19th INTERNATIONAL CONGRESS ON ACOUSTICS
MADRID, 2-7 SEPTEMBER 2007

TIME-SCALING OF CONSONANT-VOWEL TRANSITIONS USING HARMONIC-PLUS-NOISE MODEL FOR IMPROVING SPEECH PERCEPTION BY LISTENERS WITH MODERATE SENSORINEURAL IMPAIRMENT

PACS: 43.71.Ky

Jayan, A. R.; Pandey, Prem C.; Lehana, Parveen K.
EE Dept, Indian Institute of Technology Bombay, Powai Mumbai 400 076, India
{arjayan, pcpandey, lehana}@ee.iitb.ac.in

ABSTRACT
Speech produced to improve intelligibility in a difficult communication environment or for hearing impaired listeners is called clear speech. It is generally characterized by consonant segments with increased intensity and expanded formant transition duration, burst duration, and voice onset time. These characteristics are likely to reduce the effect of increased masking associated with sensorineural impairment, and may improve speech intelligibility. Our objective is to develop a technique for automated conversion of conversational speech to clear speech, by time-scale modification and intensity enhancement of consonant segments, using an analysis-synthesis platform based on harmonic plus noise model (HNM). The transition segments are automatically located by a landmark detection algorithm. Consonant-vowel and vowel-consonant transition segments are time-expanded and the steady state vowel segments are appropriately compressed to maintain the original speaking rate. Clearly intelligible speech could be produced for time-scaling factors in the range 1 to 2. The method was evaluated using VCV syllables on normal hearing subjects with simulated hearing loss. This technique can also be combined with other speech processing techniques for hearing aids, like multi-band compression, spectral modification, and binaural dichotic presentation.

INTRODUCTION
Sensorineural hearing loss is characterized by high frequency hearing loss with elevated hearing thresholds, abnormal growth of loudness with signal intensity, reduced spectral and temporal resolution, and increased spectral and temporal masking. The distortions associated with this impairment adversely affect speech perception, particularly in the presence of background noise [1], [2]. Consonant identification is quite important for speech intelligibility and it gets badly affected as the cues for identification are in transition segments which generally have low intensity and short duration. Compression amplification and frequency lowering techniques have been investigated to map the signal to the audible range of hearing impaired listeners. Dichotic presentation of signal with spectral, temporal, or combined splitting is reported to be effective for improving speech perception by reducing the effect of masking [2] - [5]. But these techniques are helpful only for persons wearing bilateral hearing aids.

Speech signal is highly redundant in terms of temporal and spectral cues required for correct phoneme identification. Even in the presence of noise and distortions, speech may remain intelligible for normal hearing listeners due to their ability to make effective use of the available cues. Hearing impaired listeners lack this ability and therefore ‘landmark’ regions in speech containing high concentration of acoustic cues are very important for them for speech perception. Identification of landmark regions and enhancement of acoustic cues to preserve them even with the distortions associated with the impairment may improve speech intelligibility for hearing impaired listeners [3].

Speech produced by a talker with an intention to improve intelligibility in a difficult communi-cation environment, such as when talking to a hearing impaired person is called ‘clear speech’. Incorporating acoustic properties of clear speech in conversational speech is reported to be effective in improving speech intelligibility for normal and impaired listeners in quiet and noisy environments. Picheny [6] reported clear speech to be about 17% more intelligible than

conversational speech, by conducting listening tests on subjects with sensorineural loss, using nonsense sentences as test material. Clear speech is generally characterized by reduced speaking rate with more frequent and lengthy pauses, non-uniformly expanded consonant segments with increased intensity, drift towards higher fundamental frequency, and increased temporal envelope modulations [7], [8]. Several investigations have been carried out to identify the features of clear speech responsible for its enhanced intelligibility, and to incorporate these features in conversational speech.

Consonant intensity relative to the neighboring vowel intensity is normally expressed as consonant vowel intensity ratio (CVR). CVR is found to be higher in clear speech than in conversational speech. Gordon-Salant [9] reported consistent improvement in recognition of CV syllables by normal and hearing impaired subjects, by 10 dB CVR enhancement and 100% expansion of consonant duration. Kennedy *et al.* [10] reported that different CVR enhancements were needed for maximizing recognition of different consonants, values generally being in the range of 8-15 dB. Hazan and Simpson [11] reported significant improvements in the identification of stops, fricatives, and nasals, by intensity modification of frication segments by 6 dB and burst segments by 12 dB. Thomas *et al*. [12], [13] experimentally evaluated improvement in speech perception by CVR and duration enhancement using synthetic syllables. CVR enhancement was found to be more effective in reducing the effect of forward masking in vowel-consonant (VC) context than in consonant-vowel (CV) context. CVR modification up to 10 dB improved identification of stop consonants without affecting vowel perception. Expansion of formant transition duration and burst duration by 50% improved consonant identification at lower SNR levels, whereas voice onset time modification resulted in degraded performance.

In all the above investigations, enhancements were performed on manually annotated speech material. The difficulties in automatically locating the landmark regions and in modifying them without introducing artifacts are the main limiting factors in intelligibility enhancement based on properties of clear speech [11]. Automated intelligibility enhancement without pre-annotated speech material can be carried out by modification of the signal in regions where it displays certain peculiar characteristics like fast spectral transitions. In an intelligibility enhancement technique reported by Colotte and Laprie [14], stop bursts and fricatives were selectively slowed down. A spectral variation function, based on mel-cepstral analysis, was used to locate the regions for enhancement, and time-scaling was performed using TD-PSOLA. Skowronski and Harris [15] reported an automated energy redistribution technique for intelligibility enhancement using voiced-unvoiced information obtained from a measure of spectral flatness. This paper presents an automated method for modifying speech during VC and CV transition segments in temporal and intensity dimensions using an analysis-modification-resynthesis system based on harmonic plus noise model (HNM) [16] - [18]. Evaluation has been carried out using vowel-consonant-vowel (VCV) syllables with vowel /a/ and stop consonants /p, b, t, d, k, g/, with listening tests involving normal hearing subjects with simulated hearing loss.

AUTOMATED MODIFICATION OF SPECTRAL TRANSITION SEGMENTS

Speech signal is digitized with sampling rate of 10 kHz and 16-bit quantization. The processing involves a landmark detection stage for locating the boundaries of segments for enhancement, an analysis-modification-resynthesis stage based on HNM, and an intensity modification stage. In the analysis-modification-resynthesis stage, the VC and CV transition segments are time-expanded, and steady state vowel segments are appropriately compressed to maintain the overall speech duration unaltered.

HNM based analysis/modification/resynthesis

In HNM, harmonic part and noise part are modeled separately and it allows time-scale modification of speech by modification of a small parameter set. Even for large time-scaling factors, the synthesized speech sounds natural, without tonal artifacts [16]. The block diagram representation of HNM analysis stage is shown in Fig. 1. Fundamental frequency F_0 is estimated every 10 ms using a normalized spectral cross-correlation function [17]. Speech segments are classified as voiced or unvoiced (V/UV) based on their harmonic structure. Analysis time instants are located pitch synchronously during voiced segments and with a constant separation of 10 ms during unvoiced segments. Voiced segments are modeled by both harmonic part and noise part, whereas unvoiced segments are simulated by noise part alone.

Parameters are estimated for each frame i extending from t_a^{i-1} to t_a^{i+1} and centered at t_a^i. Maximum voiced frequency F_m, separating the harmonic part from the noise part is located by conducting a harmonic test at each prominent peak in the magnitude spectrum. Amplitudes and phases (a, ϕ) of harmonics of F_0 up to F_m are obtained by a least-squares minimization technique. Harmonic part $s_h(n)$ is synthesized by summation of the harmonics with estimated amplitudes and phases. Noise part $s_n(n)$ is obtained by subtracting the synthesized harmonic part $s_h(n)$ from the signal $s(n)$. For both voiced and unvoiced segments, noise part is modeled with its spectral structure represented by LPC coefficients and temporal structure by its energy envelope [17], [18].

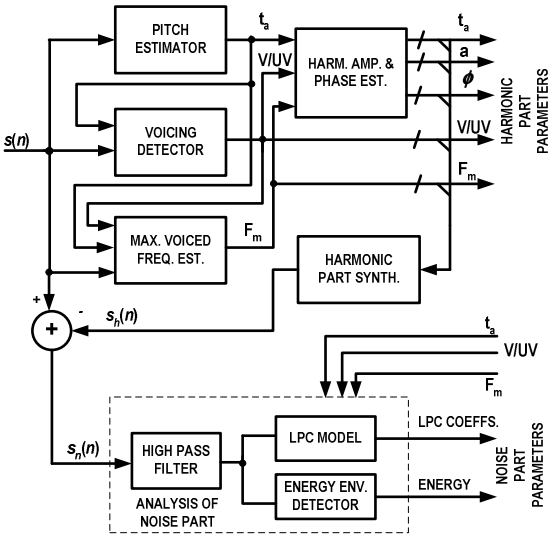

Figure 1.- HNM analysis stage

Time-scale modification is performed using a time-warping function, specified by an array of scaling factors (β), mapping the analysis time instants to synthesis time instants, maintaining the original pitch contour. HNM parameters in the time-warped scale are used for synthesizing the time-scale modified speech, as shown in Fig. 2. The harmonic part is obtained by overlap-adding a stream of short-time signals with estimated amplitudes and phases in a pitch-synchronous manner.

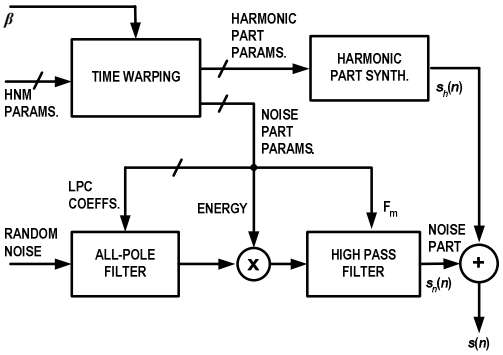

Figure 2.- Time-scaling and synthesis stage

The noise part is synthesized by filtering unit-variance Gaussian noise through a time-varying filter, formed by the LPC coefficients. The time-domain energy envelope function is applied to the synthesized noise part to make its temporal structure same as that of the original noise part. For voiced segments, frequency components below F_m are removed from the noise part using a high pass filter. The resynthesized speech is produced by addition of harmonic part with the noise part.

Locating boundaries of spectral transition segments

The method used for locating spectral transition segments is a modified form of the landmark detection algorithm reported by Liu [19]. Short-time magnitude spectra are computed using 512-point FFT on 6 ms segments with Hanning window, with 1ms shifts. The spectrum is divided into five non-overlapping bands: 0–0.4, 0.4–1.2, 1.2–2.0, 2.0–3.5, 3.5–5.0 kHz. This method is based on detecting combined variation of peak energy and centroid frequencies in these five bands. Any significant spectral transition results in a noticeable change in peak energy and centroid frequency in at least one band.

Centroid frequency of a spectral band is calculated as

$$f_c = \left(\sum_{k=k_1}^{k_2} k \left| X_k \right|^2 \Big/ \sum_{k=k_1}^{k_2} \left| X_k \right|^2 \right) \left(f_s / N \right) \quad \text{(Eq. 1)}$$

where $|X_k|$ is the magnitude of FFT component with frequency index k, and k_1 and k_2 are the lower and upper frequency indices for the band, f_s is the sampling frequency, and N is the number of points in FFT computation. Peak energy in a band is computed in dB scale as

$$E_p = 10 \log_{10} \left(\max \left[\left| X_k \right|^2 , k_1 \le k \le k_2 \right] \right)$$ (Eq. 2)

Rate-of-rise (ROR) of E_p and f_c contours are obtained by taking mean of their first difference in every 1 ms, using a 20 ms window, placed in an overlapping fashion [19]. For VC or CV transitions, ROR contours show peaks during the transition periods. To locate the simultaneous variation of energy and its frequency distribution in a band, the absolute ROR values of E_p and f_c are multiplied and normalized. The product of ROR values in the five bands are summed to get a transition index indicating the spectral transitions. This index has positive value with low amplitude variations in the vowel segments and prominent peaks during the spectral transitions for plosives. Transition segment boundaries are located by comparing this index with an empirically selected threshold.

Figure 3(a) shows the speech waveform for syllable /aba/ and E_p and f_c contours in the five bands. Normalized ROR contours of E_p, f_c, and their products are shown in Fig. 3(b). Transition index and the located boundaries of transition segments are shown in Fig. 3(c), with transition start and end points marked as negative and positive impulses.

The automated landmark detection algorithm was evaluated by applying it on manually annotated sentence materials from TIMIT database [20]. The points of maximum spectral transitions were found to closely match with the manually located segment boundaries, as shown for a sample sentence in Fig. 4.

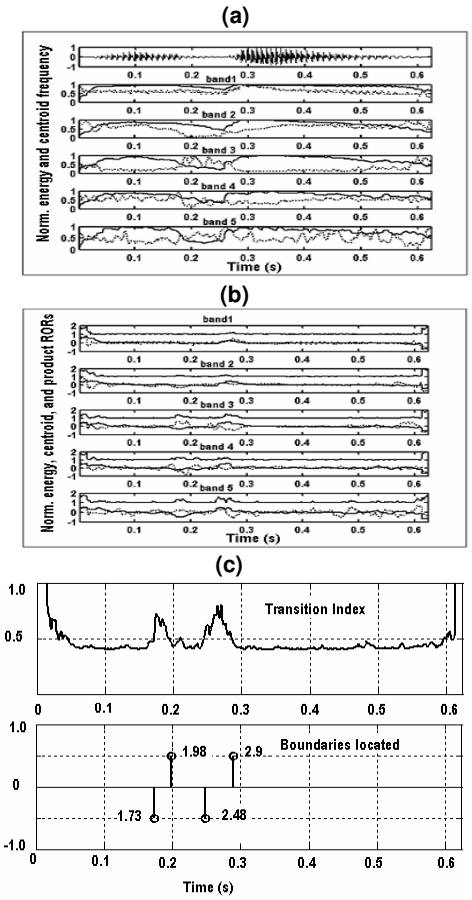

Figure 3.- Transition detection: (a) Speech signal for syllable /aba/ with contours of E_p (solid) and f_c (dotted), (b) RORs of E_p (soild), f_c (dotted), and product (top, solid), (c) Transition index and detected boundaries

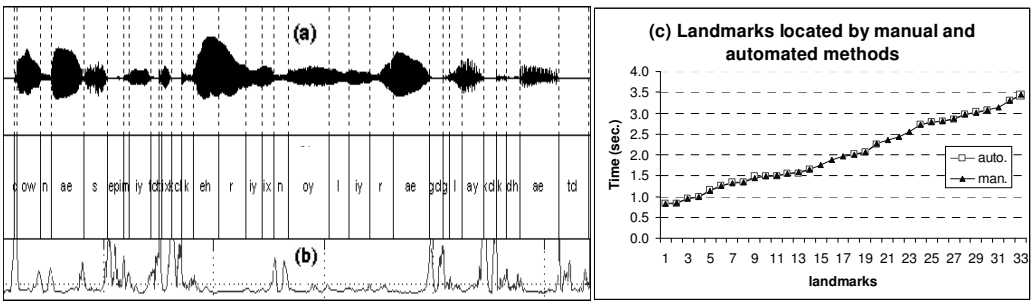

Figure 4.- Automated landmark detection: (a) Speech signal for the sentence *"don't ask me to carry'n oily rag like that"* from TIMIT database and its manual transcription, (b) transition index, and (c) matching of manual and automatically located landmarks

Intensity enhancement of consonant segments

Speech segment starting from the onset of VC transition to the end of CV transition, as located by the automated transition boundary detection, was selected for intensity enhancement. Signal waveform obtained from HNM based resynthesis was enhanced by 3, 6, and 9 dB by scaling the waveform with appropriate gain factors. Gain factors were given a trapezoidal envelope with rise

and fall times of 10 ms to eliminate occurrence of audible clicks due to sudden change in amplitude during intensity enhancement. Fig. 5 shows the spectrograms of the original, time-scaled, and time and intensity scaled VCV syllable /aba/.

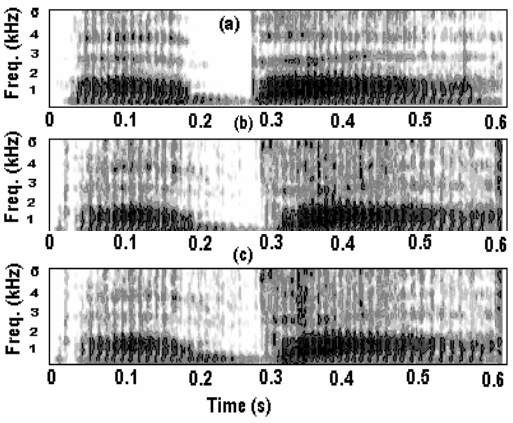

Figure 5.- Spectrograms: (a) unprocessed syllable /aba/ (b) time-scaled, and (c) time and intensity scaled for $\beta = 1.5$ and CVR enhancement = 6 dB

EXPERIMENTAL EVALUATION

Test material consisted of VCV syllables with stop consonants /p, b, t, d, k, g/ in the context of vowel /a/. Clearly intelligible speech was produced with time-scaling factors of 1 to 2. Hence time-scaling was carried out for five factors of 1.0, 1.2, 1.5, 1.8, and 2.0. Based on informal listening tests with CVR enhancement of 3, 6, and 9 dB, we selected enhancement by 6 dB. There were a total of 12 processing conditions for each VCV syllable: unprocessed (up), enhanced CVR without time-scaling (eup), time-scaled with scaling factor x (ts-x), and time scaled with enhanced CVR (ets-x).

To test intelligibility enhancement for different scaling factors, listening tests were conducted on normal hearing subjects with simulated hearing loss. Hearing impairment was simulated by adding broadband noise to the signals at 6 different SNR levels (∞, 0, -3, -6, -9, and -12 dB). This resulted in a total of 72 test conditions (12 processing conditions $\times$ 6 SNRs). Tests were conducted in a sound proof room using a computerized setup for presenting stimuli binaurally through headphones. The subjects were asked to click on one out of the six possible syllables displayed on the computer screen. In each test, 60 presentations were made, each stimulus randomly presented 10 times, with the number of consecutive presentations of the same stimulus limited to a maximum of 3. The order in which tests were conducted was randomized, with a total of 5 tests for each condition.

RESULTS AND DISCUSSION

Fig. 6 shows the percentage recognition scores for the 12 processing conditions, at different SNR levels for one subject. In case of time-scaling, recognition scores improved with the effect being more visible at lower SNR levels. Scaling factor of 1.5 was found to be most effective. The scores for CVR enhanced stimuli were comparable to the scores for time-scaling followed by CVR enhancement, indicating the importance of CVR enhancement in consonant identification.

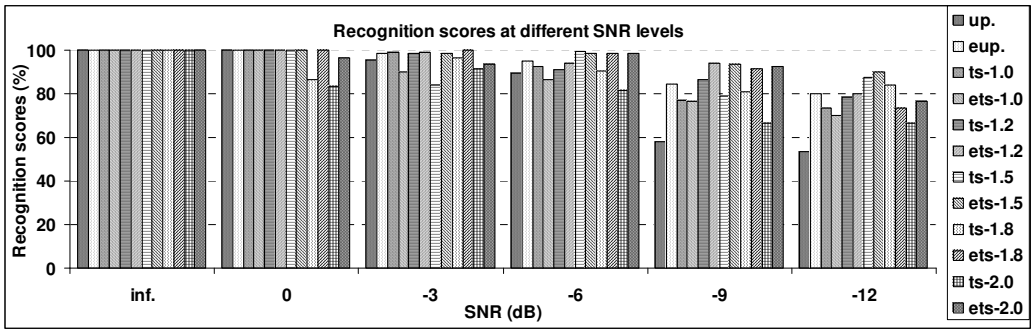

Figure 6.- Percentage recognition scores for different processing conditions

The six consonants in the syllables used as test stimuli can be fully distinguished on the basis of voicing and place features. Information transmission analysis [21] was carried out on the stimulus-response confusion matrices to get the relative information transmitted in terms of

overall, voicing, and place features. Loss in information at lower SNR levels could be fully accounted by the place feature, as the relative information transmission for voicing feature remained nearly 100%. A consistent improvement in relative information transmission was observed for the processing involving time-scaling and intensity enhancement, indicating its effectiveness in reducing the effect of masking.

CONCLUSIONS

Time-scaling of transition segments in the range 1.0-1.5 was found to give improvement in stop consonant identification particularly at lower SNR levels. Combining time-scaling with intensity enhancement resulted in improvement in scores, at all time-scaling factors. The improvements were primarily for the place feature, because voicing was not affected by lower SNR levels. The effectiveness of this technique needs to be evaluated on sentences and other test material involving variable vowel contexts and co-articulation effects.

The technique of automated transition boundary detection and HNM based analysis-synthesis can be used for combining the time-scaling and CVR enhancement with other speech processing techniques for hearing aids, like multi-band compression, spectral modification, and binaural dichotic presentation.

References

[1] B. C. J. Moore: An Introduction to the Psychology of Hearing. 4th ed. London: Academic, 1997

[2] CHABA: Speech perception aids for hearing-impaired people: Current status and needed research, Journal of the Acoustical Society of America **90, No.2** (1991) 637-678

[3] A. N. Cheeran, P. C. Pandey: Evaluation of speech processing schemes using binaural dichotic presentation to reduce the effect of masking in hearing impaired listeners. Proceedings of the 18th International Congress on Acoustics (ICA2004, Kyoto, Japan), II-1523-1526

[4] T. Arai, K. Yasu, Hodoshima: Effective speech processing for various hearing impaired listeners. Proceedings of the 18th International Congress on Acoustics (ICA 2004, Kyoto, Japan), (2004) II-1389-1392

[5] D. S. Choudhury, P. C. Pandey: Dichotic presentation of speech signal with critical band filtering for improving speech perception. Proceedings of IEEE/ICASSP 1998 (Seattle, Washington), 3601-3604

[6] M. A. Picheny, N. I. Dulrach, L. D. Braida: Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech. Journal of Speech and Hearing Research **28** (1985) 96-103

[7] M. A. Picheny, N. I. Dullach, L. D. Braida: Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech. Journal of Speech and Hearing Research **29** (1986) 434-446

[8] J. C. Krause, L. D. Braida: Acoustic properties of naturally produced clear speech at normal speaking rates. Journal of the Acoustical Society of America **115, No.1** (2004) 362-378

[9] S. Gordon-Salant: Effects of acoustic modification on consonant recognition by elderly hearing-impaired subjects. Journal of the Acoustical Society of America **81, No.4** (1987) 1199-1202

[10] E. Kennedy, H. Levitt, A. C. Neuman, M. Weiss: Consonant-vowel intensity ratios for maximizing consonant recognition by hearing-impaired listeners. Journal of the Acoustical Society of America **103, No. 2**, (1997) 1098-1114

[11] V. Hazan, A. Simpson: The effect of cue-enhancement on the intelligibility of nonsense word and sentence materials presented in noise. Speech Communication **24, No.3** (1998) 211-226

[12] T. G. Thomas: Experimental evaluation of improvement in speech perception with consonantal intensity and duration modification. Ph.D. dissertation (1996) Department of Electrical Engineering, IIT Bombay, India

[13] T. G. Thomas, P. C. Pandey, S. D. Agashe: On the importance of consonant-vowel intensity ratio in speech enhancement for the hearing impaired. Proceedings of the International Conference on Biomedical Engineering, Hong Kong (1994) 181-184

[14] Colotte, V., Laprie, Y.: Automatic enhancement of speech intelligibility. Proceedings of IEEE/ICASSP 2000, (Istanbul, Turkey) 1057–1060

[15] Skowronski, M. D., Harris, J. G: Applied principles of clear and Lombard speech for automated intelligibility enhancement in noisy environments. Speech Communication **48, No.5** (2006) 549-558

[16] J. Laroche, Y. Stylianou, E. Moulines: HNS: Speech modification based on a harmonic + noise model. Proc. IEEE International Conference on Acoustic, Speech, and Signal Processing, Minneapolis, USA (1993) 550-553

[17] Y. Stylianou: Modeling speech based on harmonic plus noise models. G. Chollet *et al*. (Eds.) Nonlinear Speech Modeling, Berlin: Springer-Verlag, (2005) 244-260

[18] P. K. Lehana, P. C. Pandey: Harmonic plus noise model based speech synthesis in Hindi and pitch modification. Proceedings of the 18th International Congress on Acoustics (ICA 2004, Kyoto, Japan) 3333-3336

[19] S. A. Liu: Landmark detection for distinctive feature based speech recognition. Journal of the Acoustical Society of America **100, No.5** (1996) 3417-3430

[20] Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., and Dahlgren, N. L.: DARPA TIMIT acoustic-phonetic continuous speech corpus. U.S. Dept. of Commerce, NIST, Gaithersburg, MD, 1993.

[21] G. E. Miller, P. E. Niceley: An analysis of perceptual confusions among some English consonants. Journal of the Acoustical Society of America **27**, **No.2** (1955) 338-352