ON THE IMPORTANCE OF CONSONANT-VOWEL INTENSITY RATIO IN SPEECH ENHANCEMENT FOR THE HEARING IMPAIRED

T.G. Thomas, P.C. Pandey, and S.D. Agashe
Department of Electrical Engineering
Indian Institute of Technology, Bombay
Powai, Bombay-400 076, India

ABSTRACT

This paper presents the results of a study on the effect of altering the consonant-vowel (C/V) intensity ratio on the perception of English stop consonants. The stimuli were consonant-vowel (CV) syllables synthesized using a modified version of the Klatt synthesizer. These were presented to normal hearing listeners under different signal-to-noise ratios (SNR) simulating hearing impairment. The results indicated a noticeable improvement in identification scores, even in the presence of noise, for the increased C/V intensity ratio stimuli over the unmodified stimuli. This suggests that C/V intensity ratio increment can play an important role in surmounting some of the speech recognition difficulties of impaired listeners.

INTRODUCTION

A promising scheme for enhancing the speech signal to improve intelligibility is based on studies of speaking clearly for the hearing impaired [1]. The assumptions that lay behind these studies were as follows: (1) speakers naturally revise their speech when speaking to impaired listeners; (2) this "clear" speech is more intelligible than "conversational" speech to impaired listeners; (3) clear speech incorporates certain consistent acoustic modifications of the speech signal; and (4) preprocessing speech with these acoustic changes might be expected to improve speech intelligibility for impaired listeners.

Investigations of the phonemic-level acoustic modifications in clear speech compared to conversational speech have shown that the duration of acoustic segments corresponding to consonants (e.g., voice-onset time, transition duration, stop gap) increases and that the C/V intensity ratio for plosives and fricatives increases. Gordon-Salant [2] evaluated the intelligibility of natural speech artificially transformed to "clear" speech by modifying these two parameters. She found that the C/V ratio increment improved consonant recognition in noise for both young and normal-hearing subjects. However, the duration increment did not improve performance for either group. Moreover, the C/V ratio increment improved recognition for most consonant place, manner, and voicing cases, without a substantial increase in consonant confusions. Subsequent studies with natural speech on hearing-impaired subjects [3,4] have yielded similar results.

While the above studies were restricted to natural speech syllables, this study considered the effect on perception of C/V ratio modification using synthesized speech syllables. The C/V ratio modified syllables were synthesized without altering the duration of the acoustic segments. The subjects had normal hearing and hearing-impairment was simulated by presenting the test syllables along with masking broadband noise. Four different C/V ratio modifications and three different signal-to-noise ratios were considered.

EXPERIMENTAL METHOD

Subjects:

Three normal-hearing subjects in the age group from 21 to 40 years were tested.

Stimuli and Apparatus:

The test syllables were CV syllables using three consonants */p, t, k/* with three cardinal vowels */a, i, u/*, thus forming a set of nine CV syllables: */pa/*, */ta/*, */ka/*, */pi/*, */ti/*, */ki/*, */pu/*, */tu/*, and */ku/*. These were synthesized using a modified version of the Klatt synthesizer [5] and stored in data files. For presentation, the data files were played back at a 10 kHz sampling rate using a PC-based data acquisition card, a 7-th order active elliptic filter, and a power amplifier. The filter maintains signal frequency components below 4.6 kHz to within 0.3 dB of no attenuation at all, while components above 5 kHz are attenuated by at least 40 dB. The stimuli were presented to the subjects at a comfortable listening level, under free-field condition.

To simulate hearing impairment, the stimuli were presented along with masking broadband noise at different levels. This noise was obtained using the speech synthesizer program and had a uniform spectrum. Each test syllable was stored under three noise conditions: no noise, 12 dB SNR, and 6 dB SNR. The duration of each stimulus was 300 ms.

Analysis and Processing of Stimuli:

The digitized waveform of the synthesized stimulus was first displayed on the PC monitor using a program written for the purpose. The user could manipulate cursors to segment the beginning and end of the consonant and vowel portions of the syllable. The consonant and vowel segments were identified after repeated visual and auditory monitoring.

The C/V intensity ratio for the syllable was determined by calculating the mean of the squared amplitudes (average power) of the sampled points within the consonant and vowel segments and taking the ratio between them. The ratio was then expressed in dB. Calculation for the vowels was based on only the initial 100 ms. Treating this synthesized syllable as the most "natural" representative, three new versions of this syllable were synthesized by modifying the C/V ratio by -6, $+6$, and $+12$ dB. Thus each CV syllable had four versions with C/V ratio modifications of -6, 0, $+6$, and $+12$ dB. Each stimulus was mixed with noise under three different SNR conditions: no noise, 12 dB SNR, and 6 dB SNR. Thus there were 9 CV syllables under 12 conditions (4 C/V ratio modifications $\times$ 3 SNRs), with a total of 108 test stimuli.

Presentation Procedure:

The stimulus presentation was carried under PC control with the subject seated in front of a terminal, using a computerized test administration program [6]. A three alternative labelling task was used to obtain identification data. The program presented stimuli in random order to each subject. Each experimental run employed a set of three stimuli with all the three consonants in an identical vowel environment, C/V intensity ratio, and SNR. Each run consisted of 5 random presentations of each stimulus for a total of 15 trials per run. The subject was instructed to identify each CV stimulus by pressing the appropriate key. Each subject underwent a total of 36 experimental runs corresponding to the three vowel contexts, three SNRs, and four C/V intensity ratios. The order of presentation of test condition was randomized across subjects.

RESULTS AND DISCUSSION

The results obtained from listening tests with three normal-hearing subjects for the three unvoiced stops /*p, t, k*/ in the CV context with three cardinal vowels /*a, i, u*/ were tabulated as percentage identification scores for different C/V ratio modifications and SNRs. As the variation in scores under different conditions for individual subjects were generally similar, these scores were averaged across the three subjects and are shown in Table 1.

The perception results for the stops in the three vowel contexts, as seen in Table 1, reveal that for the no noise condition near perfect scores were obtained for all the C/V intensity ratio modifications. The only exception was for the stop consonants in the /*u*/ context for the −6 dB C/V intensity ratio modification where the score was 67%. It was observed that under this condition, there was a perceptual confusion between /*pu*/ and /*ku*/.

For the 12 dB SNR condition in all the three vowel contexts, there was a noticeable reduction in scores when C/V intensity ratio was decreased by 6 dB. However, for the +6 and +12 dB C/V intensity ratio modifications, there were improvements in scores, with the maximum improvement being observed for stops in the /*u*/ context.

The results for the 6 dB SNR condition in all the three vowel contexts showed a similar trend as for the 12 dB SNR condition. However, the corresponding scores were somewhat lower because of the greater amount of competing noise.

The scores obtained in the /*u*/ context for both the 6 dB and 12 dB SNR conditions are significantly lower than the corresponding scores in the /*a*/ and /*i*/ contexts. This seems to imply that perception of stops by the hearing impaired, or by normal hearing subjects in the presence of masking broadband noise, is most difficult in the /*u*/ context.

The percentage identification scores of Table 1 were further averaged across the three vowel contexts and a plot of these scores is shown in Figure 1. As seen therein, near perfect scores were obtained for the no noise condition. For the 12 dB SNR condition, the score decreased by 20% when C/V intensity ratio was decreased by 6 dB. However, the score increased by 5% and 16% for C/V ratio modifications of +6 and +12 dB respectively. For the 6 dB SNR condition, the score decreased by 4% when C/V intensity ratio was decreased by 6 dB. The scores increased by 9% and 23% for C/V ratio modifications of +6 and +12 dB respectively.

The results thus indicate that for synthetic stimuli, C/V intensity ratio increment does improve perception of stops under simulated hearing loss condition. The explanation for this improvement is fairly straightforward. Energy of stop segments is weaker than that of vowel segments. By increasing C/V intensity ratio the consonantal portion is amplified with the vowel level kept constant. This reduces backward masking of the consonant by the vowel [7] and also emphasizes the otherwise weak consonant in the presence of competing noise.

REFERENCES

[1] Picheny M.A. et al., *J. Speech Hear. Res.* 26: 96-103, 1983.
[2] Gordon-Salant S., *J. Acoust. Soc. Am.* 80: 1599–1607, 1986.
[3] Gordon-Salant S., *J. Acoust. Soc. Am.* 81: 1199–1202, 1987.
[4] Montgomery A.A. and Edge R.A., *J. Speech Hear. Res.* 31: 386–393, 1988.

[5] Klatt D.H., *J. Acoust. Soc. Am.* 67: 971–995, 1980.
[6] Pandey P.C., *Ph. D. dissertation, Univ. Toronto*, 1987.
[7] Danaher E.M. et al., *Audiology* 17: 324–338, 1978.

Table 1: Percentage identification scores (averaged for three normal hearing subjects) for the stops /*p, t, k*/ under different C/V ratio modifications (CVRM) and SNRs in the context of three vowels /*a, i, u*/.

SNR (dB)	*vowel /a/*				*vowel /i/*				*vowel /u/*			
	CVRM (dB)				*CVRM (dB)*				*CVRM (dB)*			
	−6	0	6	12	−6	0	6	12	−6	0	6	12
No noise	100	100	100	100	100	96	98	89	67	100	96	96
12	59	89	89	93	70	87	93	98	31	42	51	75
6	69	67	64	89	60	73	84	93	33	33	53	60

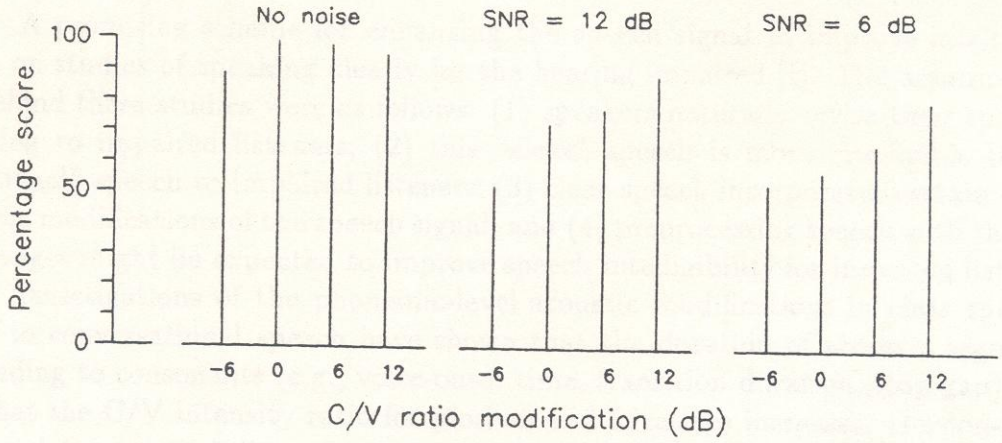

Figure 1: Percentage identification scores of Table 1 averaged across the three vowel contexts.