Paper: 2087

A Causal Deep Learning Framework for Classifying Phonemes In Cochlear Implants

Kevin M. Chu, Leslie M. Collins, Boyla O. Mainsah

Applied Machine Learning Laboratory, Department of Electrical and Computer Engineering, Duke University, Durham, NC

Introduction

- Cochlear implants (CIs) (Fig. 1) aim to restore speech perception to individuals with sensorineural hearing loss
- CI users have difficulty understanding speech in listening environments that contain reverberation and noise [1]
- CI users are more detrimentally affected than normal hearing listeners because the speech signal presented to a CI user has limited spectral resolution

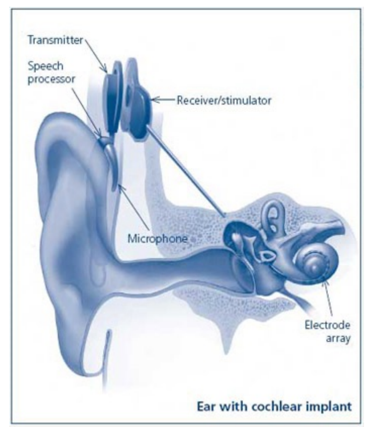

Fig. 1: Diagram of cochlear implant. Image source: Medical illustrations by NIH, Medical Arts and Photography Branch [2]

Time-Frequency Masking

- Speech enhancement technique where the time-frequency (T-F) representation of speech is multiplied by a matrix of gain values to suppress reverberation and noise [3] (Fig. 2)

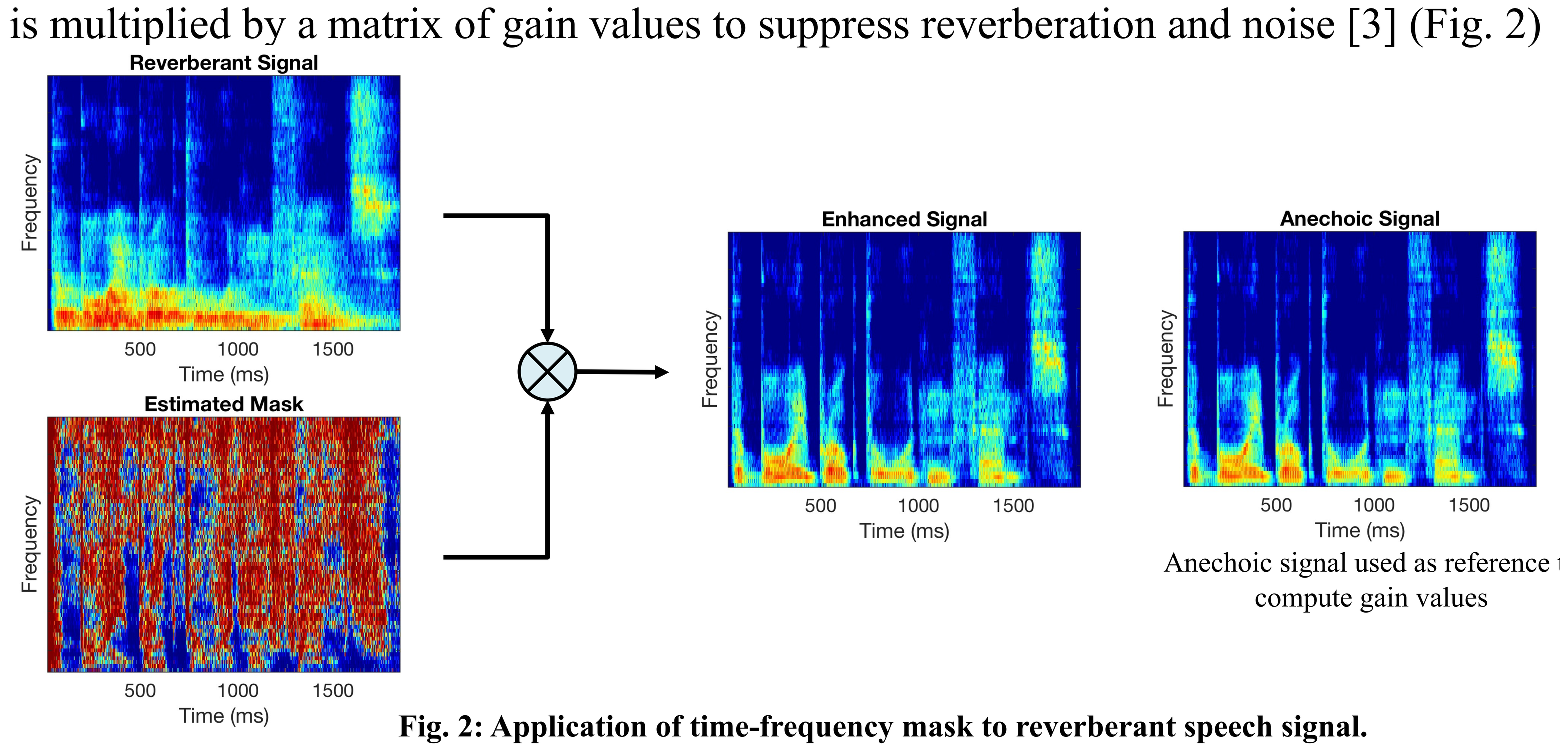

Fig. 2: Application of time-frequency mask to reverberant speech signal.

- In real-time, an algorithm must be developed to estimate mask from reverberant signal (Fig. 3)
- T-F mask estimation algorithms have limited ability to remove reverberation in low frequencies, where overall level of reverberation is higher [4]

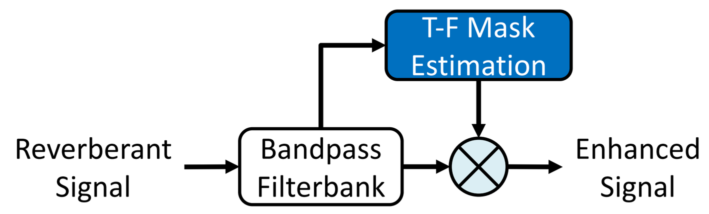

Fig. 3: Conventional phoneme-independent mask estimation

Phoneme-Based Mask Estimation

- Leverage spectro-temporal structure using phoneme-based masks, as phonemes are concentrated in specific frequency ranges (Fig. 4)
- Phoneme-based masks have improved the performance of ASR models [5], so potential benefit for CI users
- In ideal case where phoneme is known, phoneme-specific masks improve vocoded speech intelligibility compared to conventional, phoneme-independent masks (Fig. 5)

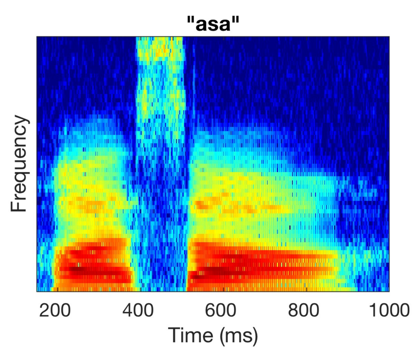

Fig. 4: Spectrogram of "asa". "a" activates low frequencies, while "s" activates high frequencies.

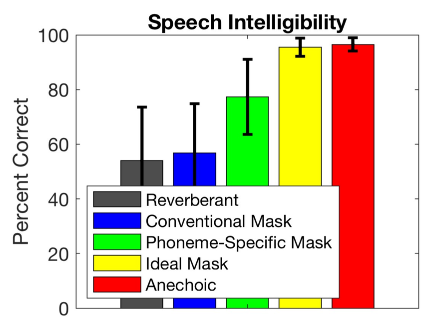

Fig. 5: Intelligibility of vocoded speech under different processing conditions. Results show mean and standard deviation in percent of correct phonemes for normal hearing listeners.

- **Goal**: develop a phoneme classification model (Fig. 6) that can categorize phonemes within the constraints of a CI:
  - Framewise
  - Causal
  - Same time-frequency resolution as a CI
  - Low parametric complexity

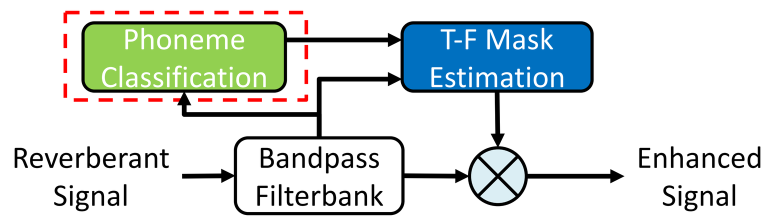

Fig. 6: Phoneme-based mask estimation framework

Classification Tasks

- Phoneme and manner of articulation (MOA) classification
- MOA describes how articulators influence airflow through vocal tract
- Phoneme classification is challenging due to confusions within same MOA (Fig. 7), which leads to confusions in classification [6]
- MOA also conveys spectral information, so potential benefit for speech enhancement algorithms with less complexity

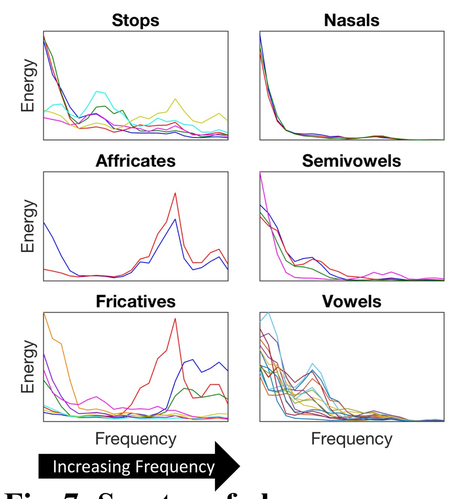

Fig. 7: Spectra of phonemes. Phonemes within the same MOA have similar spectra.

Features

- **ASR features**
  - Extracted over 25ms frames with 10ms frame shift
  - MFB-ASR (log-mel-filterbank ASR), MFCC-ASR (mel-frequency cepstral coefficient ASR)
- **CI features**
  - Extracted over 8ms frames with 2ms frame shift
  - STFT-CI (log short-time Fourier transform), ACE-CI (Advanced Combination Encoder [7]), MFB-CI

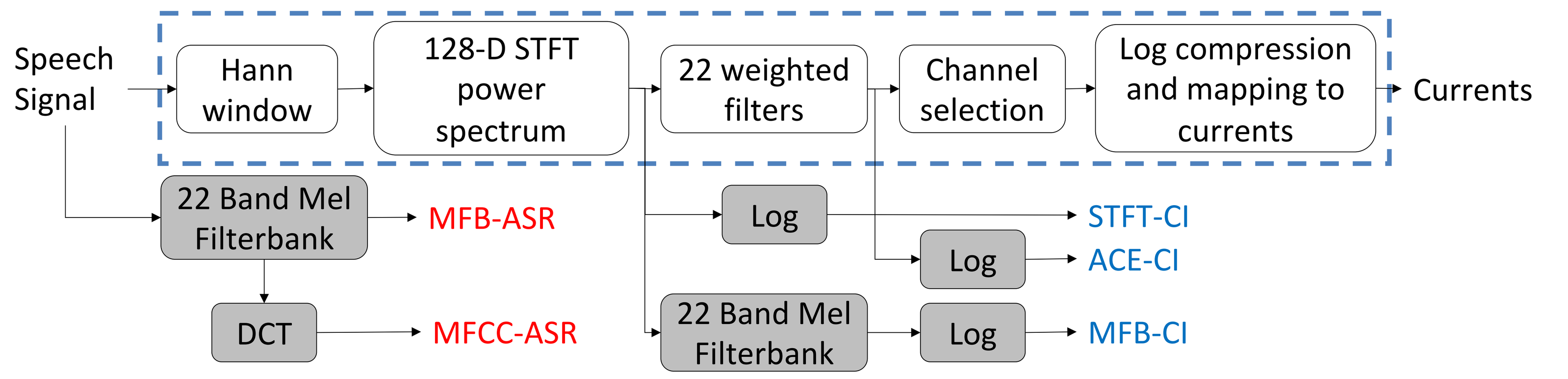

Fig. 8: Feature extraction framework. This figure shows ASR features and CI-inspired features are extracted within the ACE CI processing pipeline.

- Models: unidirectional long short-term memory (LSTM) or bidirectional long short-term memory (BLSTM) followed by softmax layer for classification

Training and Testing Datasets

- **Anechoic**: Speech was obtained from the TIMIT database [8]
- **Reverberant**: Speech signals were convolved with room impulse response functions (RIRs) from the Aachen database [9], which contains recordings from various acoustic environments as well as left and right channels
- Table 1 shows the speech stimuli and acoustic environments that were used in the training validation, and testing sets

Dataset	Speech Stimuli	Acoustic Environments
Training	TIMIT training set	• 25% anechoic • 75% reverberant (lecture hall and meeting room)
Validation	TIMIT development set	• 25% anechoic • 75% reverberant (lecture hall and meeting room)
Testing	TIMIT testing set	• Anechoic • Reverberant (office room and stairway)

Table 1: Training, validation, and testing sets. This table shows the speech stimuli and the acoustic environments that were used in the training, validation, and testing sets.

Results

Phoneme Classification

- Table 2 shows the percent of correctly identified phonemes
- CI-inspired features (LSTM-ACE-CI and LSTM-MFB-CI) outperformed ASR features (LSTM-MFB-ASR)

	Model	Anechoic	Office	Stairway
	Baseline (majority class)	25.8	25.8	25.8
CI features	LSTM-STFT-CI	62.4±0.5	48.9±0.9	45.7±1.0
	LSTM-ACE-CI	64.0±0.5	**50.8±0.6**	47.1±0.2
	LSTM-MFB-CI	**64.1±0.6**	50.6±0.4	**47.5±0.7**
ASR features	LSTM-MFB-ASR	62.6±0.3	49.5±0.3	44.6±0.9
	BLSTM-MFCC-ASR	71.1±0.2	58.9±0.5	55.9±0.4

Table 2: Percent of correctly identified phonemes. Values indicate the mean ± 1 standard deviation over five model instances trained using different random weight initializations. Bolded values indicate best performing unidirectional LSTMs.

Manner of Articulation Classification

- Table 3 shows the percent of correctly identified manners of articulation
- Higher overall accuracy than phoneme classification
- Similar trend to phoneme classification where LSTM-ACE-CI and LSTM-MFB-CI provided highest levels of performance

	Model	Anechoic	Office	Stairway
	Baseline (majority class)	37.3	37.3	37.3
CI features	LSTM-STFT-CI	82.2±0.4	70.8±0.8	68.5±0.3
	LSTM-ACE-CI	**82.9±0.1**	**72.1±0.2**	69.2±0.3
	LSTM-MFB-CI	**82.9±0.4**	72.0±0.4	**69.3±0.7**
ASR features	LSTM-MFB-ASR	81.4±0.3	70.1±0.3	66.3±0.7
	BLSTM-MFCC-ASR	85.2±0.1	77.2±0.2	74.5±0.1

Table 3: Percent of correctly identified manners of articulation. Values indicate the mean ± 1 standard deviation over five model instances trained using different random weight initializations. Bolded values indicate best performing unidirectional LSTMs.

Conclusion

- Overall goal was to develop classification model to categorize phonetic units within constraints of CI processor
- Results showed comparable levels of performance between traditional ASR features and CI-compatible features
- Future work will aim to develop phoneme-specific mask estimation algorithm where prediction from phoneme classification model is used to activate relevant mask estimation model

References

[1] K. Kokkinakis, O. Hazrati, and P. C. Loizou, "A channel-selection criterion for suppressing reverberation in cochlear implants," *J. Acoust. Soc. Am.*, vol. 129, no. 5, pp. 3221–3232, 2011.

[2] Medical Illustrations by NIH, Medical Arts and Photography Branch. "Ear with cochlear implant." [Online].

[3] R. Lyon, "A computational model of binaural localization and separation," in ICASSP '83. IEEE International Conference on Acoustics, Speech, and Signal Processing, 1983, vol. 8, pp. 1148–1151.

[4] J. M. Desmond, "Using channel-specific models to detect and remove reverberation in cochlear implants," Ph.D. Dissertation, Duke University, 2014.

[5] Z.-Q. Wang, Y. Zhao, and D. Wang, "Phoneme-specific speech separation," in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Mar. 2016, vol. 2016-May, pp. 146–150,

[6] P. Scanlon, D. P. W. Ellis, and R. B. Reilly, "Using broad phonetic group experts for improved speech recognition," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 15, no. 3, pp. 803–812, 2007.

[7] A. E. Vandali, L. A. Whitford, K. L. Plant, and G. M. Clark, "Speech perception as a function of electrical stimulation rate: Using the nucleus 24 cochlear implant system," Ear Hear., vol. 21, no. 6, pp. 608–624, 2000.

[8] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT:," Gaithersburg, MD, 1993.

[9] M. Jeub, M. Schafer, and P. Vary, "A Binaural Room Impulse Response Database for the Evaluation of Dereverberation algorithms," *Proc. 16th International Conference on Digital Signal Processing* , Santorini, Greece. 2009.

Duke UNIVERSITY | ELECTRICAL & COMPUTER ENGINEERING

This work was supported by the National Institute on Deafness and Other Communication Disorders under Award Number R01DC014290-05.

The Titan V GPU used in this work was provided by the NVIDIA GPU Grant Program.