PERCEPTION STUDIES ON THE ATTRIBUTES OF SYNTHETIC CLEAR SPEECH FOR THE HARD OF HEARING

TG Thomas

Birla Institute of Technology and Science
Pilani-Dubai Campus, Knowledge Village
Dubai,United Arab Emirates
Email:thomas@bitsdubai.com

S Chandra Sekhar

Biomedical Imaging Laboratory
Ecole Polytechnique Fédérale de Lausanne
Lausanne, Switzerland
Email:chandrasekhar.seelamantula@epfl.ch

ABSTRACT

We make a case for 'synthetic clear speech' in the context of the persons with hearing impairment. We study the acoustic attributes of 'clear speech' that enable us to understand their importance in speech perception. Our perception experiments are motivated by the growing body of research emphasizing the significance of clear speech, as opposed to conversation or conventional speech, in the context of the hearing-impaired. Specifically, we study the role of the two important attributes of clear speech namely, the consonant-to-vowel intensity ratio (CVR) and the consonant duration (CD). We have developed a computerized test administration system to study the response of the 'simulated' impaired hearing listeners to modifications in the CVR and CD. Our perception studies using the stop consonants of the English language show that the CVR and the formant transition duration play an important role in speech perception. We believe that our findings will be useful in developing appropriate speech signal processing mechanisms to improve speech perception in the hard of hearing.

1. INTRODUCTION

Speech, as we know it, is a highly complex acoustic signal containing a variety of information, organized at various levels of perception. It is a time-varying signal with its different attributes changing by different amounts, as a function of time. At the signal level, we can associate the temporal and spectral variations as a measure of its time-varying property. At the next level, humans with normal listening perceive speech in terms of acoustic-phonetic units, the knowledge of which is derived by integrating the lower-level temporal and spectral variations over time. The integrated perception of these acoustic-phonetic units, combined with the supra-segmental features, the understanding of which is driven and corrected in a feedback mechanism by the top levels of speech perception, understanding and language, together constitute what we understand of speech in totality. Thus, the speech perception and understanding mechanism is not just a *top-down approach* or a *bottom-up* one but a meaningful blend of both. This mechanism is also continuously evolving, perhaps in a subconscious manner, as we speak, understand each other, and correct ourselves.

While normal speakers and listeners are blessed with the unique ability to communicate by speech, the persons with hearing impairment are less fortunate. We do not understand what constitutes their perception of acoustics, their interpretation of speech and the abstraction of language. Moreover, these also depend on the nature of hearing impairment. Many researchers, given that they are normal listeners, believe that, as normal humans, our abilities to speak and listen grow, first in a bottom-up manner, reinforced gradually by an evolving top-down mechanism as we grow up. Finally, our percept of speech as fully developed individuals is formed by a mix of both mechanisms. It is this belief that leads us to naturally think that if a person is hard of hearing, his sophisticated top-down mechanisms are not well developed as compared to a normal listener. As a result, if we wish to enable him to perceive speech, we must target the bottom-up mechanism. We need to understand the significance of the acoustic-phonetic units and their attributes for perception. These attributes are emphasized in 'clear speech', as opposed to conversational speech.

With the above motivation, we wish to study the role of certain appropriate temporal descriptors of speech at the phoneme level. Specifically, we address the perception of English stop consonants by normal-hearing subjects with *simulated* hearing impairment.

2. A CASE FOR SYNTHETIC CLEAR SPEECH

Perception studies indicate that it is indeed possible to create more intelligible speech for the hearing-impaired by speaking clearly. By speaking clearly, we ensure that the articulators reach their target positions as much as possible and minimize the effects of co-articulation. Clear speech can be defined as the speech used by a person trying to communicate in a noisy environment, or to a foreigner, or a hearing-impaired person. The increased clarity may be obtained by changing the context, sentence structure, vocabulary, speaking rate, stress, pronunciations of individual words and speech sounds, and vocal effort. Acoustic differences have been identified in 'clear' speech which can yield consistent gains in intelligibility for both hearing-impaired listeners and people with normal hearing [1, 2, 3, 4]. Clear speech is also a manifestation of the asymmetry between naturalness of speech and intelligibility. Infact, this asymmetry is not characteristic of clear speech alone. Nusbaum et. al. [5, 6] have reported such

asymmetry in the synthetic speech produced at Motorola.

There are several spectral, temporal and their combinational attributes of clear speech that may be important in understanding how speech is perceived by the hard of hearing. Of these, the temporal properties of speech have received a lot of attention in the recent years. The most recent research in this direction has been the work of Liu and Zeng [7] where the relative contributions of speaking rate, temporal envelope and temporal fine structure, to clear speech perception have been examined. An interesting finding from their 'auditory chimeras' experiment is that the temporal envelope cue contributed more to the clear speech advantage at high signal-to-noise ratios (SNRs), whereas the temporal fine structure cue contributed more at low SNRs.

In this paper, we focus on two specific temporal attributes at the phoneme level namely, the consonant-to-vowel intensity ratio (CVR) and the consonant duration (CD). These are two parameters that are found to increase in clear speech. Subsequent to Picheny's experiments [1, 2, 3], some researchers have evaluated the intelligibility of natural speech artificially transformed to clear speech by altering either one or both of the acoustic parameters [8]. While all the studies reported improvements in recognition for CVR modification, albeit by different amounts, the effect of consonant duration has been indeterminate. This may have been because of the differences in the stimuli, subject background, the effects of signal processing etc. Furthermore, the acoustic segments associated with consonant phonemes are found to increase in a non-uniform manner and such changes cannot be aptly simulated by artificially transforming natural speech to clear speech.

The above studies on the effect of CVR modification and lengthening consonant duration used naturally-uttered-speech syllables as stimuli. Instead, by using synthetic speech material, it would be possible to manipulate the spectral, temporal, and intensity characteristics directly and in an efficient manner. The segmentation of individual phonemes would be easier and it would be possible to alter the various acoustic segments independently. With this motivation, we used a Klatt formant-based speech snythesizer [9]. The Klatt synthesizer uses the all-pole model for the human vocal tract function. It uses 39 parameters that go into determining the output, and, as many as 20 of these can be varied as a function of time. For obtaining the desired synthetic stimuli, we modified the Klatt synthesizer appropriately. We developed software routines for graphically editing the parametric tracks generating the synthetic stimuli. We had to carry out several trial-and-error adjustments to achieve perceptual distinction and good quality of the synthesized speech samples.

In the following sections, we outline the experimental methodology and provide the results of certain experiments performed to evaluate the effects of CVR modification and consonant duration modification on speech perception. We sought the help of subjects with normal hearing and the impairment was simulated. Simulating hearing impairment in normal subjects is perhaps the only way to conduct a controlled test because, as we pointed out before, we do not have a mechanism to retrieve the conceptualization of speech in a hearing impaired listener.

3. EXPERIMENTAL METHODOLOGY

The objective of our experiments is to evaluate the effect of certain types of speech processing on the perception of English stop consonants by normal-hearing subjects with simulated hearing impairment. The task of the subject is to listen to and identify the sounds presented to him at a comfortable listening level over a pair of audiometric headphones. Due to the repetitive nature of the experiments, we developed a computerized test administration system in order to automate the process.

The most common of hearing disabilities is the reduction in the acoustic dynamic range. Typically, this results from a higher threshold of hearing and a lower threshold of pain (see Fletcher-Munson curves). So, if we wish to simulate this kind of hearing impairment in a normal listener, we need to artificially reduce the acoustic dynamic range. The elevated threshold of hearing can be simulated by the addition of broadband noise [10]. The masking noise responsible for the threshold elevation is believed to be predominantly of cochlear origin [11]. In addition to simulating a reduced dynamic range, broadband random noise also approximates more closely, the loudness-growth function of listeners with sensorineural hearing loss [12]. Some researchers have employed multi-talker babble instead of broadband noise [13]. However, due to its non-stationary nature, the effective masking it may provide during syllable stimulus presentation is unpredictable. Hence, we decided to use broadband noise in our experiments.

Each experimental run consists of a number of presentations of the stimuli, in a randomized order with certain uniformity constraints. For each presentation, the response choices are displayed on the subject screen. The subject selects a response by hitting at the appropriate key. At the end of each run, the stimulus-response confusion matrix, the recognition score, and the mean response time are stored. In our preliminary experiments, we sought the help of four subjects with normal hearing. The subjects are in the age group of 21 to 35 years.

3.1. Test stimuli

We used nonsense syllables as stimuli instead of regularly-used words. This is in order to maximize the contribution of the acoustic factors to confusions and to minimize the contribution of linguistic factors. In the CVR modification experiments, our aim is to study the effects of increasing the CVR on consonant recognition in terms of the following: the vowel context, vowel impairment, consonant position in the syllable, and the response times. In addition, we also studied the effect of the CVR on the transmission of information with respect to the vowel context and the features of place and voicing. We chose the English stop consonants /p/, /t/, /k/, /b/, /d/, /g/ in the consonant-vowel (CV) and vowel-consonant (VC) contexts of the vowels /a/, /i/, /u/ as stimuli. The CV syllables used are /pa/, /ta/, /ka/, /pi/, /ti/, /ki/, /pu/, /tu/, /ku/ and the VC syllables are /ap/, /at/, /ak/, /ip/, /it/, /ik/, /up/, /ut/ and /uk/. For each stimulus, the consonant and vowel segments are identified after repeated visual and auditory monitoring. We obtain the intensities of the consonants and the vowels by com-

puting the mean of the squared amplitudes of the samples within the consonant and the vowel segments. The CVR is computed as a ratio of these intensities. We express the CVR on the decibel scale, to be consistent with the logarithmic nature of intensity perception. Treating the synthesized syllables as the most 'natural' representatives, we synthesized four new versions of each syllable by modifying the CVR by +3, +6, +9, and +12 dB. Thus, each syllable has five versions: one with no modification and the other four with CVR modifications of +3, +6, +9, and +12 dB. In our experiments on the CVR modification, we maintained the durations of the consonant and vowel segments constant.

To simulate hearing impairment, we mixed each stimulus with synthesized broadband noise under three SNR conditions: no masking noise, masking noise with 12 dB SNR and 6 dB SNR.

In the consonant duration modification experiments, our aim is to study the effect of increasing the consonant duration on the recognition of stop consonants. The acoustic features contributing to the increased duration in clear speech include the burst, formant transition, and the voice onset time (VOT). We synthesized stimuli in which we increased the duration of each of the acoustic segments by different amounts keeping the duration of the other acoustic segments unaltered.

4. PERFORMANCE MEASURES

Let $\mathcal{X}$ be the set of N stimuli $\{x_1, x_2, x_3, ..., x_N\}$ and $\mathcal{Y}$ be the set of N responses $\{y_1, y_2, y_3, ..., y_N\}$. Let $N(x_i)$, $N(y_j)$, and $N(x_i; y_j)$ denote the frequencies of the stimulus x_i, the response y_j, and the stimulus-response pair (x_i, y_j) respectively, in a sample of N observations. We can estimate the probabilities of these quantities using the notion of relative frequency of occurrence as follows:

$$p(x_i; y_j) = \frac{N(x_i; y_j)}{N} \quad (1)$$

$$p(x_i) = \frac{N(x_i)}{N} = \sum_{j=1}^{N} p(x_i; y_j) \quad (2)$$

$$p(y_j) = \frac{N(y_j)}{N} = \sum_{i=1}^{N} p(x_i; y_j) \quad (3)$$

Speech discrimination test results are usually summarized by the percentage of correct responses for many experimental runs. However, for a more detailed study of the received speech information, the results of each run are usually presented in the form of a stimulus-response confusion matrix, with the rows corresponding to the stimuli and columns corresponding to the responses. The diagonal cell entries $(i = j)$ correspond to the correct responses and the off-diagonal entries $(i \neq j)$ correspond to the confusion errors. The sum of the diagonal entries in a confusion matrix gives the empirical probability of correct responses and is known as the recognition score R_s.

The recognition score is a useful measure of accuracy but it does not provide any information on the distribution of errors. The recognition score also has the disadvantage that it is sensitive to the subject's bias. Therefore, we use

Test	Feature	Percentage relative information transmitted														
		No masking noise CVR (dB)					SNR = 12 dB CVR (dB)					SNR = 6dB CVR (dB)				
		0	3	6	9	12	0	3	6	9	12	0	3	6	9	12
CV	Overall	88	92	89	88	92	59	70	77	77	82	49	57	56	63	73
	Place	72	79	71	69	81	17	32	48	48	62	8	17	23	32	48
	Voicing	99	100	100	100	98	94	97	98	99	97	83	85	80	80	88
VC	Overall	97	100	100	99	99	77	87	88	93	97	66	71	77	83	90
	Place	97	100	100	98	98	55	75	81	88	97	40	52	67	78	96
	Voicing	99	100	100	99	100	95	96	91	96	94	88	84	83	81	81

Table 1. *Table showing the information transmission analysis result for the CVR modification test using the nine syllables in the CV and VC context. We note that the overall information transmitted as well as the transmission of the consonant features increases appreciably with increasing CVR for different SNR.*

the *information transmission analysis* [14] which provides a measure of the covariance between the stimuli and the responses. The information measures of the input stimulus $\mathcal{X}$ and the output response $\mathcal{Y}$ are given in terms of the mean logarithmic probability (MLP) by

$$I_s(\mathcal{X}) = -\sum_i p(x_i) \log_2 (p(x_i)) \text{ bits, and} \quad (4)$$

$$I_r(\mathcal{Y}) = -\sum_j p(y_j) \log_2 (p(y_j)) \text{ bits} \quad (5)$$

An MLP measure of the covariance of stimulus-response is

$$I(\mathcal{X}; \mathcal{Y}) = -\sum_i \sum_j p(x_i, y_j) \log_2 \left(\frac{p(x_i)p(y_j)}{p(x_i, y_j)} \right) \text{ bits} \quad (6)$$

The relative transmission from $\mathcal{X}$ to $\mathcal{Y}$ is given by

$$I_{tr}(\mathcal{X}; \mathcal{Y}) = \frac{I(\mathcal{X}; \mathcal{Y})}{I_s(\mathcal{X})} \quad (7)$$

The above measure of information transmission takes into account the pattern of errors and the score in a probabilistic manner. We can also apply it to the matrices derived from the original confusion matrix by grouping the stimuli in accordance with certain desired features. We can then evaluate the relative importance of these features.

In addition to the recognition scores and the information transmission analysis, we also considered the average response time as another measure of comparing the test stimuli that have been processed differently.

5. RESULTS OF THE PERCEPTION TESTS

We have shown the results of our perception experiments in Table 1 and Table 2. From these results, we note that increasing the CVR improves the recognition scores in the presence of masking broadband noise for normal-hearing subjects. However, we found that the recognition scores for stops in the syllable-final (VC) context were higher than those for stops in the syllable-initial (CV) position for all the CVR modifications. We also subjected the scores to the paired-t test to examine the statistical significance of the results. We did not observe any significant improvements when the CVR was increased in the absence of masking noise. However, for lower SNR conditions, the improvements in scores with increasing CVR were significant.

Feature	Percentage relative information transmitted								
	No masking noise			SNR = 12 dB			SNR = 6dB		
	FTD increase			FTD increase			FTD increase		
	0%	50%	100%	0%	50%	100%	0%	50%	100%
Overall	93	89	89	76	77	79	63	72	70
Place	87	78	77	61	68	73	41	60	58
Voicing	100	98	100	92	78	65	87	76	62

Table 2. *Table showing the information transmission analysis results for the formant transition duration (FTD) modification test for the stops /p/,/t/,/k/,/b/,/d/,/g/ in the CV context of the vowel /a/. We note an increase in the overall information transmitted as FTD is increased to 50%, which then decreases when FTD is increased to 100%.*

The information transmission analysis showed that both overall information transmitted as well as transmission of consonant feature increases appreciably with increasing CVR for all values of the SNR. However, the overall information transmitted as well as the transmission of place feature was seen to be superior in the VC context as compared to those in the CV context.

The results of the duration modification test showed that some subjects may find burst duration modification beneficial for perception at lower SNRs. The information transmission analysis also reveals a general increase in the overall information transmitted as the formant transition duration (FTD) is increased to 50%. When the FTD was increased to 100%, the overall information transmitted was found to decrease.

In the VOT modification test, we found that the performance of all the subjects decreased with increasing VOT for all the three noise cases. The average response time of the subjects decreased (improved) with increase in the burst duration and formant transition duration. However with increasing VOT, the average response time was generally found to increase.

6. CONCLUSION

By means of perception experiments, we have shown that the consonant-to-vowel intensity modification plays an important role in the perception of English stop consonants. The second important parameter is the formant transition duration but its influence is lesser than the former. It would be interesting to carry out similar studies using the other consonants in the language. In order to further establish the effectiveness of this approach, we envisage an analysis-by-synthesis technique in which the sub-segment boundaries are identified and the sub-segments are synthesized again after appropriate modifications. Such an approach would require a processing delay extending over several sub-segments. We know that, in speech processing for cochlear prostheses, delays of up to 120 ms in speech processing and stimulus encoding should not interfere with auditory signal processing in audio-visual comprehension of connected speech [15]. In this case, the segmentation, feature classification, and speech enhancement processes should be completed within this time.

7. REFERENCES

[1] M.A. Picheny, "Speaking clearly for the hard of hearing", Ph.D. Thesis, 1985, MIT, Cambridge, Massachusetts.

[2] M.A. Picheny, N.I. Durlach, and L.D. Braida, "Speaking clearly for the hard of hearing I: intelligibility differences between clear and conversational speech", *Jl of Speech and Hearing Research* 28:96-103, 1985.

[3] M.A. Picheny, N.I. Durlach, and L.D. Braida, "Speaking clearly for the hard of hearing II: acoustic characterisics of clear and conversational speech, *Jl. of Speech and Hearing Research*, 29:434-446, 1986.

[4] R.M. Uchanski, S.S. Choi, L.D. Braida and C.M. Reed, "Speaking clearly for the hard of hearing IV: further studies on the role of speaking rate", *Jl. of Speech, Language and Hearing Research*, Vol. 39, pp. 494-509, 1996.

[5] Nusbaum, Howard, A.L. Francis, and A.S. Henly, "Measuring the naturalness of synthetic speech," *Intl. Jl. of Speech Tech.* 1:7-19, 1995.

[6] Nusbaum, Howard, A.L. Francis, and T. Luks, "Comparative evaluation of the quality of synthetic speech produced at Motorola", Research report, Spoken Language Research Laboratory, University of Chicago, 1995.

[7] S. Liu and F-G Zeng, "Temporal properties in clear speech perception", *Jl. Acoust. Soc. Am*, Vol. 120, Issue 1, pp. 424-432, 2006.

[8] S. Gordon-Salant, "Recognition of natural and time/intensity altered CVs by young and elderly subjects with normal hearing", *Jl. Acoust. Soc. Am.*, Vol. 80, pp. 1599-1607, 1986.

[9] D.H. Klatt, "Software for a cascade/parallel formant synthesizer", *Jl. Acoust. Soc. Am.*, Vol. 67, pp. 971-975, 1980.

[10] S. DeGennaro, L.D. Braida and N.I. Durlach, "Study of multiband syllabic compression with simulated sensorineural hearing loss", *Jl. Acoust. Soc. Am.*, Vol. 69, pp. S16, 1981.

[11] H. Fletcher, "The perception of sounds by deafened persons", *Jl. Acoust. Soc. Am.*, Vol. 24, pp. 490-497, 1952.

[12] J.P.A. Lochner and J.F. Burger, "Form of the loudness function in the presence of masking noise", *Jl. Acoust. Soc. Am.*, Vol. 33, pp. 1705-1707, 1961.

[13] A.A. Montgomery and R.A. Edge, "Evaluation of two speech enhancement techniques to improve intelligibility for hearing-impaired adults", *Jl. of Speech and Hearing Research*, Vol. 31, pp 386-393, 1988.

[14] G.A. Miller and P.E. Nicely, "An analysis of perceptual confusions among some English consonants", *Jl. Acoust. Soc. Am.*, Vol. 27, pp. 338-352, 1955.

[15] P.C. Pandey, "Speech processing for cochlear prostheses", Ph.D. Thesis, University of Toronto, 1987.