IEEE-International Conference on Signal Processing, Communication and Networking.
Madras Institute of Technology, Anna University Chennai India, Jan 4-6, 2008. pp. 69-74

Automated Detection of Transition Segments for Intensity and Time-Scale Modification for Speech Intelligibility Enhancement

A. R. Jayan, P. C. Pandey, and P. K. Lehana
SPI Lab, Department of Electrical Engineering
Indian Institute of Technology Bombay
Powai Mumbai 400 076, India
{arjayan, pcpandey, lehana}@ee.iitb.ac.in

***Abstract*— Spectral transition segments serve as landmarks for the perception of consonants. In "clear speech" mode adopted by speakers to improve intelligibility in difficult communication environments, transition segments are of increased duration and intensity. Modification of conversational speech to have acoustic properties of clear speech has been reported to improve its intelligibility. This paper presents an automated method for locating spectral transition segments in speech, and to produce natural quality resynthesized speech with intensity and time-scale modified spectral transition segments. The boundaries of spectral transition segments are located using an index derived from the rate of variation of energy and centroid frequency in five non-overlapping spectral bands. Time-scale modification is performed using harmonic plus noise model (HNM) based analysis-synthesis. The overall speech duration is kept unaltered by appropriately compressing the steady state segments. Transition segments are intensity scaled by 6 dB. The effectiveness of the method was evaluated by conducting listening tests on normal hearing subjects using VCV syllables as the test material.**

***Index Terms*- Transition segment detection, Clear speech, CVR modification, Time-scale modification, Harmonic plus noise model.**

I. INTRODUCTION

The important regions in speech containing information for the correct phoneme identification are called 'landmarks,' and these are generally characterized by spectral transitions. This may be due to the ability of the human auditory system to predict spectral targets based on the transitional information [1]. It is possible to make speech more robust and intelligible by enhancing these regions by natural or synthetic methods. In "clear speech" mode, a talker attempts to make speech more intelligible (during communication with a hearing impaired listener, in a noisy environment, etc.). Clear speech is reported to be about 17 % more intelligible than conversational speech of the same speaker [2], [3].

Acoustic characteristics of clear and conversational speech show differences at global, phonological, and phonetic levels. Speaking rate in clear speech mode is nearly half of that in conversational mode. This reduction is mainly due to more frequent and lengthy pauses, and increased duration of acoustic segments. At the phonetic level, consonant segments are of increased duration and intensity. The durations of transition segments, which include voice onset time (VOT), formant transition duration, burst duration, and stop closure duration, are comparatively higher in clear speech than corresponding durations in the conversational speech. Consonant vowel intensity ratio (CVR), the ratio of consonant intensity relative to the neighboring vowel intensity is found to be higher in clear speech than in conversational speech [4], [5]. Krause and Braida [5] reported that trained speakers could produce highly intelligible speech at their normal speaking rates. Certain inherent characteristics like increased spectral energy in the 1-3 kHz region, intensity envelope with higher modulation index, changes in VOT and frequency of stop burst releases were identified as the contributors for improved intelligibility. Liu and Zeng [6] reported that at lower SNRs temporal fine structure (0.5-10 kHz) contributed more towards intelligibility, whereas at quiet and positive SNRs, temporal envelope (2-50 Hz) was more important.

Gordon-Salant [7] reported 16 % improvement in recognition of consonants by normal and hearing impaired listeners using intensity and duration altered consonant-vowel (CV) syllables as the test material. Consonant intensity was enhanced by 10 dB and duration was uniformly increased by 100 %. Hazan and Simpson [8] reported intelligibility enhancement strategies using vowel-consonant-vowel (VCV) syllables and semantically unpredictable sentences as test material. Intensity modification of frication segments (+6 dB) and burst segments (+12 dB) was reported to be effective in improving intelligibility. Thomas and Pandey [9], [10] evaluated improvement in speech perception by CVR modification (3-12 dB) and consonant duration modification (50 %, 100 %) using synthetic syllables. CVR enhancement was found to be more effective in reducing the effect of forward masking in the VC context. CVR enhancement up to 10 dB improved identification of stop consonants. Expansion of formant transition duration and burst duration by 50 % improved consonant identification at lower SNR levels, whereas VOT modification resulted in degraded performance.

In all the above investigations, modifications were performed on manually annotated speech material. Manual methods for segmentation are most accurate, but they are time consuming and tedious, particularly at the phoneme level.

Further, they obviously can not be used in automated intelligibility enhancement techniques based on the properties of clear speech.

Glass and Zue [11] reported a segmentation scheme based on the critical band filtering property of the human auditory system. Responses from the auditory nerve fibers were simulated using outputs from a set of 40 filters processed by a transduction stage. These responses were computed every 5 ms, to get a 40-dimensional feature vector, for each frame. Euclidean distances between frames were used to measure their acoustic similarity. This method detected 90 % of the boundaries with 20 ms accuracy, when tested using 500 sentences from TIMIT database [21].

Sarkar and Sreenivas [12] reported a segmentation method based on average level crossing rate (ALCR), defined as the rate of crossing of certain predefined levels by the normalized speech waveform. ALCR responds to amplitude and frequency variations in the speech signal and has valleys at locations corresponding to points of phoneme transition. The level allocation was based on an adaptive scheme using signal pdf and SNR. This method detected 78.6 % of manually located boundaries with a temporal resolution of 20 ms, when tested using 100 sentences from TIMIT database.

Alani and Deriche [13] reported a segmentation technique, capable of tracking fast and slow transitions using wavelet decomposition of the signal. Dyadic wavelet decomposition was used to split speech signal into 6 bands (0-0.25, 0.25-0.5, 0.5-1.0, 1.0-2.0, 2.0-4.0, 4.0-6.0 kHz). Short-time energy variations were computed for each band, using a 256 point Hamming window, with 25 % overlap. The strength of transition was measured by a Euclidean distance function evaluated over four consecutive bands. It was compared with an empirically selected threshold to locate the segment boundaries. This method detected 90.9 % of the manually located landmarks when evaluated using 11 speech files from TIMIT database.

These techniques for automated segmentation aim at detecting boundaries of all classes of phonemes, and are computation intensive. For the application of intelligibility enhancement, we need to detect the consonant transitions for time-scale and intensity modification. Rather than using a segmentation approach, we have used a landmark detection approach in the present investigation.

Liu [14] reported an algorithm for detecting acoustically abrupt landmarks in speech using energy variations in six frequency bands (0-0.4, 0.8-1.5, 1.2-2.0, 2.0-3.5, 3.5-5.0, 5.0-8.0 kHz). The algorithm was capable of locating glottal, sonorant, and burst onsets and offsets. Short-time spectral analysis was carried out and variations in the largest spectral component in each of the six bands were used to form energy contours. Rate-of-rise contours (ROR's) were computed by taking the derivative of energy contours, and peaks in the ROR contours were used to locate the landmarks. A two-pass strategy was used, a coarser pass to locate the vicinity of a spectral change and a finer pass to time-localize the landmarks. The algorithm detected 88 % of the total landmarks with a temporal resolution of 30 ms, when evaluated using manually annotated sentences from TIMIT database.

Automated intelligibility enhancement can be carried out by modification of the signal in regions where it displays certain peculiar characteristics like fast spectral transitions. In many applications, speech is divided into segments with boundaries placed at the time instances of major spectral changes, which correspond to major changes in the vocal tract configuration. These regions correspond approximately to the landmark regions introduced by Liu [14], where acoustic cues are concentrated [15].

Colotte and Laprie [15] reported an intelligibility enhancement technique using a spectral variation function for locating the regions for modification. The spectral variation function detected 82 % of manually located landmarks with an accuracy of 20 ms. Stop bursts and unvoiced fricatives were amplified by 4 dB and time-scale modified by factors in the range of 1.8 to 2.0. Listening tests were conducted on normal hearing listeners and they were asked to complete missing words in the sentences, with and without enhancement. Results of the listening tests showed significant improvement in missing word identification in the modified sentences.

Skowronski and Harris [16] reported a technique based on boosting of energy in the regions critical to intelligibility. A measure of spectral flatness defined as the ratio of geometric mean to arithmetic mean of the spectrum of the speech signal was used for locating the regions for modification. Listening tests were conducted on 25 subjects using isolated utterances of confusable words from 16 speakers. The enhancement improved intelligibility for 9 speakers without degrading intelligibility for the remaining speakers.

This paper presents an automated method for improving speech intelligibility, utilizing the acoustic properties of clear speech. Figure 1 shows block diagram representation of the method. The regions for modification are located by a landmark detector, which detects boundaries of transition segments. The detected transition segments are time-expanded using a harmonic plus noise model (HNM) based time-scale modification stage [17], [18]. The overall duration is kept unaltered by appropriately compressing the steady state vowel segments [20]. Intensity scaling stage performs intensity medication of transition segments. Evaluation has been carried out using 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.

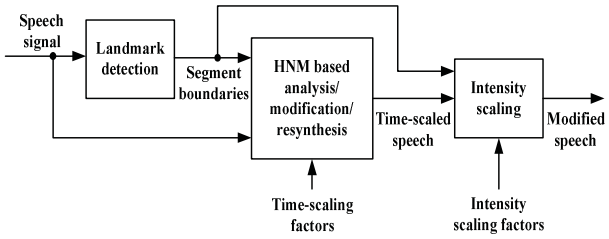

Fig. 1. Automated landmark detection and speech modification system.

II AUTOMATED DETECTION OF TRANSITION SEGMENTS

Landmarks are information rich areas in an utterance and the speech perception process focuses on landmarks to get the acoustic cues necessary for correct phoneme identification. Landmarks are classified into abrupt, non-abrupt and vocalic. An estimate of phonetically balanced sentences in the TIMIT database has been reported to have 68 % acoustically abrupt landmarks, 29 % vocalic landmarks and 3 % non-abrupt landmarks [14]. Acoustically abrupt landmarks are associated with consonants, which involve formation of a tight constriction and a release in the vocal tract, by movement of a primary articulator. For semivowels, narrowing of the constriction and its gradual release, result in non-abrupt landmarks.

In addition to energy variations, our method uses variation of centroid frequencies in the spectral bands, which contain information regarding the formant transitions. By combining the ROR functions of energy and centroid frequency, it is possible to track the energy and formant frequency variations in spectral bands. The combined ROR contours were added to get a single parameter called transition index, as an indication of the overall spectral variation.

A. Transition Segment Detection Technique

Short-time magnitude spectra are computed using 512-point FFT on 6 ms segments selected using Hanning window. The input speech is zero padded before FFT computation to get a spectral resolution of 20 Hz for a sampling rate of 10 k Sa/s. The short analysis window length gives a spectral envelope with the effect of harmonics suppressed, and frames are taken every 1 ms to permit tracking of any abrupt variations in the spectrum. 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. Band 1 primarily monitors glottal vibrations, bands 2 -5 detect closures and releases of consonants, and onsets and offsets of aspiration and frication noise associated with stops, fricatives, and affricates [14]. Our method is based on the assumption that a significant spectral transition results in a noticeable change in energy and centroid frequency in at least one of these bands.

A 20-point moving average is used for getting smoothened spectral components $|X_n(k)|$, from which the peak energy E_p and centroid frequency f_c contours are computed. Peak of log energy values in $|X_n(k)|$ in each band *b* is taken, with *n* spaced every 1ms, to form an energy contour for band *b* for frame *n*, and is given by

$$E_p(b,n) = 10\log_{10}\left(\max\left[\left|X_n(k)\right|^2\right]\right),\ k_1 \le k \le k_2 \quad (1)$$

where k_1 and k_2 are the lower and upper frequency indices for the band *b*. Centroid frequency of a spectral band *b*, for frame *n* is calculated as

$$f_c(b,n) = \left(\sum_{k=k_1}^{k_2} k\left|X_n(k)\right|^2 \Big/ \sum_{k=k_1}^{k_2}\left|X_n(k)\right|^2\right)\left(\frac{f_s}{N}\right) \quad (2)$$

where f_s is the sampling frequency, and N is the number of points in FFT computation.

ROR's of E_p and f_c contours are obtained by taking their first difference, every 1 ms, using a 50 ms time-step. Energy and centroid frequency ROR's for a band *b* and frame *n* are defined by

$$E'_p(b,n) = E_p(b,n+K) - E_p(b,n-K) \quad (3)$$

$$f'_c(b,n) = f_c(b,n+K) - f_c(b,n-K) \quad (4)$$

where K = 25, corresponding to a 50 ms time-step. These ROR functions are normalized to the 0-1 range by shifting and scaling as $E'_{pn}(b,n)$ and $f'_{cn}(b,n)$. To locate the simultaneous variation of energy and frequency in a band, the absolute ROR's $E'_{pn}(b,n)$ and $f'_{cn}(b,n)$ are multiplied, and these product ROR's are averaged across bands to get the transition index

$$T_r(n) = (1/5)\sum_{b=1}^{5} E'_{pn}(b,n) f'_{cn}(b,n) \quad (5)$$

Transition segment boundaries are located by comparing this index with an empirically selected threshold.

B. Results of Transition Segment Detection

Figure 2(a) shows the speech waveform *x*(*n*) for syllable /aka/ with E_p and f_c contours in the five bands. ROR contours E'_{pn} and f'_{cn} are shown in Fig. 3. Figure 4(a) shows the waveform *x*(*n*) for syllable */aka/*, with the smoothened spectrogram (from $|X_n(k)|$) in Fig. 4(b), transition index $T_r(n)$ in Fig. 4(c), and the located boundaries of transition segments in Fig. 4(d). The transitions corresponding to the onset and offset of vowel segment */a/* and the release burst of the consonant */k/* result in peaks in the transition index contour, and these are indicated by impulses in Fig. 4(c). Transition start and end points marked as negative and positive impulses in Fig. 4(d). The labels in Fig. 4(c) and 4 (d) indicate the locations of landmarks in seconds.

C. Results of Evaluation of Landmark Detector

The method of landmark detection was evaluated in terms of detection rates and temporal resolution using 50 manually annotated sentences (5 speakers × 10 sentences) from TIMIT database. Figure 5 shows the waveform of a portion of a sentence, manually annotated landmarks, and the detected landmarks. Segment boundaries involving abrupt transitions are detected accurately with good temporal resolution. Non-abrupt transitions involving semivowel to vowel transition (*/l/* to */a/*) got deleted and it is labeled as a single segment (label 14). The detection rates of the algorithm for different classes of phonemes are listed in Table 1, with the number of tokens for each class given in brackets. It is seen that detection rates for abrupt landmarks (stops and fricatives) is high (94-95 %) for 30 ms time resolution.

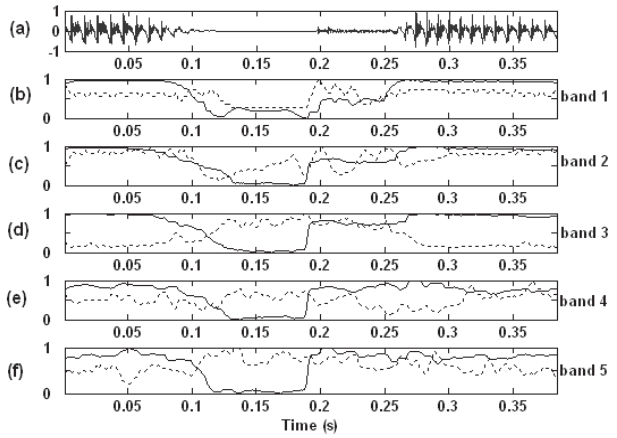

Fig. 2. Signal waveform for VCV syllable /aka/ and corresponding E_p (solid) and f_c (dotted) contours.

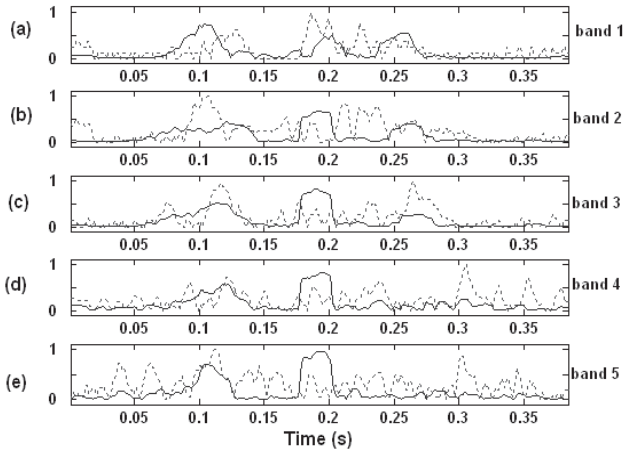

Fig. 3. ROR contours of E'_{pn} (solid) and f'_{cn} (dotted).

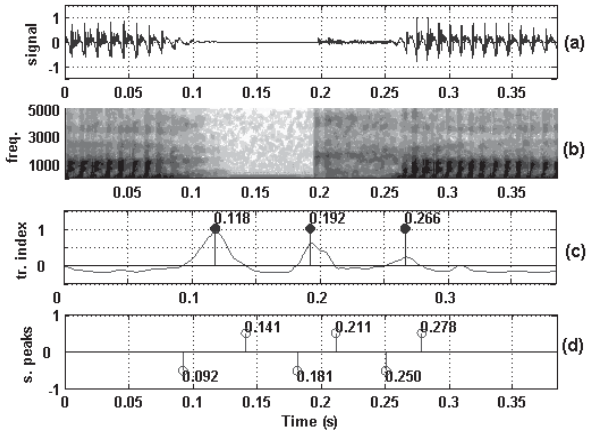

Fig. 4. (a) Signal waveform for VCV syllable /aka/ (b) Spectrogram, (c) Transition index $T_r(n)$, and (d) transition boundaries detected.

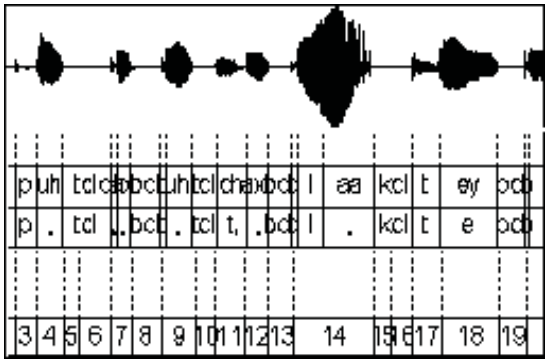

Fig. 5. (a) Waveform of a portion of the sentence '*put the butcher block table'*, (b) manually annotated (TIMIT) landmarks, and (c) automatically detected landmarks. Manual annotation: "bcl"- */b/* closure onset, "b"- */b/* release burst, etc. Automatic detection: landmarks numbered as 5, 6,..etc.

TABLE I
DETECTION RATES FOR TIMIT SENTENCES

Phoneme class	30 ms Det. (%)	20 ms Det. (%)	10 ms Det. (%)
Stop (548)	94	75	62
Fricative (266)	95	85	76
Nasal (154)	79	66	53
Vowel (614)	77	66	58
Sem. vowel (213)	42	33	26
Overall det. (%)	81.2	73.2	57.6

III TIME-SCALE AND INTENSITY MODIFICATION

Speech signal was digitized with sampling rate of 10 k sa/s and 16-bit quantization. The processing involved an analysis-modification-resynthesis stage based on HNM, and an intensity modification stage. The boundaries of segments for enhancement, detected by the landmark detection stage were used for time and intensity scaling.

A. HNM based Time-Scale Modification

In HNM based analysis-synthesis of speech, 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 [17].

Block diagram of HNM analysis stage is shown in Fig. 6. Fundamental frequency F_0 is estimated by the pitch detector using a normalized spectral cross-correlation function [18]. Speech segments are classified as voiced or unvoiced (V/UV) based on their harmonic structure, by the voicing detector. Analysis time instants are located pitch synchronously during voiced segments and at 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 [18], [19].

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. The β values are estimated automatically, so as to have the required expansion $(\beta > 1)$ during transition segments and compression $(\beta < 1)$ during steady state segments so as to maintain the overall speech duration unaltered. Steady state segment boundaries are located on either side of the transition segments with transition index below the threshold value for a minimum duration of 60 ms. For a time-scale expansion factor β_{tr} and transition segment boundaries (t_s, t_e), the compression factor β_{st} required for the steady state segment with boundaries (s_s, s_e) was calculated as

$$\beta_{st} = 1 - \left[(\beta_{tr} - 1)(t_e - t_s)/(s_e - s_s) \right] \quad (6)$$

HNM parameters in the time-warped scale are used for synthesizing the time-scale modified speech, as shown in Fig. 7. The harmonic part is obtained by overlap-adding a stream of short-time signals with estimated amplitudes and phases in a pitch-synchronous manner. 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.

B. Intensity Enhancement of Transition 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. 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.

C. Results of Experimental Evaluation

Experimental evaluation has been carried out using VCV syllables with vowel */a/* and stop consonants */p*, *b*, *t*, *d*, *k*, *g/*. Listening tests were conducted on normal hearing subjects with hearing loss simulated by adding broadband noise at 6 different SNR levels (∞, 0, -3, -6, -9, and -12 dB). For time-scaling factors above 2.0, expansion of stops and too much compression of the vowel segments resulted in perceptual artifacts. Five time-scaling factors (ts-*x* with *x* = 1.0, 1.2, 1.5, 1.8, and 2.0) were selected for evaluation. Based on informal listening tests with CVR enhancement of 3, 6, and 9 dB, consonant intensity enhancement by 6 dB was selected for evaluation.

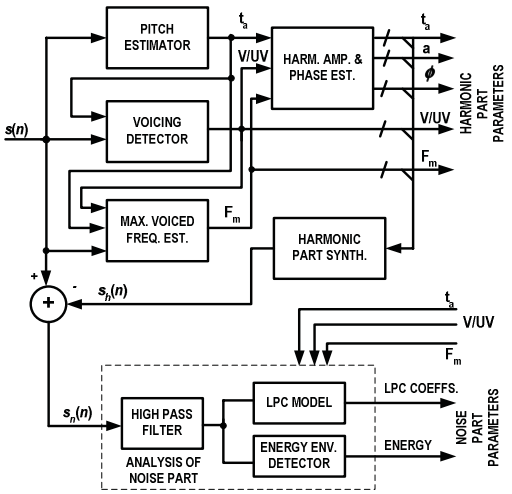

Fig. 6. HNM based analysis stage, adapted from [19].

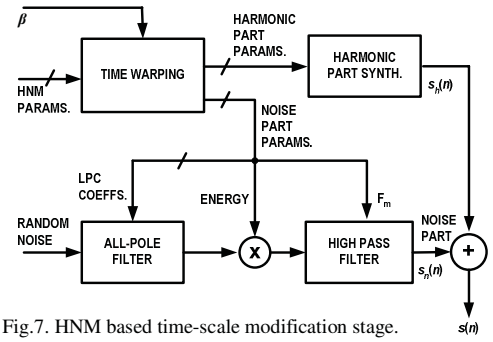

Fig.7. HNM based time-scale modification stage.

There were a total of 12 processing conditions for each VCV syllable: unprocessed (up), enhanced CVR without time-scaling (e), time-scaled with scaling factor *x* (ts-*x*), and enhanced CVR with time scaling (ets-*x*). This resulted in 72 test conditions (12 processing conditions × 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 options on the computer screen. Each test had 60 presentations, with each stimulus randomly presented 10 times. Consecutive

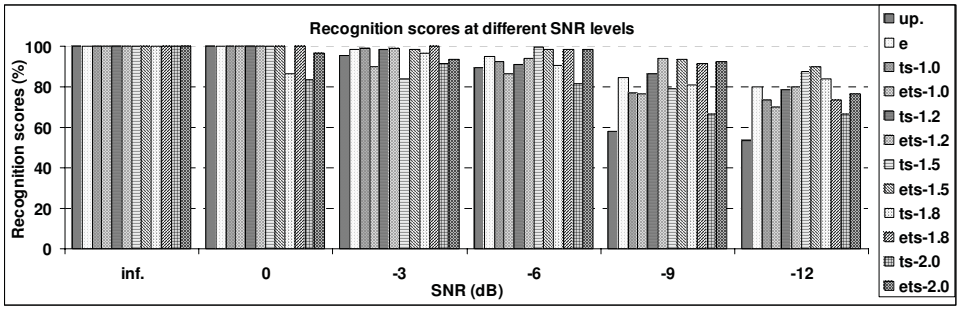

Fig. 8. Percentage recognition scores for different processing conditions.

presentations of the same stimulus were limited to a maximum of 3. The order in which tests were conducted was randomized, with a total of 5 tests for each condition.

Figure 8 shows the percentage recognition scores 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. At -9 dB SNR level, the recognition score for the unprocessed syllables decreased to 60 %. Time-scaling improved the score substantially, with transition scaling factor of 1.2-1.8 appearing to be the optimal range. The scores for intensity enhanced stimuli were comparable to the scores for time-scaling followed by intensity enhancement, indicating the importance of CVR enhancement in consonant identification.

IV. CONCLUSIONS

An automated method is presented for detection of landmarks characterized by spectral transitions, and enhancement of these segments by intensity and time-scale modification, without increasing the overall speaking rate. The transition segment boundaries are located using the rate of variation of energy and centroid frequency in five non-overlapping spectral bands. Time-scaling and CVR enhancement is performed using HNM based approach which introduces less perceptual artifacts. Listening tests with normal hearing subjects and broadband noise added to speech showed that under poor SNR conditions, processing improved recognition scores. Further experiments are to be conducted to find the optimum scaling factors and the extent of intelligibility improvement for different test materials and to investigate the effectiveness of the technique for subjects with moderate sensorineural loss. Combination of this technique with other speech processing schemes for improving speech perception for the hearing impaired listeners also needs to be investigated.

REFERENCES

[1] S. Furui, "On the role of spectral transition for speech perception," *J. Acoust. Soc. Am*., 80(4): 1016-1025, 1986.

[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," *J. Speech Hear. Res*. 28: 96-103, 1985.

[3] K. L. Payton, R. M. Uchanski, and L. D. Braida, "Intelligibility of conversational and clear speech in noise and reverberation with normal and impaired hearing," *J. Acoust. Soc. Am.*, 95(3): 1581-1592, 1994.

[4] M. A. Picheny, N. I. Durlach, and L. D. Braida, "Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech," *J. Speech Hear. Res.*, 298: 434-446, Dec. 1986

[5] J. C. Krause, and L. D. Braida, "Acoustic properties of naturally produced clear speech at normal speaking rates," *J. Acoust. Soc. Am.*, 115(1): 362-378, 2004.

[6] S. Liu, F. G. Zeng, "Temporal properties in clear speech perception," *J. Acoust. Soc. Am*., 120(1): 424-432, 2006.

[7] S. Gordon-Salant, "Recognition of natural and time/intensity altered CVs by young and elderly subjects with normal hearing," *J. Acoust. Soc. Am.*, 80(6): 1599-1607, 1986.

[8] V. Hazan, and A. Simpson, "The effect of cue-enhancement on the intelligibility of nonsense word and sentence materials presented in noise," *Speech Comm*., 24: 211-226, 1998.

[9] 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

[10] 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 Int. Con. Biomedical Engineering, Hong Kong (1994) 181-184

[11] J. R. Glass, and V. W. Zue, "Multi-level acoustic segmentation of continuous speech," in *Proc. ICASSP*, 429–432, 1988.

[12] A. Sarkar, and T. V. Sreenivas, "Automatic speech segmentation using average level crossing rate information," in *Proc. ICASSP*, 397–399, 2005.

[13] A. Alani, and M. Deriche, "A novel approach to speech segmentation using the wavelet transform," *Fifth Int. Symp. Signal Processing and its Applications. (ISSPA'99)*, 127-129, 1999.

[14] S. A. Liu, "Landmark detection for distinctive feature based speech recognition," *J. Acoust. Soc. Am*., 100(5): 3417-3430, 1996.

[15] V. Colotte, Y. Laprie, "Automatic enhancement of speech intelligibility," in *Proc. ICASSP*, vol.2: 1057–1060, 2000.

[16] M. D. Skowronski, J. G. Harris, "Applied principles of clear and Lombard speech for automated intelligibility enhancement in noisy environments," *Speech Comm.*, 48: 549-558, 2006.

[17] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," *IEEE Trans. Speech, Audio Proc., 9(1)*: 21-29, 2001.

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

[19] P. K. Lehana, and P. C. Pandey, "Harmonic plus noise model based speech synthesis in Hindi and pitch modification," *Proc. 16th Int. Congress Acoust*. (ICA2004, Kyoto, Japan), pp. 3333-3336, 2004.

[20] A. R. Jayan, P. C. Pandey, and P. K. Lehana, "Time-scaling of consonant-vowel transitions using harmonic plus noise model for improving speech perception by listeners with moderate sensorineural impairment,", in *Proc. 19th Int. Congress Acoustics* (ICA 2007), Madrid, paper no. CAS-03-006, 2007.

[21] J. S. Garofolo, *et al.*, "DARPA TIMIT acoustic-phonetic continuous speech corpus," U.S. Dept. of Commerce, NIST, Gaithersburg, MD, 1993.