Detection of Acoustic Landmarks with High Resolution for Speech Processing

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

Abstract- **Earlier Investigations have shown that speech processing to incorporate certain acoustic characteristics of clear speech in conversational speech can improve its intelligibility under adverse listening conditions. This processing needs detection of acoustic landmarks, the important regions in speech containing cues for phoneme identification. This paper presents a method for landmark detection in which energy and centroid frequency variations in a number of frequency bands are used to detect the landmarks, followed by a wavelet based decomposition for improving the time localization of the landmarks. The method was able to detect 98.1 % of release bursts in VCV syllables with a temporal resolution of 3 ms, and 86 % of stop consonant landmarks in sentence material with 20 ms resolution.**

I. INTRODUCTION

Speech produced by a speaker with an intention to improve intelligibility in a difficult communication environment (talking to a hearing impaired listener, noisy background, etc) is called "clear speech", and it is reported to be about 17 % more intelligible than conversational style speech [1]. This intelligibility advantage is applicable for normal hearing and hearing impaired listeners under quiet and adverse listening conditions. Acoustic properties of clear speech are different from those of conversational speech at the sentence, word, and phonemic levels. Clear speech is characterized by reduced speaking rate, more frequent and lengthy pauses, increase in fundamental frequency and its excursions, less number of sound deletions, increased consonantal segment intensity and duration, and more targeted vowel formants [2]-[4].

Investigations on acoustic properties of clear speech have identified certain features responsible for its improved intelligibility. Processing techniques for incorporating these features in conversational speech are reported to be effective in improving its intelligibility under adverse listening conditions. Speaking rate reduction, time-scale modification of spectral transition segments along with reduction of steady state segments, and enhancement of consonant-vowel intensity ratio (CVR) are some of the acoustic modifications reported to have positive impact on intelligibility [5]-[7].

Automatic enhancement of speech intelligibility has become more relevant in the present scenario where communication devices like mobile phones are commonly used in environments with varying types and levels of background noise. Use of directional microphones and speech enhancement techniques (adaptive filtering, spectral subtraction, etc.) attempt to reduce the noise at the talker end, but are not helpful under adverse listening conditions. Processing based on the properties of clear speech provides a way of improving speech perception in the presence of background noise at the listener's end, as well as for persons with moderate sensorineural hearing loss.

In most of the earlier investigations on intelligibility enhancement by using properties of clear speech, processing of conversational speech was carried out by modification of selected regions, manually located by inspection of speech waveforms and spectrograms. These annotation methods are time consuming and tedious and cannot be employed in real-time processing. Speech intelligibility enhancement making use of the acoustic properties of clear speech consists of an automatic landmark detection stage for selecting the acoustically salient regions, followed by a speech modification stage. The difficulty in accurately locating the regions for modification in an automated fashion is one of the limiting factors in the use of these techniques [7].

Landmarks are information rich areas in speech waveform, with concentration of acoustic cues for phoneme identification [8]. They generally coincide with regions of major spectral changes. Acoustically abrupt landmarks are produced by movement of a primary articulator or by sudden changes in the sound by glottal or velo-pharyngeal activity. Closures and releases of stops, fricatives, and nasals are acoustically abrupt. In an estimation on TIMIT database [9], about 68 % of the total landmarks were found to be acoustically abrupt, 29 % were vocalic, and 3 % were non-abrupt as in the case of semivowels and vowel-to-vowel transitions.

Landmark detectors with high detection rates and moderate temporal resolution may be adequate for applications like feature extraction for supporting speech recognition. But for intelligibility enhancement techniques in which specific modifications are applied on short-duration sub-phonemic segments, temporal resolution of detected landmarks are very important. In a technique reported by Colotte and Laprie [10], a spectral variation function based on mel-cepstral analysis was used to locate the regions for enhancement. It detected 82 % of the manually located landmarks with a temporal resolution of 20 ms. Stop bursts and unvoiced fricatives were automatically located and enhanced by amplification and time-scale modification.

Liu [8] reported an algorithm for detecting acoustically abrupt landmarks 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 the rate-of-rise contours were computed by taking the derivative of energy of the largest spectral component in each of the six bands. Peaks in these 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 temporal resolution with which landmarks were detected was evaluated by comparing with manually annotated speech material from TIMIT database. The algorithm detected 88 % of the total landmarks with a temporal resolution of 30 ms, 83 % with 20 ms, 73 % with 10 ms, and 44 % with 5 ms.

Niyogi and Sondhi [11] reported a landmark detector for stop consonants using their phonetic properties. Three parameters extracted from the short-time spectrum were used: log energy, energy above 3 kHz, and a measure of spectral flatness. Linear and non-linear operators optimized for minimizing an empirical risk function were found better for extracting parameters, compared to derivative of log energy used in [8].

Salomon *et al*. [12] used temporal parameters and spectral information to locate acoustically abrupt landmarks. A Hilbert transform based envelope operator was used for extracting parameters, which performed better than simple smoothening operation, in capturing the abrupt information. As in [8], derivative of log energy was used to locate onsets and offsets. Adaptive time steps (5 ms for stop bursts, 30 ms for frication, and 2 pitch periods in periodic regions) were used to improve temporal resolution.

Alani and Deriche [13] reported a segmentation technique using dyadic wavelet decomposition of 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). Energy variations in these bands were used to detect the segment boundaries. This method was able to track fast transitions associated with stops as well as slow transitions associated with lengthy vowel sounds. This method detected 90.9 % of the manually located landmarks, when evaluated using speech files from TIMIT database,

In an earlier investigation [14], we used properties of clear speech for improving speech perception by listeners with sensorineural impairment. A modified form of Liu's landmark detection algorithm [8] was used for locating the boundaries of vowel-consonant transition segments, specifically for stops. Speech modification was performed by expanding the transition segments using harmonic plus noise model (HNM) based analysis-synthesis [15]-[17]. The overall speech duration was kept unaltered by appropriately compressing the steady-state vowel segments. This method resulted in an improvement in recognition scores for VCV syllables by ~20 %, particularly at the lower SNR levels (below -6 dB), for time-scaling factors in the range of 1.5 to 1.8. The temporal resolution of the landmark detector used in [14] was found inadequate in detecting sub-phonemic events in VCV syllables and sentence material. The method is modified by adding a second pass using discrete wavelet transform based decomposition, for improving the time localization of the landmarks detected in the first pass. Multi-resolution feature of wavelet transforms facilitates time localization of sub-band spectral variations. The next section explains the technique and it is followed by results of evaluation, using VCV and sentence material.

II. Landmark Detection

Liu's method [8] for landmark detection used rate-of-rise (ROR) of peak energy in different frequency bands. For improving detection rates and temporal resolution, specifically for stop consonants, we use band centroids in addition to peak energy. Due to relatively high SNR near formant peaks in spectrum, these two parameters are less likely to be affected by addition of noise. In the first pass, ROR's are computed for peak energy and centroids and these are combined to get a single parameter called transition index, as indicator of the overall spectral variations. Locations of peaks in the transition index are marked as possible landmarks in the first pass of the method. In the second pass, wavelet decomposition is performed in a short duration window (40 ms), centered on these landmarks. Short-time energy and zero crossing rates are extracted from the lower decomposition levels (corresponding to higher frequency contents), and ROR's are computed using short time-steps (3 ms). Landmarks are relocated to the prominent peak locations in ROR's, in the second pass, to improve the temporal resolution.

A. Pass 1: Landmark detection

The spectrum is divided into five 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 consonant closures and releases, onsets and offsets of aspiration/frication associated with stops, fricatives, and affricates [8]. Landmark detection is based on detecting combined variation of peak energy E_p and centroid frequency f_c in the five bands. Any significant spectral transition results in a noticeable change in peak energy and centroid frequency in at least some of the spectral bands.

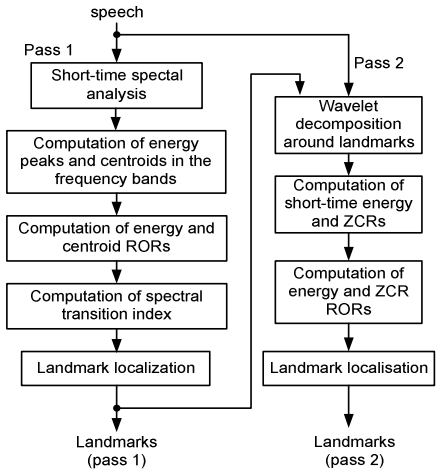

Fig. 1. Processing stages for landmark detection.

Speech is sampled at 10 k samples/s and short-time magnitude spectra are computed using 512-point FFT on 6 ms frames, using Hanning window. The short window length gives a spectral envelope with the effect of harmonics suppressed. The frames are taken every 1 ms to track any abrupt spectral variations in the signal. A 20-point moving average is used along the frequency index *k*, for obtaining smoothened spectral components $|X_n(k)|$, with n spaced every 1 ms. The log of largest spectral component in each band is used to form energy contours, $E_p(b,n)$ for band *b* and frame *n* 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 contour is calculated as

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

where f_S is the sampling frequency, and N is the number of points in FFT computation. ROR of contours of E_p and f_c are obtained as the magnitude of the first difference, every 1 ms.

$$E_p'(b,n) = \left|E_p(b,n+K) - E_p(b,n-K)\right| \quad (3)$$

$$f_C'(b,n) = \left|f_C(b,n+K) - f_C(b,n-K)\right| \quad (4)$$

In the first pass, to improve detection rates, a high value of *K* (=25) is used to get a 50 ms time-step in the first difference computation. These ROR functions are normalized to the 0-1 range (by shifting and scaling) as $E_{pn}'(b,n)$ and $f_{cn}'(b,n)$. These normalized functions are used to get the energy based transition index

$$T_e(n) = (1/5)\sum_{b=1}^{5} E_{pn}'(b,n) \quad (5)$$

and energy and centroid based transition index

$$T_{ec}(n) = (1/5)\sum_{b=1}^{5} E_{pn}'(b,n) f_{cn}'(b,n) \quad (6)$$

These indices have positive values with low amplitude variations in the vowel segments and prominent peaks during the spectral transitions for plosives. Transition segment boundaries are located by comparing either of these indices with an empirically selected threshold. The waveform and ROR contours in lower three bands of the VCV syllable /uka/ are shown in Fig. 2. During the /k/ release burst, the ROR's in energy show flat headed peaks compared to centroid ROR's, illustrating the possibility of improvement in temporal resolution in localizing the landmarks by combining energy and centroid ROR's. Figure 3 shows the results of landmark detection for the syllable /uka/. Waveform *x*(*n*), spectrogram, and the combined transition index $T_{ec}(n)$ are shown in the parts (a), (b), and (c) respectively. It is seen that in $T_{ec}(n)$, steady state vowel and silence segments show near-zero values, whereas spectral transitions corresponding to the onsets of vowels, stop closure, and release burst result in peaks.

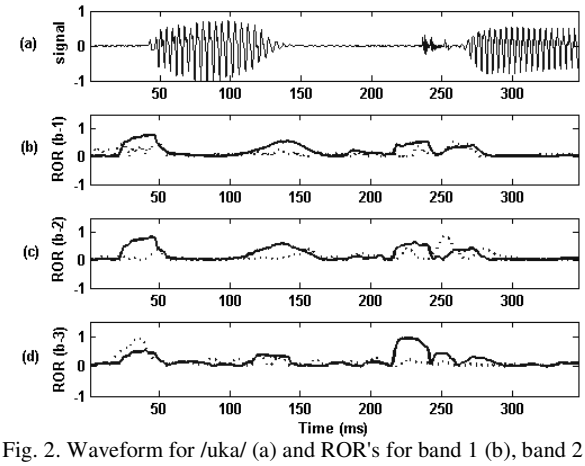

Fig. 2. Waveform for /uka/ (a) and ROR's for band 1 (b), band 2 (c), and band 3(d). Solid: $E_{pn}'(b,n)$, dotted: $f_{cn}'(b,n)$.

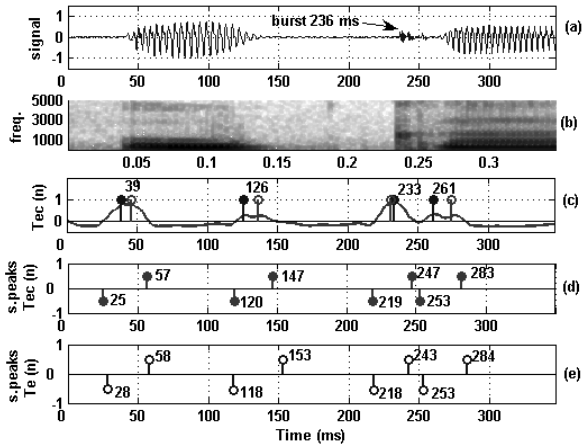

Fig. 3. Processing results for /uka/ : Waveform(a), spectrogram (b) transition index T_{ec} (c), and transition segment boundaries using T_{ec} (d), and using T_e (e).

In Fig. 3(c), the peaks, localizing the landmarks, are marked as impulses with filled dots. The transition index $T_e(n)$, based on energy alone, is not plotted here, but its peaks are indicated as impulses with unfilled dots. The manually located position of release burst is at 236.0 ms. The locations obtained from peaks in $T_{ec}(n)$ and $T_e(n)$ are at 233 ms and 231 ms, with errors of 3 ms and 5 ms respectively. The boundaries of transition segments are located by threshold comparison and these are marked in Fig. 3(d), and 3(e) as negative impulses for start of transition and positive impulses for end of transition (with filled and unfilled dots as obtained from the two indices $T_{ec}(n)$ and $T_e(n)$). It is seen that landmarks located using energy based transition index result in late detection of vowel onsets and offsets, but early detection of the release bursts.

B. Pass I1: Localizing landmarks

To improve the temporal resolution of the method, a second pass in which analysis was performed on 40 ms window

centered at the landmark location obtained in Pass 1. The signal in this window is decomposed into 6 levels (l = 1 to 6) using discrete Meyer wavelet [18], [19]. The high frequency contents (above ~1 kHz) in the lower two levels are reconstructed and short-time energy $E(l,n)$ and zero crossing rates $Z(l,n)$ of these two levels (l=1, 2) are computed every 1 ms, using a short-time window of length 3 ms. The ROR contours of short-time energy $E'(l,n)$ and zero crossing rates $Z'(l,n)$ are formed by taking their derivatives with a time-step of 3 ms. As in Pass 1, these ROR's are normalized and used to get a transition index,

$$T_{ez}(n) = (1/2)\sum_{l=1}^{2} E_n'(l,n) Z_n'(l,n) \qquad (7)$$

where $E_n'(l,n)$ and $Z_n'(l,n)$ are the normalized ROR's for the windowed segment. Figure 4 shows the windowed segment around the closure release burst for the waveform of /uka/, ROR's of the reconstructed signals in the lower two levels, and the transition index derived from the ROR's. Windowed segment of 40 ms duration (213 to 253 ms) was taken for the burst location at 233.0 ms obtained in Pass 1. Transition index contour in Pass 2 shows a prominent peak at 236.4 ms, giving an excellent match to manual location of the burst at 236.0 ms.

III. Evaluation Results

The landmark detection technique was evaluated in terms of detection rates and temporal resolution using VCV syllables and manually annotated continuous speech material from TIMIT database.

A. Evaluation using VCV syllables

VCV syllables recorded from 2 speakers (one male and one female), consisting of unvoiced stops /p/, /t/, and /k/, in the context of vowels /a/, /i/, and /u/ were used for evaluation.

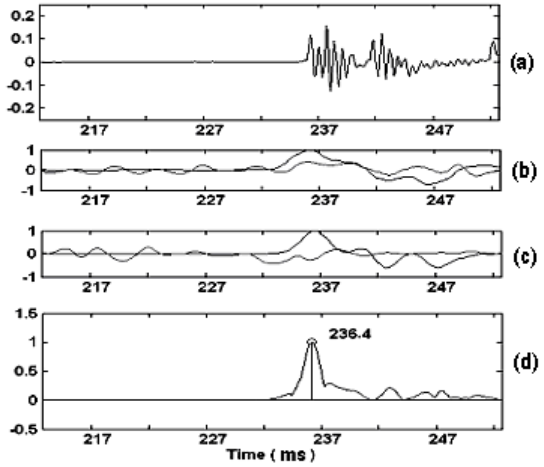

Fig. 4. (a) Windowed segment used in second pass, (b) energy and ZCR ROR's of level 1, (c) ROR's of level 2, and (d) transition index T_{ez} computed from ROR's in (b) and (c).

There were a total of 54 utterances (2 speakers × 3 initial vowels × 3 stops × 3 final vowels). The positions of the peaks corresponding to release bursts (first peak in the transition index contour after the onset of stop closure located by the method) were compared with the locations obtained by manual inspection of the waveforms and spectrograms.

The number of bursts missed in the detection for different time resolution limits are given in Table 1 and Table 2, for Pass 1 and Pass 2 respectively. Each cell gives the number of errors for a stop for an initial vowel context, out of a maximum of 18 detections. The last row in each table gives the percentage detection, for the total of 54 bursts. For a resolution limit of 30 ms, Pass 1 is able to achieve 100 % detection. The detection marginally decreases at 20 and 10 ms. But there is a severe decrease in detection rate for resolution limit of 5 ms. By introducing Pass 2, 100 % detection is achieved for a resolution limit of 5 ms. Out of these, 98.1 % of the release bursts were detected with a resolution of 3 ms.

TABLE 1
PASS 1: ERRORS IN RELEASE BURST DETECTION

Stop	30 ms			20 ms			10 ms			5 ms		
	Initial vowel			Initial vowel			Initial vowel			Initial vowel		
	a	i	u	a	i	u	a	i	u	a	i	u
/p/	-	-	-	-	-	-	-	-	-	1	1	2
/t/	-	-	-	-	-	-	-	-	-	1	1	2
/k/	-	-	-	1	-	-	1	-	1	3	3	3
Det.%	100			98.1			96.3			68.5		

TABLE II
PASS 2: ERRORS IN RELEASE BURST DETECTION

Stop	10 ms			7 ms			5 ms			3 ms		
	Initial vowel			Initial vowel			Initial vowel			Initial vowel		
	a	i	u	a	i	u	a	i	u	a	i	u
/p/	-	-	-	-	-	-	-	-	-	-	-	-
/t/	-	-	-	-	-	-	-	-	-	-	1	-
/k/	-	-	-	-	-	-	-	-	-	-	-	-
Det.%	100			100			100			98.1		

B) Evaluation using sentences

The technique was applied for landmark detection in 50 manually annotated sentences (5 speakers × 10 sentences) from the TIMIT database. Figure 5 shows the waveform of a sentence, along with manual and automatically detected landmarks. The closure symbols for the stops b, d, g, p, t, k are bcl, dcl, gcl, pcl, tcl, and kcl, respectively. Landmarks involving abrupt transitions are detected accurately with good temporal resolution. Non-abrupt landmarks 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 method for Pass 1 and Pass 2, for different classes of phonemes are listed in Table 3, with the number of tokens for each class given in brackets. It is seen that detection rates for abrupt landmarks (stops and fricatives) is very high (94-95 %)

for 30 ms time resolution. Pass 2 improves the overall detection rates by about 2 %.

The temporal resolution of the method in locating the onset of closure and burst locations for stop consonants was evaluated using 418 tokens present in the same set of sentences. The averaged localization errors for different types of stop landmarks for both passes are shown in Fig. 6. The second pass reduces the localization error, averaged across all types of landmarks by about 2 ms.

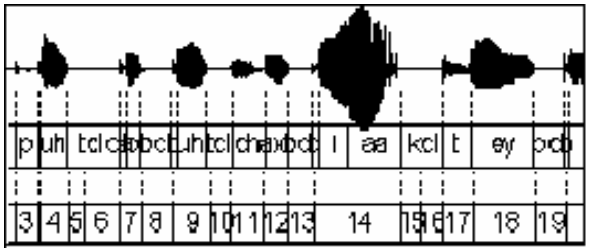

Fig. 5. (a) Waveform of the sentence '*put the butcher block table'*, (b) manual (TIMIT) landmarks, and (c) detected landmarks. Manual annotation: "bcl"- /b/ closure onset, "b"- /b/ release burst, etc. Automatic detection: landmarks numbered as 5, 6,..etc.

TABLE III
DETECTION RATES FOR TIMIT SENTENCES

Phoneme class	30ms		20 ms		10 ms	
	Det. (%)		Det. (%)		Det. (%)	
Pass	1	2	1	2	1	2
Stop (548)	94	96	82	86	62	66
Fricative (266)	95	95	90	90	76	79
Nasal (154)	80	79	70	70	53	51
Vowel (614)	77	79	70	71	58	57
S. vowel (213)	69	70	68	67	60	61
Overall det. (%)	84.1	85.7	76.4	78.0	61.7	63.0

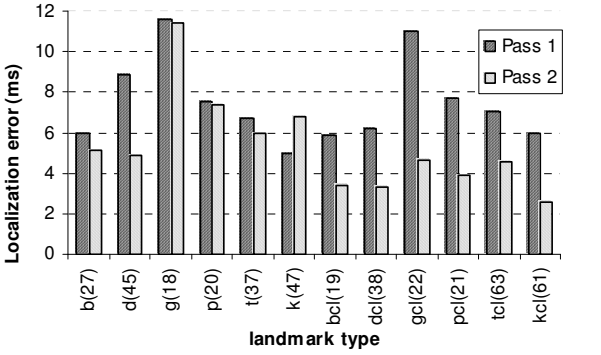

Fig. 6. Avg. temporal resolution of acoustic landmarks.

IV. CONCLUSION

An acoustic landmark detection technique is described with a two pass strategy to improve the detection rates and temporal resolution of landmarks, especially for stop consonants. Energy and centroid frequency variations in spectral bands are used to locate the landmarks and they are refined by wavelet transform decomposition. By restricting the window length and time-steps used, it was possible to improve temporal resolution. The method needs to be evaluated in terms of detection rates and temporal resolution in the presence of various kinds of noise at different levels to ensure its practical usefulness for locating landmarks in conversational speech.

REFERENCES

[1] 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.*, vol. 28, pp. 96-103, Mar. 1985.

[2] 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.*, vol. 298, pp. 434-446, Dec. 1986.

[3] F. R. Chen, "Acoustic characteristics and intelligibility of clear and conversational speech at the segmental level," *M.S. Dissertation*, MIT, Cambridge, 1980.

[4] J. C. Krause and L. D. Braida, "Acoustic properties of naturally produced clear speech at normal speaking rates," *J. Acoust. Soc. Am.*, vol. 115, no. 1, pp. 362-378, Jan. 2004.

[5] S. Gordon-Salant, "Effects of acoustic modification on consonant recognition by elderly hearing-impaired subjects," *J. Acoust. Soc. Am.*, vol. 81, no.4, pp.1199-1202, 1987.

[6] T. G. Thomas, "Experimental evaluation of improvement in speech perception with consonantal intensity and duration modification," *Ph.D. Dissertation*, Dept. of Elect. Engg, IIT Bombay, 1996.

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

[8] S. A. Liu, "Landmark detection for distinctive feature based speech recognition," *J. Acoust. Soc. Am.*, vol. 100, no. 5, pp. 3417-3430, 1996.

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

[10] V. Colotte and Y. Laprie, "Automatic enhancement of speech intelligibility," in *Proc. ICASSP'2000*, Istanbul, Turkey, 2000, pp. 1057–1060.

[11] P. Niyogi and M. M. Sondhi, "Detecting stop consonants in continuous speech," *J. Acoust. Soc. Am.*, vol. 111, no. 2, 2002, pp. 1063-1076.

[12] A. Salomon, C. Y. Espy-Wilson, and O. Deshmukh, "Detection of speech landmarks: Use of temporal information," *J. Acoust. Soc. Am.*, vol. 115, no. 3, 2002, pp. 1296-1305.

[13] A. Alani and M. Deriche, "A novel approach to speech segmentation using the wavelet transform," in *Proc. 5th Int. Symp. Signal Processing and its Applications.* (*ISSPA'99*), 127-129, 1999.

[14] 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.

[15] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis", *IEEE Trans. On Speech and Audio Processing,* vol. 9., pp. 21-29, Jan. 2001.

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

[17] P. K. Lehana and P. C. Pandey, "Harmonic plus noise model based speech synthesis in Hindi and pitch modification," in *Proc. 18th Int. Congress Acoustics* (ICA 2004), Kyoto, Japan, 2004, pp. 3333-3336.

[18] S. Mallat, *A Wavelet Tour of Signal Processing*. 2nd ed., New Delhi: Reed Elsevier, 2006.

[19] I. Daubechies, *Ten Lectures on Wavelets,* Philadelphia: SIAM, 1992.