Automatic Speech Recognition with Primarily Temporal Envelope Information

Payton Lin[1], *Fei Chen*[2], *Syu Siang Wang*[1], *Ying Hui Lai*[1], *Yu Tsao*[1]

[1] Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
[2] Division of Speech & Hearing Sciences, The University of Hong Kong, Hong Kong
{paytonlin,sypdbhee,jackylai,yu.tsao}@citi.sinica.edu.tw, feichen1@hku.hk

Abstract

The aim of this study is to devise a computational method to predict cochlear implant (CI) speech recognition. Here, we describe a high-throughput screening system for optimizing CI speech processing strategies using hidden Markov model (HMM)-based automatic speech recognition (ASR). Word accuracy was computed on vocoded CI speech synthesized from primarily multi-channel temporal envelope information. The ASR performance increased with the number of channels in a similar manner displayed in human recognition scores. Results showed the computational method of HMM-based ASR offers better process control for comparing signal carrier type. Training-test mismatch reduction provided a novel platform for reevaluating the relative contributions of spectral and temporal cues to human speech recognition.

Index Terms: Cochlear implant, HMM-based ASR, vocoder

1. Introduction

The cochlear implant (CI) is a surgically implanted device that can restore partial hearing by electrically stimulating the auditory nerve. Speech processing strategies extract temporal envelope information from bandpass filter outputs to modulate pulse trains at specified electrodes [1]. Speech recognition can be achieved with primarily amplitude and temporal cues, but CI users have obtained various degrees of benefit in their perception of speech from electrical stimulation. Considerable effort has been made to determine the appropriate number and placement of electrodes. Acoustic vocoder simulations derived from CI speech processing strategies have been presented to normal-hearing (NH) listeners to predict optimal performance [2]. The presentation of a dynamic temporal pattern in only a few broad spectral regions was sufficient for recognition of speech. For CI users, average performance increased with the number of electrodes, but no change was observed as the number of electrodes was increased from 7 to 20 [3]. These results indicate that CI users are not able to make full use of the spectral information on all electrodes.

Recent studies have shown that temporal waveform envelope provides significant information for hidden Markov model (HMM)-based automatic speech recognition (ASR) [4, 5]. The recognition of cochlear implant-like spectrally reduced speech (SRS) was computed using mel frequency cepstral coefficients (MFCC) as acoustic features. The triphone HMMs were trained on TI-digits, a clean speech training database used for small vocabulary tasks [6]. High levels of ASR word accuracy was achieved using 8-, 16-, and 24-subband SRS. The present study expands on previous research in efforts to determine optimal speech processing strategies for CI users. Synthesized vocoded testing sets approximate the effects of reduced spectral analysis and source segregation in the auditory periphery and brainstem [7]. Training sets can be manipulated to approximate the effects of impaired lexical, prosodic, and phonetic analysis in the cortex. Postoperative language development for postlingually deaf CI recipients is simulated with large vocabulary speech databases, context-dependent acoustic models, and tri-gram language models.

A critical issue for optimizing vocoder simulations has been determining the appropriate signal carrier type. The extracted envelope signal has been used to modulate band-limited white noise (noise-vocoder) [2]. However, CI users reported that electric signals sounded more like “beep tones,” leading to simulations resynthesized as the sum of sine waves at the center of the channels (tone-vocoder) [8]. When presented to NH subjects, tests of sentence intelligibility showed no significant differences between tone-vocoders and noise-vocoders. However, the effects of subject training from short-term listening practice and task familiarization from sequential test order has been questioned [8, 9]. For instance, tone-vocoded sentences were more intelligible than noise-vocoded sentences when using different speech material, randomized test order, and providing no feedback [9]. Therefore, the computational method of HMM-based ASR offers better process control for comparing signal carrier type.

Training-test mismatch reduction also provides a novel platform for reevaluating the relative contributions of spectral and temporal cues to human speech recognition. Human psychoacoustic tests are often inconclusive due to intersubject variability. For instance, CI performance actually decreased with increasing numbers of electrodes for experienced CI users with the longest duration of implant usage (ex: subject N5 and N2) [3]. This unexplained trend can be accounted for by systematically varying ASR system training to differentiate the acoustic front end from HMM-based temporal pattern matching. Similarly, a comparison of training sets will reveal the extent of linguistic or “top-down” knowledge contributing to NH ceiling effects at low number of channels [2, 8].

This paper is organized as follows: Section 2 describes the HMM-based ASR system, Section 3 determines the appropriate signal carrier type for simulating CI speech processors, and also evaluates the effects of human subject training. Section 4 concludes and discusses future research for improved speech processing.

2. System description

2.1. Speech material

Speech data was selected from Aurora-4, a standardized database used for evaluating large vocabulary continuous speech recognition (LVCSR) performance [10, 11]. Two training sets were provided in Aurora-4. The clean-condition training set was adopted as the original training material. In addition, Aurora-4 contains fourteen test sets. The clean testing data were used as the original testing data for preparing all testing materials. Speech utterances in Aurora-4 were acquired from the Wall Street Journal (WSJ0) corpus [12]. The 16 kHz sampling rate condition was used throughout the experiments.

2.2. Vocoder synthesis algorithm

The input speech materials were first processed through a pre-emphasis filter (first-order Butterworth high-pass filter at 1200 Hz), and then band-passed filtered into N frequency bands between 200 and 7,000 Hz using fourth-order Butterworth filters. The equivalent rectangular bandwidth scale was used to allocate the N channels with the specific bandwidth [13]. The temporal envelope from each frequency band was extracted by half-wave rectification and low-pass filtering (fourth-order Butterworth filter with a cutoff frequency of 160 Hz). For implementation of the noise-vocoder, the envelope of each frequency band was used to modulate white noise, followed by band-limiting using the same Butterworth band-pass filters, as shown in Figure 1. For implementation of the tone-vocoder, the envelope of each frequency band was used to modulate a sine-wave generated at the center frequency of the analysis band. The envelope-modulated noises or sine-waves of each frequency band were summed and then normalized to yield the same root-mean-square amplitude as the input speech.

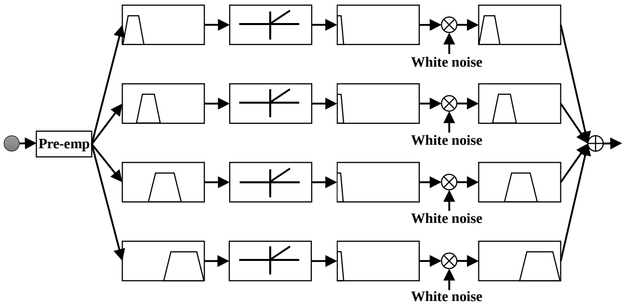

Figure 1: *Noise-vocoder synthesis with N=4 channels.*

2.3. HMM-Based ASR system training

Figure 2 shows the first experimental protocol. The testing set included 166 clean speech utterances, which were processed with either noise-vocoder or tone-vocoder synthesis parameters and decomposed into different numbers of channel N = [1, 2, 4, 6, 8, 16, 24].

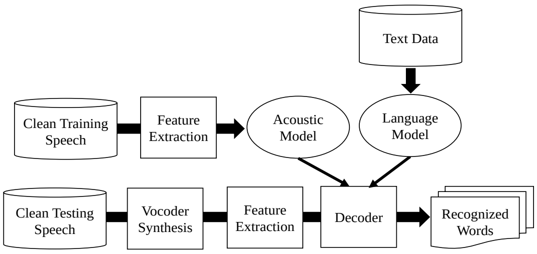

Figure 2: *Overview of first experimental protocol*.

The training set included 7,138 clean speech utterances, which were used to train acoustic models. Each utterance in the training and testing sets was converted into a sequence of MFCC vectors. Each vector included 13 static components plus their first- and second- order time derivatives. The frame length and shift were set to 32 ms and 10 ms, respectively. Cepstral mean subtraction (CMS) [14] was applied for each feature vector. The maximum likelihood (ML) training was applied to estimate a set of context-dependent triphone acoustic models. Each triphone was characterized by an HMM, which consists of three states, with eight Gaussian mixtures per state. A single-pass Viterbi beam search-based decoder was used along with a standard 5K lexicon [15] and tri-gram language model. The language model was prepared using the MIT language modeling (MITLM) toolkit [16] based on the reference transcription of training utterances. The hidden Markov model toolkit (HTK) [17] was adopted for the training and recognition processes.

Figure 3 shows the second experimental protocol. The synthesized tone-vocoded testing sets from the first experimental protocol were used. Four tone-vocoded training sets were prepared in total. Speech material from these sets were synthesized tone-vocoded speech using N=4, 8, and 16 channels. In addition, a multi-style training set was prepared by randomly selecting 1/3 speech data from each of the other vocoded training sets at N=4, 8, and 16, thus summing to 7,138 utterances in total. With these four vocoded training sets we thus trained four acoustic models using the HTK toolkit.

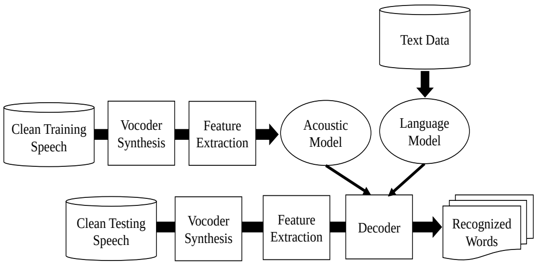

Figure 3: *Overview of second experimental protocol*.

3. Experimental Results

3.1. Differences between tone and noise carriers

The sine-wave carrier of the tone-vocoder is a single frequency component and thus has a fixed amplitude envelope. Spectral sidebands are generated from amplitude modulation that reflect the spectral content of the envelope. The band-limited white-noise carrier of the noise-vocoder only provides information on the spectral distribution of energy and has a rapidly fluctuating envelope. The carrier spectra of noise-bands mask side bands as shown in Figure 4.

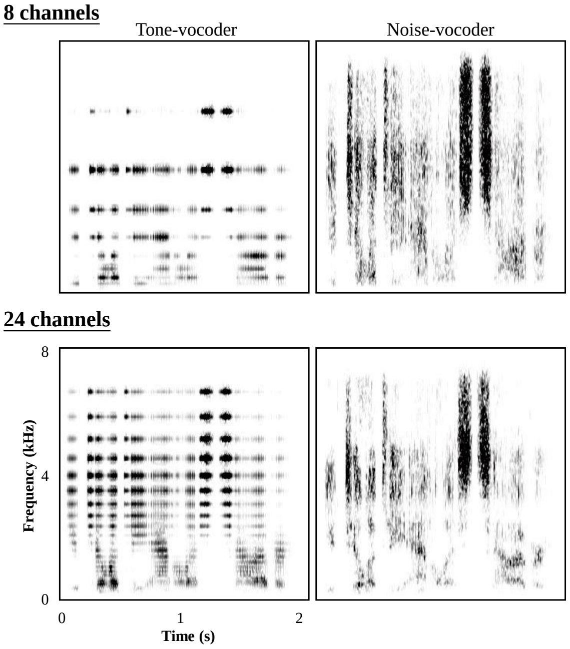

Figure 4: *Spectrograms of a vocoded speech utterance*.

Therefore, better modulation detection is expected for the tone-vocoder compared to the noise-vocoder.

3.2. Effects of the number of channels

3.2.1. First experimental protocol

Figure 5 displays ASR performance for vocoded testing speech as a function of the number of channels. The triphone HMMs were trained on the Aurora-4 clean-condition speech data. The ASR performance advantage of tone-vocoders over noise-vocoders at N=6 and 8 support several existing theories from both CI and NH speech recognition experiments [3, 9]. The sidebands of the tone-vocoder provided a periodic temporal structure that enhanced discrimination capabilities of acoustic units. In contrast, sidebands of noise-band carriers were masked by carrier spectra, limiting the delicate structure of triphones and speech attributes. Therefore, the tone-vocoder contributed relative temporal envelope information that reduced mismatch in the feature extraction stage at N=6 and 8.

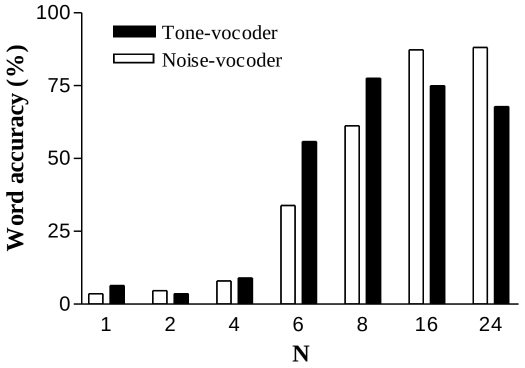

Figure 5: *ASR performance (word accuracy in %) computed on 166 speech utterances synthesized from seven values for the number of channels N, and two carrier types for vocoder.*

Training-test mismatch occurred for the tone-vocoder as the number of channels increased from N=8 to 24, with a reduction of word accuracy from 77.50% to 67.81%. As more channels were implemented, unresolved sidebands of amplitude modulation likely distorted low frequency bands. These ASR results account for unexplained CI occurrences [3], where sentence recognition scores for subjects N5 and N2 decreased when implementing more than 7 electrodes. Several speech processors currently stimulate only a subset of the bands with maximal activities (e.g., 8 of the 22 electrodes) in order to avoid simultaneous stimulation between different electrodes. The triangular filter characteristics of mel-cepstral analysis support the efficacy of these n-of-m strategies such as SPEAK, where typically the six channels with the largest amplitudes are selected to activate six corresponding electrodes [18]. Moreover, the distortions observed in ASR feature-space [19] at high number of channels establish the tone-vocoder front end of ASR signal-space as the most accurate simulation for actual CI performance.

However, the original aim of previous research has been to determine how to effectively use all the spectral information on all the available electrodes [3, 8]. For the continuous interleaved sampling (CIS) strategy, the number of analysis bands is the same as the number of electrodes [1]. Figure 5 also shows ASR performance for the noise-vocoder increased monotonically as the number of channels increased from N=1 to 24, with word accuracy improving from 3.54% to 88.10%. The 24-channel noise-vocoder displayed word accuracy nearly equivalent to wideband clean testing speech (92.67%). Therefore, the contiguous bands simulated by white-noise carriers could substantially improve electrode-to-nerve interactions.

3.2.2. Clinical implications

Several factors must be considered in order to transfer the benefits of noise-vocoders to the electrode-to-nerve interface. The stimulating electrodes of cochlear implants deliver a broad electrical current field while presenting temporal information with sequences of pulses at high rates of stimulation [1]. Electrical stimulation produces highly synchronous activity across auditory nerve fibers, depolarizing the peripheral processes and spiral ganglion cells at high intensity levels. Forward masking experiments displayed excitation patterns that had a spatial band pass characteristic with a peak in the region of the masked electrode [20]. Electrical temporal modulation transfer functions (TMTFs) indicated that high carrier levels allow CI listeners to detect much smaller amplitude modulations than achieved by normal hearing [21].

The ASR results support two conclusions. The place code of electric pitch should imitate the spatial separation of the white-noise carrier in order to make full use of all of the existing electrodes on modern arrays. The temporal code of electric pitch should employ the pitch-related periodic temporal fluctuations of sine-wave carrier sidebands at N=6 and 8 channels in order to imitate the phase-locking of action potentials. Figure 6 illustrates the underlying physiological mechanisms of the tone-vocoder front end as related to channel interactions during electric acoustic stimulation (EAS). Results demonstrate the dependency of the area and temporal pattern of neural responses to the distance of current flow from the stimulating electrode [22].

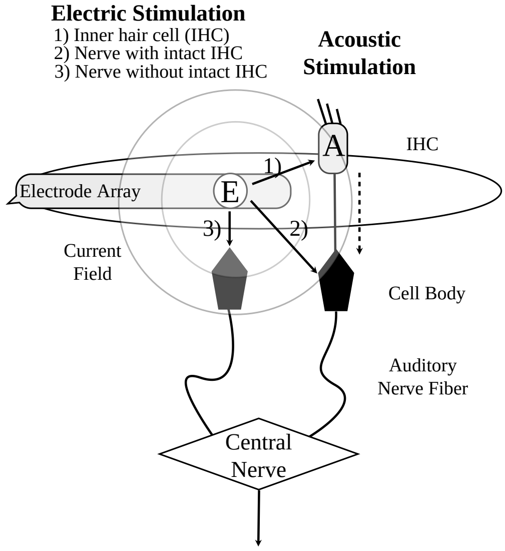

Figure 6: *Place and temporal code in vocoder signal-space.*

3.3. Effects of training

3.3.1. Second experimental protocol

The present study expands on previous research that used SRS for small-vocabulary tasks [4, 5]. Speech data was selected from Aurora-4, a standardized LVCSR database that better represents postlingually deaf CI recipients. Word accuracies computed on both noise-vocoders and tone-vocoders were much lower at N=4 compared to the SRS correspondence with the equivalent 160 Hz AM bandwidth. The maximum length of seven-digit sequences in TI-digits [7] likely caused SRS ceiling effects in saturated ASR performance at low numbers of channels. While the material selected for training databases has been known to notably influence subsequent ASR performance [23, 24], the second experimental protocol was designed to compare the effects of human subject training [8, 9].

Linguistic or "top-down" knowledge from short-term listening practice and visual feedback has been questioned for causing NH ceiling effects at low number of channels [2, 8, 9]. The ASR system evaluates the effects of training by differentiating the acoustic front end from HMM-based temporal pattern matching. Table 1 shows training-test mismatch reduction occurred when the vocoded testing speech shared equivalent numbers of channels as the vocoded training sets. The N-matched condition at N=4 improved word accuracy by 77.13% compared to the clean-condition training set results.

In addition, the effects of subject experience and sequential test order has been questioned [8, 9]. For instance, NH subjects were presented test conditions in increasing order of difficulty (e.g., N=9, followed- by N=8, followed by N=7, etc.). The multi-style training set evaluates the effects of task familiarization by combining speech material randomly from tone-vocoded training sets at N=4, 8, and 16. Table 1 demonstrates how the accumulation of subject training during the testing procedure can enhance human speech recognition scores.

		Training Set			
	N	**4**	**8**	**16**	**Multi-**
Testing Set	**4**	86.08	5.89	17.46	83.02
	8	5.56	91.38	89.87	89.17
	16	13.15	88.03	90.76	89.47

Table 1: *ASR performance (word accuracy in %). The multi-style training set was prepared from 1/3 speech data from each of the other three vocoded training sets at N=4, 8, and 16.*

Learning effects and differences in duration of implant use have also hindered speech recognition tests for CI subjects [3]. In order to evaluate speech recognition as a function of the number of electrodes, CI subjects were required to wear experimental speech processors for 2 days each before testing. While no difference was reported for average performance with 4-electrode and 20-electrode processors, sentence intelligibility scores for several individual subjects decreased with increasing numbers of electrodes.

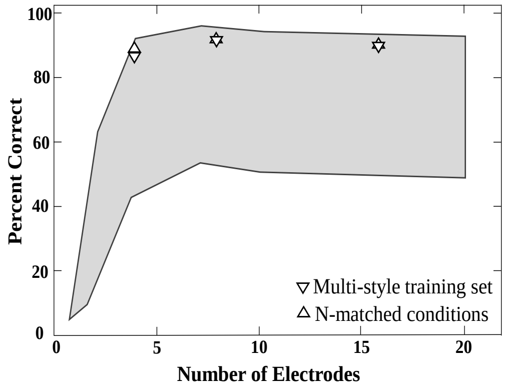

Figure 7: *Comparison of the multi-style training set and N-matched conditions with the range of scores from CI subjects from the Fishman et al. (1997) study (filled area).*

Figure 7 compares the range of results of the 11 CI subjects from the Fishman et al. study with ASR performance (% correct) computed on the tone-vocoded training sets. In the first experimental protocol, the tone-vocoder front end was established as the most accurate simulation of actual CI performance by accounting for how sentence recognition scores could possibly decrease for 2 of the 11 CI subjects when implementing more than 7 electrodes. A comparison of the effects of ASR training further elucidates the best performance of the CI listeners or "star" performer effect of subjects training on testing material prior to examination. The Bayesian model of human concept learning [25] should be accounted for when reevaluating psychoacoustic testing.

4. Conclusions

The present study proposes HMM-based ASR as an effective screening system for optimizing speech processing strategies for individual CI users. The compactness of automated systems allows higher efficiency, better process control, and faster analysis time for predicting CI speech recognition. Moreover, computational simulation provides a safer testing platform while alleviating patient workload, time commitment, and travel cost. Future work will address EAS and current-state-of-the-art deep learning based ASR systems.

5. Acknowledgements

This research was supported by Seed Funding for Basic Research, The University of Hong Kong, and by National Science Council of Taiwan under contracts NSC101-2221-E-001-020-MY3.

6. References

[1] Wilson, B.S., Finley, C.C., Lawson, D.T., Wolford, R.D., Eddington, D.K. and Rabinowitz, W.M., "Better speech recognition with cochlear implants," Nature, 352, pp. 236–238, 1991.

[2] Shannon, R.V., Zeng, F.-G., Kamath, V., Wygonski, J. and Ekelid, M., "Speech recognition with primarily temporal cues," Science, 270, pp. 303–304, 1995.

[3] Fishman, K.E., Shannon, R.V. and Slattery, W.H., "Speech recognition as a function of the number of electrodes used in the SPEAK cochlear implant speech processor," J. Speech Hear. Res., 40, pp. 1201-1215, 1997.

[4] Do, C.-T., Pastor, D. and Goalic, A., "On the recognition of cochlear implant-like spectrally reduced speech with MFCC and HMM-based ASR", IEEE Trans. on Audio, Speech, and Language Processing, 18 (5), pp. 1065–1068, 2010.

[5] Do, C.-T., Pastor, D. and Goalic, A., "A novel framework for noise robust ASR using cochlear implant-like spectrally reduced speech," Speech Commun., 51(1), pp. 119-133, 2012.

[6] Leonard, R., "A database for speaker-independent digit recognition," in Proc. IEEE ICASSP, 9, pp. 328-331, 1984.

[7] Morgan, N., Bourlard, H., and Hermansky, H., "Automatic speech recognition: an auditory perspective," in Speech Processing in the Auditory System, New York: Springer, 2004.

[8] Dorman, M. F., Loizou, P. C. and Rainey, D., "Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs," J. Acoust. Soc. Am., 102 (4), pp. 2403–2411, 1997.

[9] Whitmal, N.A., Poissant, S.F., Freyman, R.L. and Helfer, K.S., "Speech intelligibility in cochlear implant simulations: Effects of carrier type, interfering noise, and subject experience," J. Acoust. Soc. Am., 122 (4), pp. 2376–2388, 2007.

[10] Parihar, N. and Picone, J., "Aurora working group: Dsr front end lvcsr evaluation au/384/02," in Institute for Signal and

Information Processing Report, 2002.

[11] Parihar, N., Picone, J., Pearce, D. and Hirsch, H. G., "Performance analysis of the Aurora large vocabulary baseline system," in Proc. EUSIPCO'04, 553–556, 2004.

[12] Paul, D. B. and Baker J. M., "The design for the wall street journal-based CSR corpus," in Proc. ICSLP'92, 357-362, 1992.

[13] Glasberg, B. and Moore B., "Derivation of auditory filter shapes from notched-noise data," Hear. Res., 47, pp. 103–138, 1990.

[14] Furui, S., "Cepstral analysis technique for automatic speaker verification," IEEE Trans. Acoustics, Speech, Signal Processing, 29, pp. 254-272, 1981.

[15] "The CMU Pronouncing Dictionary," Carnegie Mellon University, Pittsburg, Pennsylvania, USA, June, 2001.

[16] Hsu, B.-J., and Glass J., "Iterative language model estimation: Efficient data structure and algorithms," in Proc. INTERSPEECH, pp. 841-844, 2008.

[17] Young, S., Evermann, G., Gales, M., Hain T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V. and Woodland, P., "The HTK Book (for HTK Version 3.3)." Cambridge University Engineering Department, 2005.

[18] McDermott, H.J., Mckay. C.M. and Vandali, A.E., "A new portable sound processor for the University of Melbourne/Nucleus Limited multielectrode cochlear implant," J. Acoust. Soc. Am., 91 (6), pp. 3367–3371, 1992.

[19] Sankar, A., and Lee, C.-H., "A maximum-likelihood approach to stochastic matching for robust speech recognition," IEEE Trans. Speech Audio Processing, 4 (3), pp. 190-202, 1996.

[20] Chatterjee, M., and Shannon, R.V., "Forward masked excitation patterns in multielectrode electrical stimulation," J. Acoust. Soc. Am., 105, pp. 2565-2572, 1998.

[21] Shannon, R.V., "Temporal modulation transfer functions in patients with cochlear implants," J. Acoust. Soc. Am., 91 (4), pp. 2156-2164, 1992.

[22] Lin, P., Turner, C.W., Gantz, B.J., Djalilian, H.R., and Zeng, F.-G, "Ipsilateral masking between acoustic and electric stimulations," J. Acoust. Soc. Am., 130 (2), pp. 858-865, 2011.

[23] Lippmann, R. P., Martin, E. A. and Paul D. B., "Multi-style training for robust isolated-word speech recognition," in Proc. ICASSP, pp. 705–708, 1987.

[24] Huang, X., Acero, A., and Hon, H.-W., Spoken Language Processing: A Guide to Theory, Algorithm and System Development, Prentice Hall, 2001.

[25] Tenebaum, J.B., "Bayesian modeling of human concept learning," Advances in Neural Information Processing Systems, 11, Cambridge, MIT Press, pp. 59-65, 1999.