RELATING THE ACOUSTIC SPACE OF VOWELS TO THE PERCEPTUAL SPACE IN COCHLEAR IMPLANT SIMULATIONS

Chuping Liu[1] *and Qian-Jie Fu*[2,3]

[1]Department of Electrical Engineering, [2]Department of Biomedical Engineering,
University of Southern California, Los Angeles, CA, 90007
[3]Department of Auditory Implants and Perception, House Ear Institute
2100 West Third Street, Los Angeles, CA, 90057, USA
Email: chupingl@usc.edu

ABSTRACT

In automatic speech recognition, speech features are measured and used as reference templates for machine learning and recognition. The differences between these features can be used to calculate relative acoustic distances between phonemes. However, when speech signals are spectrally degraded, as in electric hearing with cochlear implants, it is unclear whether these acoustic distances can predict speech recognition performance. The present study measured acoustic distances between spectrally degraded vowel tokens and investigated the relation between acoustic vowel space and perceptual vowel space. After processing vowel tokens using a cochlear implant simulation, Mel-frequency cepstrum was extracted from each token; features were then time aligned and the weighted Euclidean distance was calculated between all tokens. Results demonstrated a significant correlation between vowel perception data and averaged acoustic distance between vowel tokens, for a variety of experimental conditions. These results suggest that acoustic distance between phonemes may well predict recognition performance of spectrally degraded speech.

1. INTRODUCTION

Cochlear Implants (CIs) represent speech signals by using the temporal envelope extracted from frequency analysis bands to modulate pulse trains delivered to appropriate implanted electrodes. CI users' perception performance on such degraded speech stimulation typically not only depends on the front end CI speech processor, but also on the cochlear and central brain mapping. Variable factors such as number of electrodes, acoustic input frequency range, stimulation rate and envelope cutoff frequency differs in CI speech processors. In addition, individuals greatly differ in electrode insertion depth and healthy neural population.

Previous CI speech perception studies have systematically explored some speech processor parameters while fixing others [1, 5]. These parametric studies are not only time consuming, but results are often difficult to interpret because of interactions between fixed and varied parameters. However, if the signal degradation due to these variable factors can be quantified, the above disadvantages might be lessened or avoided. Further, quantification of degraded speech can be potentially correlated to perception data; hence it might help to evaluate the performance of individual patients according to the specifics of their implant device, etiology and parameter settings.

The spectrally degraded speech patterns experienced by CI users can be simulated using an acoustic noise-band vocoder [1]. CI patient performance has been consistently shown to be comparable to that of normal-hearing (NH) subjects listening to similar processing conditions. In automatic speech recognition (ASR) field, the most popular front-end feature extraction method is Mel-cepstrum coefficients, which can yield high recognition rate. This implies that Mel-cepstrum coefficients may well expand parameter space and identify acoustic differences between speech signals, although machine learning may not always employ perception cues. Using this ASR feature extraction technique to map acoustic space processed by a CI simulation, perceptual data can then be compared to acoustic distance data.

The present study evaluated whether perceptual space of spectrally degraded speech via CI processing can be predicted by acoustic space of processed phonemes. The acoustic distance between vowel tokens processed by a CI simulation was measured for several speech processing conditions. These acoustic spaces were compared to perceptual data from NH listeners listening to the same CI simulation.

2. METHODS

2.1 Test materials

Tokens used for both acoustic analysis and closed-set vowel recognition tests were digitized natural productions drawn from speech samples collected by Hillenbrand et al. [2]. There were 12 phonemes in the stimulus set, including 10 monophthongs and 2 diphthongs, presented in a /h/vowel/d/ context (heed, hid, head, had, who'd, hood, hod, hud, hawed, heard, hoed, hayed). All stimuli were normalized to have the same long-term root mean square (RMS) values. Acoustic spaces of variable experiments were measured from 2 male and 2 female talkers while recognition tests were measured from 5 males and 5 females.

2.2 CI simulation: degraded speech synthesis

In acoustic CI simulation, the effects of several processing conditions were investigated, including the number of frequency bands, frequency band partitions, spectral smearing and temporal envelope cutoff frequency. A noise-band vocoder was used to simulate a CI speech processor fitted with the Continuously Interleaved Sampling (CIS) strategy [3]. The processor was implemented as follows. The signal was first processed through a pre-emphasis filter (high-pass with a cut off frequency of 1200 Hz and a slope of 6 dB/octave). An input frequency range was band-passed into a number of frequency analysis bands. The temporal envelope was extracted from each frequency band by half-wave rectification and low-pass filtering. The envelope of each band was used to modulate a wideband noise, which was then spectrally limited by the same bandpass filter as the one used in the original analysis band. Finally, the modulated carriers of each band were summed and the overall level was adjusted to be the same RMS level as the original speech. Unless otherwise noted, for all conditions, four spectral bands were used, the overall input frequency range was 100 – 4000 Hz, the analysis and carrier band filter slopes were 24 dB/octave and the temporal envelope filter cutoff frequency was 160 Hz. Specifically, the number of frequency bands, the slope and the distribution of the analysis filters, and the cut-off frequency of the envelope filter depended on experimental conditions. Further details of speech processor parameters are described below.

2.1.1 The number of spectral channels

In CI speech processing, perhaps the most important parameter is the number of spectral channels. Previous studies with both CI listeners and NH subjects listening to a CI simulation have shown that performance generally improves with increasing numbers of spectral channels [1, 4, 5]. In the present study, changes in acoustic space due to spectral resolution were compared to NH listener's perceptual data. A number of spectral resolution conditions were analyzed: 8-, 6-, 4-, 3-, 2-, and 1-channel speech, as well as unprocessed speech. The input frequency range was linearly divided by number of frequency bands for each test condition.

2.1.2 Frequency allocation

Another important speech processor parameter is the frequency allocation, which determines the assignment of acoustic frequencies to the place of stimulation in the cochlea. Several studies have shown that speech recognition can be significantly affected by frequency allocation, especially when the number of frequency bands is relatively small [5]. In the present experiment, the effect of frequency allocation on the acoustic space of 4-channel processed speech was analyzed. Seven frequency allocation tables were generated according to Eq. 1:

$$f(k) = F\frac{10^{\frac{pXk}{N}} - \alpha}{10^{pX} - \alpha}, \qquad k = 1,2,...N \qquad (1)$$

where k is the channel number, α is a constant, N is the total number of frequency bands, X is the cochlear extent (in mm) relative to the maximum frequency F (Hz), and p is the frequency warping factor (ranging between 0.01 and 0.06 in 0.01 steps; when p=0.01, the frequency-to-place mapping was nearly linear and when p=0.06, the mapping was nearly logarithmic). X and F are related according to Eq. 2:

$$X = \frac{1}{0.06}\log_{10}(\frac{F}{165.4} + \alpha) \qquad (2)$$

which is merely a reverse form of the place-frequency mapping proposed by Greenwood [6]. Six frequency allocation conditions were tested (P1 – P6, corresponding to the range of frequency warping factors p). For comparison, a linear frequency allocation was tested (P0). The above frequency allocation method was applied to both the analysis and carrier bands.

2.1.3 Spectral smearing

When spectral details in speech signals are smeared (because of channel/electrode interactions), CI users' effective spectral resolution can be further reduced. Several studies have shown that as the amount of spectral smearing increased, recognition performance reduced [5]. In the present experiment, the effect of spectral smearing on the acoustic space of processed vowels was analyzed. Spectral smearing was approximated by varying the degrees of overlapping between carrier bands. The frequency allocation was fixed at condition P2 as described above in section 2.1.2. The analysis band filter slope was fixed (36 dB/octave), while the carrier band slope was varied between 36 dB/octave (no spectral smearing) and 6 dB/octave (spectrally smeared).

2.1.4 Temporal smearing

In CI speech processors, temporal envelope is typically extracted from each frequency analysis band by half-wave rectification and low-pass envelope filtering. Previous studies have shown that slowly varying temporal components (< 20 Hz) provide most useful phonetic information to CI and NH listeners, even with spectrally degraded speech [1,7]. In the present experiment, the effect of temporal smearing on the acoustic space of processed vowels was analyzed. The corner frequencies of 4 spectral bands were 300, 713, 1509, 3043, and 6000 Hz. Different degrees of temporal smearing were simulated by varying the cutoff frequency of envelope filter (640, 160, 40, 20, to 10 Hz), thereby limiting the available temporal cues.

2.3 Acoustic distance measurement and space definition

Acoustic distance between two tokens was defined as the least-cost mapping of time-aligned and path-weighted Euclidean distance of their Mel-cepstrum coefficients, as shown in Eq. 3:

$$D(S1, S2) = \min_{F}\left[\frac{\sum_{k=1}^{K} d(c(k)) \cdot w(k)}{\sum_{k=1}^{K} w(k)}\right] \qquad (3)$$

where *d(c(k))* is the Euclidean distance between two aligned Mel-cepstrum coefficient vectors, *w(k)* is a nonnegative path-weighting coefficient and *F* is a time warping function. The denominator $\sum w(k)$ is employed to compensate path length [8].

For each speech processing condition, acoustic distance between each vowel pair was measured and entered into an acoustic confusion matrix. Acoustic space for each processing condition was defined as the averaged value of the acoustic confusion matrix, which reflects the acoustic space volume that the speech processing condition expands. In parallel, perceptual space was defined as the averaged overall percent correct under each experimental condition, which reflects the actual perceptual volume achieved from experiments. Hence, bigger acoustic space volume implies a better distinction among vowels, which will potentially result in less confusion in speech perception (higher percent correct rate). For comparison purposes, both acoustic space and perceptual space were normalized to one. The normalized acoustic space was computed by dividing the acoustic space of spectrally degraded speech by the acoustic space of the original speech; while the normalized perceptual space was computed by rationalized arcsine transformation [10].

3. RESULTS AND DISCUSSION

3.1 The effect of the number of bands

Figure 1 shows the normalized acoustic space as a function of the number of spectral bands. For comparison purposes, normalized perceptual space for NH subjects listening to the same conditions is also shown in Figure 1 (see right axis). As the number of spectral channels increase, the acoustic space steadily expands. On average, as the number of channels was doubled, the acoustic space was expanded by about 21%, which is comparable to that of perception space under 8 channels (22%). Statistical analysis revealed a significant correlation between the normalized acoustic space and the perceptual space (r^2=0.974, p<0.0001).

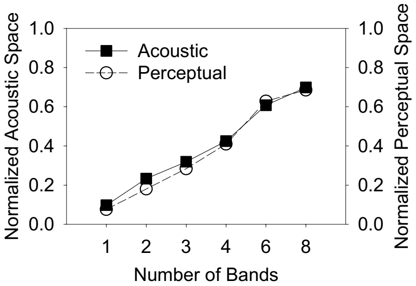

Figure 1: The effect of number of frequency bands

3.2 The effect of frequency allocation

Figure 2 shows the normalized acoustic space and perceptual space as a function of the frequency allocation. Both the peaks of acoustic space and perceptual space were found when frequency warping factor *p* = 0.02 (condition P2). The frequency allocation had a small but significant effect on acoustic space as the frequency allocation became more linear or more logarithmic. Statistical analysis again revealed a significant correlation between acoustic space and perceptual space (r^2=0.653, p=0.028).

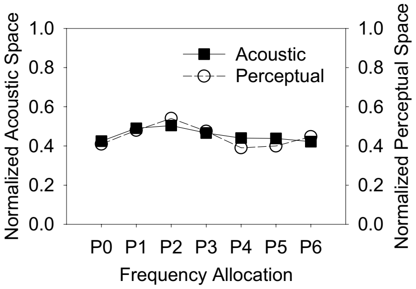

Figure 2: The effect of frequency allocation

3.3 The effect of spectral smearing

Figure 3 shows the normalized acoustic space and perceptual space as a function of the amount of spectral smearing. As the slope of the carrier bandpass filters became shallower (increasing the degree of spectral smearing), the acoustic space gradually reduced. Similarly, NH subjects' recognition performance worsened as the amount of spectral smearing increased. Statistical analysis revealed a significant correlation between the normalized acoustic space and the perceptual space (r^2=0.975, p=0.012).

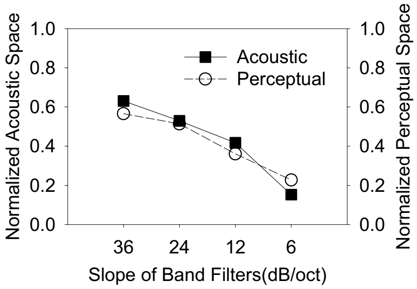

Figure 3: The effect of spectral smearing

3.4 The effect of temporal envelope cutoff frequency

The acoustic space for vowel tokens as a function of the amount of temporal envelope information is shown in Figure 4. Similar to the perception data, there was no significant change of acoustic space when the cutoff frequency of envelope filter was 10 Hz or above.

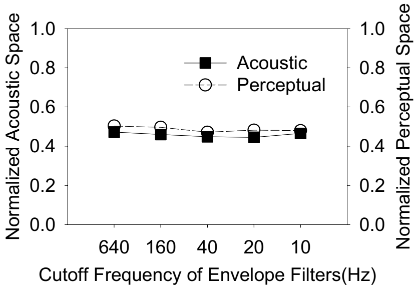

Figure 4: The effect of temporal envelope

4. GENERAL DISCUSSION AND CONCLUSIONS

Results from the present study demonstrated that most spectrum-related speech processor parameters significantly affect both acoustic space and perceptual data. A significant correlation between acoustic space and perceptual space was observed for all experimental conditions.

It is also interesting to note that speech processor parameters that had no effect on recognition performance (e.g., amount of temporal information) also did not significantly affect acoustic vowel space. Furthermore, confusion patterns between phonemes were matched between the acoustic and perceptual distances. For example, acoustic distance of the vowel pair "had/head" was generally smallest, while distance of the vowel pair "had/heed" was greatest and was more than 12 times greater than the "had/head" distance. These acoustic distances agree with perceptual confusions, which showed that "had" is more likely to be confused with "head" rather than "heed."

The results indicate that measuring acoustic space using dynamic time warped Mel-cepstrum coefficients could nicely predict perception data for a variety of parameter settings in a CI speech processor. These results are in agreement with recent research by Remus [9], but provide better perceptually related acoustic distances for most spectrally related parameters in CI speech processing.

In the present study, simulations investigated did not include spectral shift and compression, which is normally associated with CIs due to limited electrode insertion depth. Further research on it may shed light on the underlying mechanism of recognizing spectrally degraded and mismatched speech.

5. REFERENCES

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

[2] J. Hillenbrand, L. Getty, M. Clark, and K. Wheeler, "Acoustic characteristics of American English vowels," J. Acoust. Soc. Am. 97, pp. 3099-3111, 1994.

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

[4] K. Fishman, R.V. Shannon, and W.H. Slattery, "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.

[5] Q.-J. Fu, "Speech pattern recognition in electric hearing," Ph.D. Dissertation, University of Southern California, 1997.

[6] D.D. Greenwood, "A cochlear frequency-position function for several species -- 29 years later," J. Acoust. Soc. Am. 87, pp. 2592-2605, 1990.

[7] Q.-J. Fu and R.V. Shannon, "Effects of stimulation rate on phoneme recognition in cochlear implant users," J. Acoust. Soc. Am. 107, pp. 589-597, 2000.

[8] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," IEEE Trans. on Acoustic, Speech, Signal Processing Vol. ASSP-26, pp. 43-49, 1978.

[9] J.J. Remus and L.M. Collins, "Vowel and consonant confusion in noise by cochlear implant subjects: predicting performance using signal processing techniques," Proc. of 2004 IEEE International Conference on Acoustic, Speech, Signal Processing IV, pp. 13-16, 2004.

[10] G.A. Studebaker, A "rationalized" arcsin transformation, J. Speech Hear Res, 28(3), pp. 455-462, 1985.