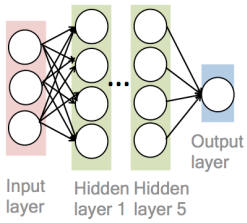
In this study, in order to observe whether the model can extract clean speech from speech with noisy background (ie. noisy speech) for CI and return a better SE performance result than LPS based Deep neural network (DNN), we analyze the quantitative evaluation of speech quality (PESQ) [5] and the short-time objective intelligibility (STOI) score [6] to evaluate the speech quality and intelligibility. We utilize the Voice Operated recorder (Vocoder) to simulate the enhanced speech into the sound CI patients heard, which mimics the electronic sounds that deconstruct the utterances. Thus, we can conduct hearing test with normal hearing subjects. The more characters in several mandarin utterances of the enhanced speech are correctly answered, the more efficient the performance shows. Experimental results show both successful quantitative and hearing test results. The FCN model can not only reduce the number of parameters but also extract relatively cleaner speech even after voice processing into vocoded speech than DNN.

II. SPEECH ENHANCEMENT MODEL

A. LPS based DNN

The model in figure 1 is a tradtional neural model with 5 hidden fully connected layers. Each fully connected layer has 2048 neurons with input size is 7 frames. The Overlap sets to 256 points. The DNN has a characteristics of large amount of parameters which it is likely to be reduced.

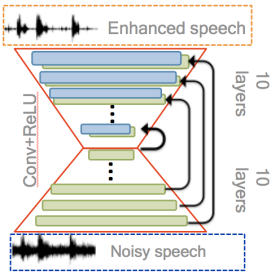
FIGURE 1. Structure of DNN model



B. *Raw waveform FCN*

In order to preserve the important phase components in speech, inputting raw waveform speech is chosen. FCN model is an improved version of Convolutional Neural network (CNN). CNN normally consists of convolutional layers and max-pooling layers. Convolutional layer serves to apply filters and extract features while Max-pooling layers make the output of convolution networks translational invariant. [7] In speech enhancement, the information and structure of the speech is very important for understanding the meaning, however the characteristics of pooling suffers the details of textures and structures. [3] As a result, we do not apply pooling layer. The difference between FCN and CNN is that the fully connected layer in the CNN model has been given rid of in the FCN model. Since the characteristics of raw waveform, the hidden fully connected layer is hard to learn the weights. [4] The FCN model in figure 2 we proposed is Encoder has 10 convolutional layers and Decoder also has 10 convolutional layers with 16384 points in a frame. Overlap sets to 256 points.

FIGURE 2. Structure of encoder-decoder FCN model



III. VOCODED SPEECH

Vocoder is a voice processing system that can analyze and resynthesize human voice signals. In past decades, vocoder has had a profound impact on the development of CI research; that is, vocoder-based speech simulations have been widely adopted to predict the general pattern of speech recognition performance for CI users [8] The speech processed after vocoder is view as the sound heard by CI patients. The advantage of using vocoder is to avoid the difficulty of conducting hearing test on real CI patients. There are two main types of vocoder implementations for CI experiments—namely, tone vocoder and noise vocoder [9] Many studies use noise vocoder as simulating CI speech. [10]

IV. HEARING TEST

Hearing test interface is implemented on Matlab. Hearing test interface includes check box for tester to tick the wrong answers from subjects and calculates the correctness directly after subjects finish test. By playing vocoded enhanced speech from different models, the correctness of answers from subject evaluates the performance of denoising. Before hearing test, tester should play a certain amount of vocoded clean speech to subjects since the subjects are under assumption that they are not familiar with vocoded speech.

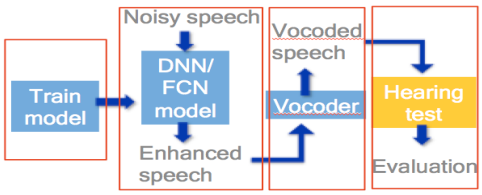
V. EXPERIMENTS

A. *Experiment setup*

In our experiments, the Mandarin version of Hearing in Noise Test (MHINT) corpus [11] was used to prepare the testing datasets and training datasets. The prepared dataset contains 200 utterances artificially mixed with 9 different noise types (boy1_30s, boy_ankh, buccaneer1, cafeteria_babble, car_noise_idle_noise_60mph, crowd-party-adult-med, girl-3, girl-4, pinknoise) at 5 different SNRs (-10 dB, -5 dB, 0 dB, 5 dB, 10 dB), resulting in a total of (200 x 9 x 5) utterances. To form the testing set, it adopted 2 types of interference noise sources. These 2 noises are not seen in the training set, including 2 girl talkers and baby cry. In total there are 120 clean utterances mixed with the 2 noise types at -3dB and -6 dB SNRs.

LPS-based DNN model and raw waveform FCN model are chosen for comparison. The evaluation method is by both quantitative result and hearing test result. Figure 3 is the procedure of experiment. First, train model and input the testing data in model, utilize the generated enhanced speech from model into vocoder. After processing enhanced speech as vocoded speech, conduct the hearing test to test the intelligibility. In the hearing test, tester plays 5 utterances randomly from the vocoded speech of DNN, FCN model as well as the vocoded noisy speech. Each utterance contains 10 characters and therefore the full score for each model at each SNR is 5x10=50. 8 subjects who are native mandarin speakers with normal hearing and not familiar with the testing data utterances have conducted the hearing test.

FIGURE 3. EXPERIMENT PROCEDURE



B. *Experimental Results*

TABLE I. STOI AND PESQ RESULTS

	DNN (LPS)		FCN (Waveform)	
SNR(dB)	STOI	PESQ	STOI	PESQ
-3 dB	0.6169	1.1604	0.6870	1.3020
-6 dB	0.5781	1.1301	0.6426	1.2401
Average	0.5975	1.1453	**0.6648**	**1.2710**

Table II. Noisy background Results

	Subjects			
	Subject 1	Subject 2	Subject 3	Subject 4
2 girls - 3dB	27	8	0	4
2 girls - 6dB	14	3	10	3
4 talkers - 3dB	4	10	27	1
4 talkers - 6dB	10	20	5	10
Average	13.5	10.25	10.5	4.5
	Subject 5	Subject 6	Subject 7	Subject 8
2 girls - 3dB	2	15	24	16
2 girls - 6dB	12	13	15	3
4 talkers - 3dB	0	19	0	26
4 talkers - 6dB	0	15	0	15
Average	3.5	15.5	9.75	15

TABLE III. DNN DENOISED RESULTS

	Subjects			
	Subject 1	Subject 2	Subject 3	Subject 4
2 girls - 3dB	23	26	29	29
2 girls - 6dB	21	30	24	12
4 talkers - 3dB	11	24	11	7
4 talkers - 6dB	10	21	11	13
Average	16.25	25.25	18.75	15.25
	Subject 5	Subject 6	Subject 7	Subject 8
2 girls - 3dB	18	9	10	45
2 girls - 6dB	9	29	20	11
4 talkers - 3dB	0	24	7	31
4 talkers - 6dB	18	9	10	6
Average	11.25	17.75	11.75	23.25

TABLE IV. FCN DENOISED RESULTS

	Subjects			
	Subject 1	Subject 2	Subject 3	Subject 4
2 girls - 3dB	34	31	20	21
2 girls - 6dB	26	21	33	14
4 talkers - 3dB	26	46	15	24
4 talkers - 6dB	17	37	2	17
Average	25.75	33.75	17.5	19
	Subject 5	Subject 6	Subject 7	Subject 8
2 girls - 3dB	25	23	29	35
2 girls - 6dB	18	5	8	9
4 talkers - 3dB	12	23	25	43
4 talkers - 6dB	10	18	21	37
Average	16.25	17.25	20.75	31

TABLE V. MEDIAN

	Median		
	Noisy	*DNN Denoised*	*FCN Denoised*
2 girls -3dB	11.5	24.5	27
2 girls -6dB	11	20.5	16
4 talkers -3dB	7	11	24.5
4 talkers -6dB	10	10.5	17.5

VI. STATISTICAL ANALYSIS

Note for statistical analysis of this preliminary data, we use the Wilcoxon signed-rank test. Because we have a small sample size of 8, we cannot assume the distribution of the means is normal. However, we are assuming each subject is independent from one another that one answers independently from the influence of others and the distribution from the null and alternative hypothesis are symmetric. Since we want to discount for individuals who are on the extremes, we use median for a robust analysis. Note that robust means the few extreme values will not affect the data drastically and skew the statistic.

We use the difference between the correct characters from FCN model minus DNN model for each subject. The hypotheses above are two sided, we are using significance level at alpha (type I error) as 0.05 to find the critical value.

We consider individual differences by denoting the noisy data as base. We subtract number of correctly answered characters for noisy data from the correct characters for FCN or DNN model. This provides a fair comparison for the results between the models. While focusing on the median of each data result, calculating the average on each subject indicates the individual differences and comparable across the three models.

VII. Results and Discussion

Notice raw waveform based FCN scores higher than LPS-based DNN with 0.6648 in STOI and 1.2710 in PESQ (Table I). Under the same SNRs of the testing dataset, FCN model shows better speech quality and intelligibility under computational calculation than the previous DNN model.

Considering the experimental results, for the noise type of two girls talking at -3 dB and -6 dB, FCN and DNN models produced similar results as well as for four talkers at -6 dB. However, for 4 talkers at -3 dB, the FCN model is statistically significantly better than the DNN model (Table V).

Specifically, for two girls talking in the background at -3 dB, the W-statistic is 12. The critical value of W for N = 8 at $p \leq 0.05$ is 3. Therefore, the result is not significant at $p \leq 0.05$. Similarly for -6 dB, the W-statistic is 14.5. Therefore, the result is not significant at $p \leq 0.05$. For four talkers in the background, at -6 dB, the W-statistic is 7.5, which is also not statistically significant. However, for four talkers in the background, -3 dB has the W-statistic at 1. The critical value of W for N = 8 at $p \leq 0.05$ is 5. Therefore, the result is significant at $p \leq 0.05$.

The difference in median for two talkers noise using result from FCN minus DNN models at -3 dB is 2.5 characters and at -6 dB is -4.5 characters. For four talkers noise, the difference in median at -3dB is 13.5 characters and at -6 dB is 7.5 characters.

For human talking as background noise, FCN model shows an improvement particularly in four talkers in the background at -3dB. In addition, when we examine the full analysis, the data indicates there is no significant difference in the model. Since the FCN model requires less parameters to operate and indicates either no significant difference or improvement in comparison to the DNN model, we can continue to build on FCN model as an indicator for an improved speech denoising processor for CI users.

Noting the individual differences (Table II-IV), on average, the FCN model increased subject responses with two individuals indicating DNN model had better performance for them in the increase of about 1 character, which is not a clear indicator of the better performance of DNN model for those particular subjects.

Prior to this trial, we also considered other non stationary noises such as the sound of baby crying in the background, from that experiment, both models performed significantly worse than noisy with all data as negative values (correct character count for model – noisy data). This indicates that the sound of baby crying with the high frequency changes is much easier to interpret the correct sentence in the background and denoising such noise, is ineffective. However, with the models for two girls talking and four girls talking, this model is effective to extract the sentence. The non stationary noise with relatively close range of frequency as oppose to baby crying at sharper, higher tones, is much better suited for this denoising model.

To further this trial, we want to continue to gain a larger sample size, testing more subjects, to have a larger data collection to calculate statistical significance. We can also verify for other non-stationary noises that CI users may encounter to identify and optimize the function of the FCN model.

VIII. Conclusion

FCN has higher PESQ AND STOI score than LPS based DNN in terms of standardized quantitative evaluation. Moreover, the numbers of parameters in network is dramatically reduced in FCN model and this could benefit in both faster calculation and less hard-drive spaced needed for CI users. Due to good quantitative result, if the hearing test results show similarity, the FCN model still concludes as a better model. Denoising human talking, the non-stationary background noise, FCN model shows an improvement and in this case could be considered to outperform DNN. Therefore, for future prospect, we can optimize the FCN and conduct further testing with more subjects.

Acknowledgment

We would like to express our gratitude to the support by our advisor Dr. Yu Tsao. This project would not be possible without his advice. We would like to thank to Mr. Szu Wei Fu and Sian Fong Liao for his help on implementation the models.

References

[1] Y . Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 32, pp. 1109-1121, 1984.

[2] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in INTERSPEECH, pp. 436-440.I, 2013.

[3] Fu, Szu-Wei, Yu Tsao, and Xugang Lu. "SNR-Aware Convolutional Neural Network Modeling for Speech Enhancement." In *INTERSPEECH*, pp. 3768-3772. 2016.

[4] Fu, Szu-Wei, Yu Tsao, Xugang Lu, and Hisashi Kawai. "Raw Waveform-based Speech Enhancement by Fully Convolutional Networks." *arXiv preprint arXiv:1703.02205*, 2017.

[5] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," ITU-T Recommendation, p. 862, 2001.

[6] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, pp. 2125-2136, 2011.

[7] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, and G. Penn, "Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition," in International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4277-4280, 2012.

[8] F. Chen, and A. H. Lau, "Effect of vocoder type to Mandarin speech recognition in cochlear implant simulation," in Proc. ISCSLP, pp. 551-554,2014.

[9] N. A. Whitmal, S. F. Poissant, R. L. Freyman, and K. S. Helfer, "Speech intelligibility in cochlear implant simulations: Effects of carrier type, interfering noise, and subject experience," The Journal of the Acoustical Society of America, vol. 122, no. 4, pp. 2376-2388, 2007.

[10] Lai, Ying-Hui, Fei Chen, Syu-Siang Wang, Xugang Lu, Yu Tsao, and Chin-Hui Lee. "A deep denoising autoencoder approach to improving the intelligibility of vocoded speech in cochlear implant simulation." IEEE Transactions on Biomedical Engineering, 2016.

[11] L. L. Wong, S. D. Soli, S. Liu N. Han, and M.-W. Huang, "Development of the Mandarin hearing in noise test (MHINT),"Ear and hearing, vol. 28(2), pp 70S-74S, 2007.