Fully convolutional network (FCN) model to extract clear speech signals on non-stationary noises of human conversations for cochlear implants

Tsai, Yi-Ting, University of Hong Kong ;
Lauren Diana Liao, University of California, San Diego

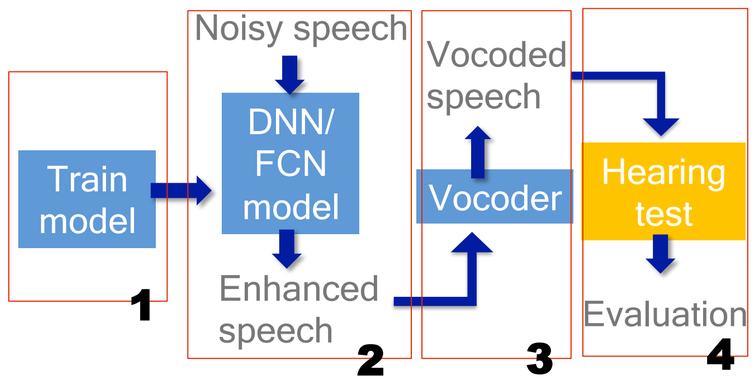
INTRODUCTION

Cochlear implant (CI) electronically stimulates the nerve to help those with severe hearing lost. Under noisy backgrounds, however, 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 makes use of deep learning and focus on 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 performance with Log power spectrum (LPS) based Deep neural network (DNN) model by conducting hearing test of enhanced speech which simulated in CI.

OBJECTIVES

Purpose a raw waveform based FCN model for denoising speech with human conversations noises, which has the same or better performance as LPS based DNN on quantitative result and hearing test result.

PROCEDURE

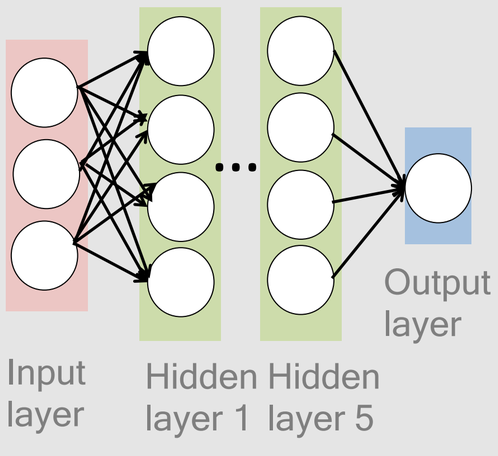


1. Given the data set and train FCN/DNN model
2. Generate testing data set
3. Transfer enhanced speech result into vocoded speech
4. Conduct hearing test and evaluate

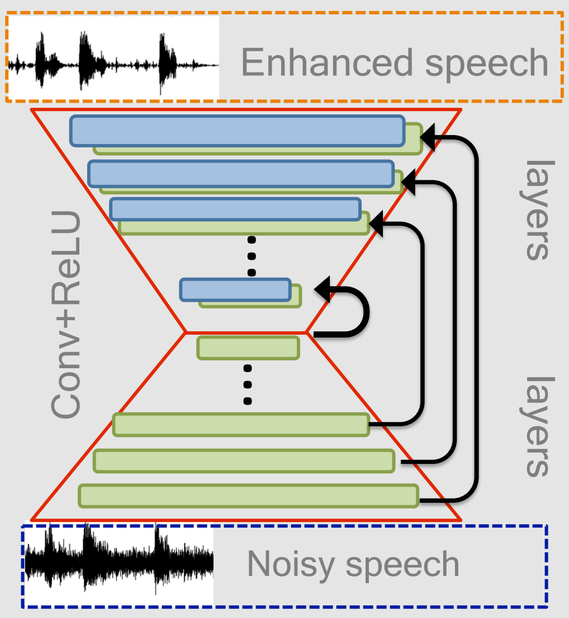
METHODS

A. Models:

Log power spectrum based **DNN**. A fully connected network with 5 hidden layers.



Raw waveform based **FCN**
Encoder-decoder architecture, Encoder and decoder have 10 convolutional layers each.



B. Training dataset:

200 utterances artificially mixed with 9 different noise types at 5 different signal to noise ratio (SNRs) (-10 dB, -5 dB, 0 dB, 5 dB, 10 dB), resulting in a total of (200 x 9 x 5) utterances.

C. Testing dataset:

2 noises are not seen in the training set- 2 girls talkers and 4 mix gender talkers. In total there are 120 clean utterances mixed with the 2 noise types at -3dB and -6 dB SNRs.

D. Vocoded speech:

Vocoder *(Voice Operated reCOrDER)* is a voice processing system that can simulate the voice heard by CI users. The testing dataset results were vocoded and used in the hearing test.

E. Hearing test:

Subject: 8 native mandarin speakers with normal hearing and not familiar with the utterances.
Utterance: 120 utterances, 3-4 seconds each sentence, spoken by native mandarin speakers.
Randomly pick each 10 utterances from DNN or FCN and original noisy speech playing to the subject and analyze the correctness.

RESULTS

A. Quantitative result

To evaluate the performance of the proposed models, the perceptual evaluation of speech quality (PESQ) and the short-time objective intelligibility (STOI) scores were used to evaluate the speech quality and intelligibility, respectively.

	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**

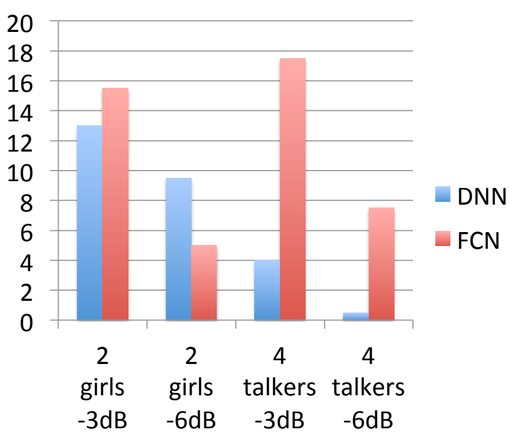
Since the fully connected layers were removed, the number of weights involved in FCN was approximately only 10% when compared to that involved in DNN.

B. Hearing test result

The figure in the table is the median correct characters they have answered and heard during the hearing test. The full score for each noise type at different SNR is 50 characters, since we randomly chose 5 Mandarin utterances for each noise type and each utterance includes 10 characters. The bar chart is derived from denoting noisy speech as base. This provides a fair comparison for the results between the models.

	i. Noisy speech	ii. Enhanced speech by DNN	iii. Enhanced speech by FCN
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

Number of correctly answered characters
(Speech from DNN/FCN – Noisy)



CONCLUSIONS

- 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.
- 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.

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.

REFERENCES

Y.-H. Lai, F. Chen, S.-S. Wang, X. Lu, Y. Tsao, and C.-H.Lee, "A deep denoising autoencoder approach to improving the intelligibility of vocoded speech in cochlear implant simulation,",2016
S.P, A.B, J.S, "SEGAN: Speech Enhancement generative adversarial network"
S-W Fu, Y. Tsao, X. L, H. K, "Raw waveform-based speech enhancement by fully convolutional networks", 2017.

THE UNIVERSITY OF HONG KONG
DEPARTMENT OF COMPUTER SCIENCE
UC San Diego Mathematics