RECOGNITION OF SPECTRALLY DEGRADED SPEECH IN NOISE WITH NONLINEAR AMPLITUDE MAPPING

Qian-Jie Fu and Robert V. Shannon

Department of Auditory Implants and Perception
House Ear Institute, 2100 West Third Street
Los Angeles, CA 90057

ABSTRACT

The present study measured phoneme recognition as a function of signal-to-noise level under conditions of spectral smearing and nonlinear amplitude mapping. Speech sounds were divided into 16 analysis bands. The envelope was extracted from each band by half-wave rectification and low-pass filtering and was then distorted by a power-law transformation whose exponents varied from a strongly compressive (p=0.3) to a strongly expanded value (p=3.0). This distorted envelope was used to modulate a noise which was spectrally limited by the same analysis filters. Results showed that phoneme recognition scores in quiet were reduced only slightly with either expanded or compressed amplitude mapping. As the level of background noise was increased, performance deteriorated more rapidly for both compressed and linear mapping than for the expanded mapping. These results indicate that, although an expansive amplitude mapping may slightly reduce performance in quiet, it may be beneficial in noisy listening conditions.

1. INTRODUCTION

Cochlear implants transform speech sounds into electrical signals that directly stimulate remaining auditory nerve fibers and can partially restore the speech sensations of profoundly deaf listeners. Modern multichannel cochlear implants divide speech sounds into multiple frequency bands and extract the temporal envelope information from each band. Then the acoustic envelope amplitude is converted into electric amplitude which is delivered to electrodes located in the different places within the cochlea. To recreate the tonotopic distribution of activity within the normal cochlea, the envelopes from low frequency bands are delivered to electrodes located near the apex and the envelopes from high frequency bands are delivered to basal electrodes. The improvement of speech performance from single-channel to multichannel device demonstrates a clear utilization of place cues in cochlear implant users [1].

In quiet conditions, most cochlear implant users with the latest implant device can achieve a high level of speech performance. However, performance deteriorates significantly in noisy environments [5, 8] even for the best cochlear implant user. The cause of the noise susceptibility of cochlear implant users has been investigated recently. Fu et al. [3] measured the recognition of spectrally degraded vowels and consonants as a function of signal-to-noise ratio in both normal-hearing subjects and cochlear implant users. The results showed that as the spectral information was reduced, speech recognition deteriorated only slightly in quiet conditions, but recognition deteriorated significantly more in noisy conditions. The performance of the best cochlear implant users was similar to that of normal-hearing subjects listening to a similar level of spectral reduction, suggesting that those implant listeners were making optimal use of the spectral cues available. As the spectral resolution was reduced the performance in noise decreased, demonstrating that the limited spectral resolution is a key factor causing the noise susceptibility. However, some of the cochlear implant listeners had poorer performance than processor-matched normal-hearing subjects, suggesting that those implant listeners were not receiving as many spectral channels of information as their number of electrodes, due to unknown factors. One possible additional factor is the loudness mapping function between acoustic amplitude and electric current.

Amplitudes in normal speech can range over 40 to 60 dB. However, implant listeners typically have dynamic ranges of only 6 to 15 dB in electric current, requiring the acoustic range to be compressed into the electric range. Fu and Shannon [2] measured vowel and consonant recognition as a function of the exponent of a power-function nonlinearity in both cochlear implant users and normal-hearing listeners. They found that, for both acoustic and implant listeners, the best performance was obtained when normal loudness was preserved. Performance deteriorated slightly when the amplitude mapping function was either more compressive or more expansive. Thus, instantaneous amplitude nonlinearity only has a minor effect on phoneme recognition in quiet.

The goal of the present study was to understand the effects of nonlinear amplitude mapping on recognition of spectrally degraded speech in noise. The recognition of vowels and consonants was measured in five normal hearing listeners as a function of signal-to-noise ratio, with the exponent of the amplitude-mapping power function as a parameter.

2. METHODS

2.1 Subjects

Five normal-hearing subjects between the ages of 25 and 35 years served as subjects. All subjects had thresholds better than 15 dB HL at audiometric test frequencies from 250 to 8000 Hz and all were native speakers of American English.

2.2 Test materials and procedures

Speech recognition was assessed for medial vowels and consonants. Vowel recognition was measured in a 12-alternative identification paradigm, including 10 monophthongs and 2 diphthongs, presented in a /h/-vowel-/d/ context (heed, hawed, head, who'd, hid, hood, hud, had, heard, hoed, hod, hayed). The

tokens for these closed-set tests were digitized natural productions from 5 male, 5 female, 5 children, drawn from the material collected by Hillenbrand et al. [4]. Consonant recognition was measured in a 16-alternative identification paradigm, for the consonants /b d g p t k l m n f s ʃ v z j θ/, presented in an /a/-consonant-/a/ context. Two repetitions of each of the 16 consonants were produced by three speakers (1 male, 2 female) for a total of 96 tokens (16 consonants * 3 talkers * 2 repeats). All test materials were stored on computer disk and were output via custom software to a 16 bit D/A converter (TDT DD1) at a 16-kHz sampling rate. Speech sounds were presented using a Tucker-Davis-Technologies (TDT) AP2 array processor in a host PC connected via an optical interface.

Each test block included 180 tokens for vowel recognition or 96 tokens for consonant recognition. A stimulus token was randomly chosen from all 180 tokens in vowel recognition and from 96 tokens in consonant recognition and presented to the subject. Following the presentation of each token, the subject responded by pressing one of 12 buttons in the vowel test or one of 16 buttons in the consonant test, each marked with one of the possible responses.

All subjects started with a training session. Two extreme mappings were used as training conditions. Each training session included 8 consecutive test blocks with the same mapping condition and the same speech material. Feedback was provided. The order of training conditions (two mapping conditions and vowel/consonant tests) was randomized across subjects. Subjects started the test sessions after all training conditions were finished. In the test sessions, the order of S/N ratio conditions was randomized. The order of the five mapping conditions, and the order of the vowel and consonant tests, was counterbalanced across subjects. No feedback was provided in test sessions.

2.3 Signal processing

The speech signal was mixed with simplified speech spectrum-shaped noise (constant spectrum level below 800 Hz and 10-dB/octave role-off above 800 Hz). The signal-to-noise ratio (S/N) was defined as the difference, in decibels, between the RMS levels of the whole speech token and the noise. The speech signal was mixed with the noise at S/N levels of 24 dB, 18 dB, 12 dB, 6 dB, 0 dB, -6 dB, -12 dB, for a total of 8 conditions in addition to the original speech.

The spectrally degraded speech stimuli were implemented as follows. The unprocessed speech with the desired S/N level was first pre-emphasized using a first-order Butterworth high-pass filter with a cutoff frequency of 1200 Hz, and then band-pass filtered into 16 frequency bands using eighth-order Butterworth filters. The corner frequencies (3 dB down) of the bands were at 300, 379, 473, 583, 713, 866, 1046, 1259, 1509, 1804, 2152, 2561, 3043, 3612, 4281, 5070, and 6000 Hz. The envelope in each band was extracted by half-wave rectification and low-pass filtering (eighth-order Butterworth) with a 160-Hz cutoff frequency. The envelope was then distorted by a power-law transformation, applied to envelope amplitudes between the maximum envelope value and the noise floor. The exponent of the power function varied from a strongly compressive (p=0.3) to a strongly expanded value (p=3.0). This distorted envelope of each band was used to modulate a wideband noise, which was then spectrally limited by the same bandpass filter used for that analysis band. The output from all bands were then summed, tokens were equated in rms energy, and presented to the listeners diotically through Sennheiser HDA200 headphones at 70 dBA.

3. RESULTS

Figure 1 shows the mean scores of vowel and consonant recognition as a function of the number of training blocks for the extremely compressed and expanded conditions.

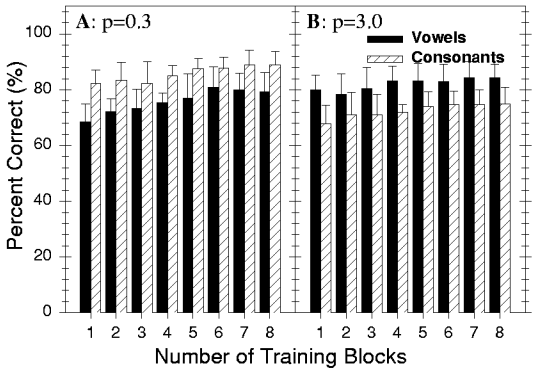

Figure 1. Percent correct of vowel and consonant recognition as a function of the number of training blocks. (A) p=0.3; (B) p=3.0. Error bars represent +/- one standard deviation.

For the compressed condition (p=0.3), the vowel score increased from 68.4% to 80.8%, and consonant scores increased from 82.2% to 88.9% over the eight training sessions, but these increases were not significant [$F(7,32)=2.20$, $p=0.06$ for vowels; $F(7,32)=0.62$, $p=0.74$ for consonants]. However, there was a significant interaction between training and subjects, reflecting a large increase with training for some subjects and no change with training for others [$F(4,35)=8.31$, $p<0.001$ for vowels; $F(4,35)=2.82$, $p=0.04$ for consonants]. For the expanded condition, a 4.4% and 7.0% improvement was observed in vowel and consonant recognition, respectively, but these differences were also not significant [$F(7,32)=1.49$, $p=0.21$ for vowels; $F(7,32)=0.86$, $p=0.55$ for consonants]. Again, there was a significant interaction between subjects and training [$F(4,35)=21.34$, $p<0.001$ for vowels; $F(4,35)=6.39$, $p<0.001$ for consonants].

Figure 2 shows the mean vowel and consonant recognition scores as a function of the exponent of the power function in quiet and noise condition. In the quiet condition (filled circles), both vowel and consonant scores decreased slightly when either a compressed or expanded mapping was applied. Vowels were relatively more tolerant to expansion while consonants were more tolerant to compression. There was a significant effect of amplitude mapping on recognition of vowels [$F(4,20)=11.49$, $p<0.001$] and consonants [$F(4,20)=23.67$, $p<0.001$]. Post-hoc tests according to Tukey HSD multiple comparisons showed that only the extreme compression (p=0.3) significantly reduced the performance in vowel recognition relative to linear mapping (p=1.0). Consonant recognition deteriorated significantly in all mapping conditions except the moderate compression (p=0.5). In noise conditions, amplitude mapping had a significant impact on vowel and consonant recognition at all signal-to-noise levels.

Post-hoc Tukey HSD tests showed no significant performance drop in conditions with expansive mapping relative to linear mapping. Indeed, the extreme expansion (p=3.0) significantly improved the vowel recognition scores at high noise levels (-6 dB SNR). Post-hoc Tukey HSD tests also showed a significant performance drop in all conditions at all noise levels with compressive mappings relative to those with linear mappings.

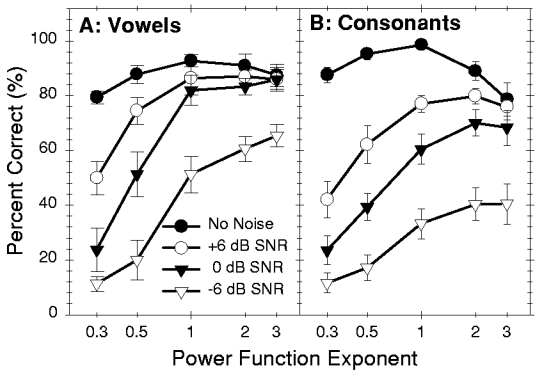

Figure 2. Recognition scores of vowels and consonants as a function of the exponent of the power function in quiet and noise condition. (A) Vowels; (B) Consonants. Error bars represent +/- one standard deviation.

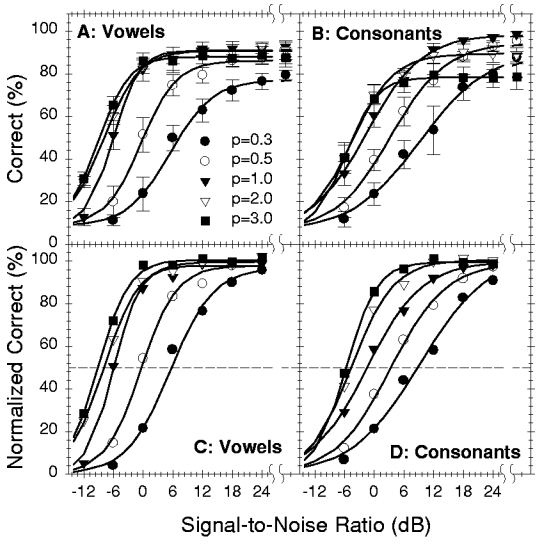

Figure 3. Recognition scores of vowels and consonants as a function of signal-to-noise ratio. (A) Vowel scores; (B) Consonant scores; (C) Normalized vowel scores; (D) Normalized consonant scores. The solid lines represent the fitting curve based on the Equation 1 and experimental data. The dashed lines represent 50% levels. Error bars represent +/- one standard deviation.

Figures 3A and 3B show the mean scores of vowel and consonant recognition, respectively, as a function of S/N ratio with different amplitude mappings. Both vowel and consonant scores gradually decreased as signal-to-noise (S/N) ratio decreased for all mapping conditions. Figures 3C and 3D show the normalized performance on vowels and consonants, respectively, as a function of S/N ratio, relative to the performance in quiet. The dashed lines in Figures 3C and 3D indicate 50% of the normalized score after correction for chance. The S/N level that produced this 50% level of performance was defined as the phoneme recognition threshold (PRT).

The data of Figure 3 were fit by a simple sigmoidal model.

$$\%C = P_0 + (Q - P_0)/(1+\exp(-\beta(x-PRT))) \quad (1)$$

where Q is the percent correct in quiet, PRT is the phoneme recognition threshold in dB, x is the S/N ratio in dB, P_0 is the chance level (6.25% for consonants, 8.33% for vowels), and β is related to the slope of the function at PRT. Figure 4 shows the PRTs and slopes as a function of the power function exponents. The fits of this function to the data were uniformly excellent, with all r^2 values better than 0.99. The PRT for both vowels and consonants improved significantly as the mapping function changed from a compressive mapping to an expanded mapping [$F(4,40)=190.14$, $p<0.001$]. The slopes of the vowel and consonant functions at PRT also changed significantly as a function of the mapping exponent [$F(4,40)=8.03$, $p<0.001$].

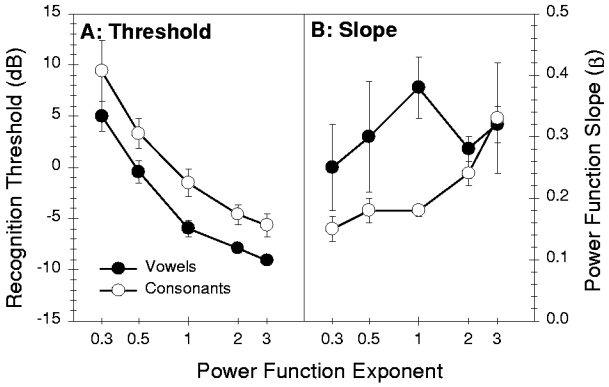

Figure 4. Phoneme recognition threshold and the slope of vowel and consonant recognition as a function of the power function exponents. (A) Phoneme recognition threshold; (B) The slope (β) of the power function. Error bars represent +/- one standard deviation.

4. DISCUSSION

In normal hearing the signal of interest (speech) and the interfering noise are processed through the same sensory channels in the normal cochlea. However, in patients with hearing impairment or in deaf patients with cochlear implants the signal stream must be pre-processed to remediate the impaired sensory processing. Processing strategies need to accommodate not only ideal conditions of listening in quiet, but also real-world conditions of listening in high noise levels. Some modern digital hearing aids and all cochlear implants use instantaneous nonlinear amplitude compression to restore normal loudness sensations to the listener with impaired sensory processing. The present study measured phoneme recognition in quiet and in noise as a function of the nonlinear amplitude mapping. Spectral resolution was reduced to 16 bands to simulate the reduced spectral resolution in an impaired cochlea. Although the results are most directly applicable to signal processing for hearing-

impaired listeners, they also have interesting implications for normal hearing in noisy conditions.

The present results replicate the finding [2, 6] that amplitude nonlinearity in quiet has only a minor effect on phoneme recognition. However, as the S/N level decreases, the effect of amplitude nonlinearity becomes dramatic and asymmetric: expansive mappings are only mildly effected by noise, while compressive mappings are strongly effected. One implication of this result is that expansive mappings may be better overall for mixed quiet and noisy conditions. The expansive exponents may be slightly poorer in quiet conditions, but would still allow reasonably good speech recognition in noise. In contrast, compressive mappings would allow a similar level of speech recognition in quiet, but would be considerably worse in noise. An interesting implication is that a processor with an expansive nonlinearity may improve speech recognition in noise compared to no processing even for normal-hearing subjects.

The results in the present experiment indicated that the improvement by learning was subject-dependent as well as stimulus-dependent. The improvement in the present study was much less than that reported by Licklider and Pollack [6], which may be simply due to the difference in test materials. The improvement of consonant recognition was similar for either compressed or expanded speech. However, more improvement in vowel recognition was observed for the compressed speech than the expanded speech. Some subjects improved more than 20% after 8 training blocks, while other subjects showed no improvement with the same training. The variation of training effects across subjects was unexpectedly large. Possible reasons may include the motivation of subjects or training procedure.

Although amplitude distortion has only a small affect on speech intelligibility in quiet [3, 6], Thomas and Niederjohn [9,10] found that amplitude-compressed speech was recognized at a much higher level than uncompressed speech at high noise levels. This result appears to be contradictory to the results in the present study, which showed a devastating effect of amplitude compression on speech intelligibility in noise. In Thomas and Niederjohn's experiments amplitude compression was applied to the noise-free speech to which uncompressed noise was then added. These earlier methods are applicable where the noise-free speech is available for processing, prior to the introduction of noise. However, their method is not appropriate for most listening situations in everyday life where the speech and noise are added together before the processing can be applied.

The present results also show an interesting difference between vowel and consonant recognition. In the quiet condition, the influence of amplitude compression on vowel and consonant recognition was similar. However, consonant recognition deteriorated much faster than vowel recognition for expanded speech. Further analysis showed that performance on the manner cues suffered most [7]. Amplitude mapping had a similar impact on the recognition pattern of the PRTs for vowels and consonants although the slope of the sigmoidal functions was different.

The data in the present study showed that the PRT was highly dependent on amplitude mapping. Slightly expansive mapping may be better overall in combined quiet and noisy listening conditions. Compressive mapping functions may be satisfactory in quiet, but result in a large decrease in performance in noise. This suggests that at least part of the high variability in cochlear implant users may be due to non-optimal amplitude mapping. Implant listeners who have an amplitude mapping function that is too compressive would be at a disadvantage in noise compared to implant listeners with expansive loudness mappings. The asymmetry of these results suggests that a slightly expansive mapping might be the best choice for overall listening conditions.

5. SUMMARY AND CONCLUSIONS

Nonlinear amplitude mapping produced only a mild decrement in speech recognition in quiet, but could produce a large decrement in noise. Expansive nonlinear mapping provides better overall performance in noise than linear or compressive mapping.

6. ACKNOWLEDGMENTS

We wish to thank Professor James Hillenbrand for allowing us to use the multi-talker vowel test materials. The assistance of Xiaosong Wang in collecting data was also gratefully acknowledged. The research was supported by NIDCD.

7. REFERENCES

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

[2] Fu, Q.-J., and Shannon, R.V. (1998). "Effects of amplitude nonlinearity on speech recognition by cochlear implant users and normal-hearing listeners," J. Acoust. Soc. Am., 104(4).

[3] Fu, Q.-J., Shannon, R.V., Wang, X. (1998). "Effects of noise and spectral resolution on vowel and consonant recognition: Acoustic and electric hearing," J. Acoust. Soc. Am., submitted.

[4] Hillenbrand J., Getty, L.A., Clark, M.J., and Wheeler K. (1995). "Acoustic characteristics of American English vowels," J. Acoust. Soc. Am. 97, 3099-3111.

[5] Hochberg, I., Boothroyd, A., Weiss, M., Hellman, S. (1992). "Effects of noise and noise suppression on speech perception by cochlear implant users," Ear and Hearing 13, 263-271.

[6] Licklider, J.C.R. and I. Pollack (1948). "effects of differentiation, integration, and infinite peak clipping upon the intelligibility of speech," J. Acoust. Soc. Am. 20, 42-51.

[7] Miller, G. and Nicely, P. (1955). "An analysis of perceptual confusions among some English consonants." J. Acoust. Soc. Am. 27, 338-352.

[8] Müller-Deiler, J., Schmidt, B.J., and Rudert, H. (1995). "Effects of noise on speech discrimination in cochlear implant patients," Ann. Otol. Rhinol. Laryngol. 166, 303-306.

[9] Niederjohn, R.J. and J.H. Grotelueschen. "The enhancement of speech intelligibility in high noise levels by high-pass filtering followed by rapid amplitude compression," IEEE Transactions on Acoustic, Speech, and Signal Processing, Vol. 24, pp. 277-282, Aug. 1976.

[10] Thomas, I.B. and R.J. Niederjohn. "The intelligibility of filtered-clipped speech in noise," Journal of the Audio Engineering Society, Vol. 18, pp. 299-303, June, 1970.