DETECTION OF STOP LANDMARKS USING GAUSSIAN MIXTURE MODELING OF SPEECH SPECTRUM

A. R. Jayan and P. C. Pandey

Department of Electrical Engineering
Indian Institute of Technology Bombay
Powai, Mumbai, 400 076, India
{arjayan, pcpandey} @ ee.iitb.ac.in

ABSTRACT

Perception of speech under adverse listening conditions may be improved by processing it to incorporate properties of clear speech. It needs automated detection of stop landmarks and enhancement of bursts and transition segments. A technique for accurate detection of stop landmarks in continuous speech based on parameters derived from Gaussian mixture modeling of log magnitude spectrum, a voicing onset-offset detector, and a spectral flatness measure is presented. Applying the technique on sentences from the TIMIT database resulted in burst detection rates of 98, 97, 95, 90, and 73 % at temporal accuracies of 30, 20, 15, 10, and 5 ms respectively.

Index Terms— Speech enhancement, acoustic landmark detection, Gaussian mixture modeling.

1. INTRODUCTION

The acoustic landmarks in speech signal are characterized by a concentration of cues important for speech perception. A stop consonant typically has three landmarks: closure of the vocal tract, closure release burst, and onset of voicing [1], [2]. Depending on the phonetic context, all the three landmarks may or may not be present. As the stops are transient sounds with low energy, their perception gets severely affected due to masking by background noise, or increased spectral and temporal masking associated with sensorineural hearing impairment [3], [4]. Under adverse listening conditions, speakers naturally adopt a speaking style called "clear speech", which is reported to be about 17 % more intelligible than conversational speech [5], [6]. Several intelligibility enhancement techniques utilizing properties of clear speech, based on detection and enhancement of frication bursts and transition segments have been reported [7]–[10]. Detection of landmarks with good temporal accuracy (alignment with actual landmarks) and without insertions (false detections) is critical for these techniques.

Rate-of-rise measures based on the first difference of parameters obtained from a set of fixed bands (e.g. band energy, spectral centroid, spectral peak, etc.) have been used to locate landmarks, [1], [2], [11], [12]. These methods provide good detection rates (~90 %) with moderate temporal accuracies (20-30 ms), but detection rates fall below 50 % for temporal accuracies of the order of 5 ms. Use of distance measures with time-steps adapted to the type of landmark being detected and use of optimum filtering have been reported to improve the temporal accuracy for stop landmarks [2], [12]. However, use of fixed bands severely constrains the extracted parameters. In case of more than one spectral prominence in a band or a spectral prominence spread across bands, band parameters may not meaningfully represent the spectral variations. Further, the use of fixed bands may introduce speaker dependent variability in the performance of the landmark detector [11]. Using a spectral modeling approach which can adapt to the dynamic nature of the spectrum, and using the variations in more than one parameter derived from the model, detection rate and temporal accuracy of a landmark detector may be improved.

A Gaussian mixture model (GMM) of the short-time speech spectrum provides a parametric representation of the spectral envelope using a weighted sum of Gaussian functions. An approximation with a small error can be obtained, for all classes of sounds, using a small number of Gaussian components in the mixture model. Zolfaghari and Robinson [13] used a GMM based parametric scheme for extracting formant-like features. GMM parameters have been used for improving speech recognition in noisy environments and for performing spectral modifications [14]–[17]. We applied GMM parameters for landmark detection in VCV utterances [18]. The detection rates for stop release bursts in VCV utterances from 3 male and 3 female speakers were 94%, 80%, and 60% for temporal accuracies of 30 ms, 10 ms, and 5 ms respectively. No speaker dependence in the performance was observed, but the detection rates decreased for conversational speech.

This paper presents a technique, based on Gaussian mixture modeling of the short-time speech spectrum, for detection of stop landmarks in continuous speech. The objective is to improve the temporal accuracy and to reduce the number of insertions, so that the technique can be used for enhancing the burst and transition segments to improve speech intelligibility under adverse listening conditions. The technique is evaluated by comparing the detected landmarks with the manually marked ones in VCV utterances and TIMIT sentences.

2. GAUSSIAN MIXTURE MODELING OF SHORT-TIME SPEECH SPECTRUM

Our technique for detection of stop landmarks uses parameters obtained by Gaussian mixture modeling of short-time log magnitude spectrum. The approximation errors in modeling the log

magnitude spectra are smaller than those in modeling the squared magnitude or the magnitude spectra. Further, the first difference of the log magnitudes involves relative values and gain normalization across the utterances is not necessary.

The log magnitude spectrum is computed, for speech signal acquired at 10 kHz, using 512-point DFT on 6 ms Hanning windowed frames, every 1 ms. The short length window suppresses the pitch harmonics which may cause the Gaussian components to track non-formant peaks [13], [16]. The high frame rate helps in tracking fast spectral variations. The magnitude spectrum is smoothed by a 50-point median filter, along the frequency index k. The smoothed log magnitude spectrum $S_n(k)$ is approximated by a weighted sum of M Gaussian functions as

$$\hat{S}_n(k) = \sum_{g=1}^{M} w_{gn}\, G(\mu_{gn}, \sigma_{gn}, k)\,, \tag{1}$$

where w_{gn}, μ_{gn}, and σ_{gn}^2 represent the weight, mean, and variance respectively of the g th Gaussian in the mixture model for frame n.

The GMM parameters are estimated using expectation maximization (EM) algorithm [15], [19]. With a given initialization, the EM algorithm iteratively computes the maximum likelihood estimates of the model parameters from the magnitude spectrum. The smoothed log magnitude spectrum is viewed as a histogram with rectangular bins placed at each frequency index k. Iterations are started with an initial set of parameters. The probability $p(g \mid k)$ that frequency measurement k comes from the Gaussian component g, is evaluated as

$$p(g \mid k) = w_{gn}\, G(\mu_{gn}, \sigma_{gn}, k) \;/\; \sum_{g=1}^{M} w_{gn}\, G(\mu_{gn}, \sigma_{gn}, k)\,. \tag{2}$$

The new mixture weights, means, and variances are calculated as

$$\hat{w}_{gn} = \sum_{k=1}^{N/2} S_n(k)\, p(g \mid k) \;/\; \sum_{k=1}^{N/2} S_n(k)\,, \tag{3}$$

$$\hat{\mu}_{gn} = \sum_{k=1}^{N/2} k\, S_n(k)\, p(g \mid k) \;/\; \sum_{k=1}^{N/2} S_n(k)\, p(g \mid k)\,, \tag{4}$$

$$\hat{\sigma}_{gn}^2 = \sum_{k=1}^{N/2} (k - \hat{\mu}_{gn})^2 S_n(k)\, p(g \mid k) \;/\; \sum_{k=1}^{N/2} S_n(k)\, p(g \mid k)\,. \tag{5}$$

These parameters are used in the next iteration. Here N denotes number of points in the DFT computation. The iterations are continued until the changes in the parameter values in successive iterations become less than a set threshold or the number of iterations reaches a set limit. A reasonably good approximation of the speech spectrum is possible with 4 to 5 Gaussian components [17], [18]. As the approximation errors for the two do not significantly differ, we have used the 4-component model.

The choice of initial parameters for the EM algorithm affects the solutions obtained and the number of iterations needed. Use of parameters obtained for the previous frame as the initialization for the current frame results in comparatively smoother parameter tracks in a smaller number of iterations, but the estimated values respond poorly to rapid spectral changes. The mixture weights were initialized with equal values. The means and the variances were initialized with values corresponding to the average vowel formant frequencies and extreme bandwidths respectively (first: 600, 160; second: 1200, 200; third: 2400, 300; fourth: 3600, 400 Hz) as given in [20]. These initialization values resulted in parameters tracks responding to the spectral changes in the speech signals from a number of male and female speakers. For sampling frequency F_s and N-point DFT, the initialization values for the g th component are obtained from the values of the average formant F_g and extreme bandwidth B_g by using the correspondence $\mu_g = NF_g/F_s$ and $\sigma_g = N(B_g/2.35)/F_s$. The number of iterations is set at 12 because no significant decrease in approximation error is observed by increasing the number of iterations. Figure 1 shows modeling for a 6 ms segment of vowel /a/ spoken by a male speaker. As seen in this case, the peaks in the GMM approximation generally matched the resonance peaks in the spectrum.

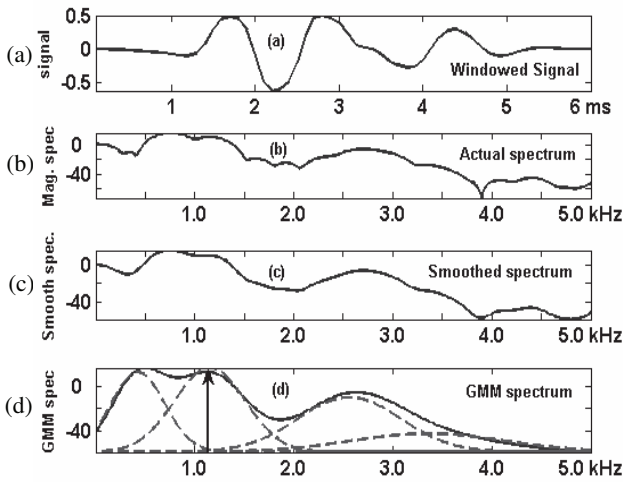

Fig. 1. Modeling of a segment of vowel /a/: (a) windowed segment of 6 ms, (b) log magnitude spectrum (in dB), (c) smoothed spectrum (in dB), (d) GMM approximated spectrum with dotted lines indicating the individual Gaussian components.

3. DETECTION OF STOP LANDMARKS

A rate of change function defined on the parameters obtained by modeling of the smoothed log magnitude spectrum using 4 Gaussian components, along with a voicing onset-offset detector [1] and a spectral flatness measure [9], is used for detection of stop landmarks.

As compared to the Gaussian weights, the amplitudes of the GMM approximated envelope at the four mean values were found to be more consistently related to the spectral changes. The means $\mu_g(n)$, square root of the variances $\sigma_g(n)$, and amplitude of the GMM envelope at the mean locations $A_g(n)$ for the four Gaussian components are processed by 30-point median filtering, in order to smoothen the tracks during steady state segments without significantly smearing the variations corresponding to abrupt spectral transitions. These smoothed parameters $A'_g(n)$, $\mu'_g(n)$ and $\sigma'_g(n)$ are used for calculating a rate-of-change (ROC) function given as

$$r_c(n) = r_A(n)\, r_\mu(n)\, r_\sigma(n) / R\,, \tag{6}$$

where

$$r_A(n) = \sum_{g=1}^{4} \left| A'_g(n) - A'_g(n - n_s) \right|, \tag{7}$$

$$r_\mu(n) = \sum_{g=1}^{4} \left| \mu'_g(n) - \mu'_g(n - n_s) \right|, \tag{8}$$

$$r_\sigma(n) = \sum_{g=1}^{4} \left| \sigma'_g(n) - \sigma'_g(n - n_s) \right|, \tag{9}$$

and R is used to scale the maximum value of r_c to 1. The time step n_s corresponds to 2 ms. The use of short time-step suppresses relatively slow spectral variations associated with semivowels, voicing offsets, etc. The product operation on individual ROCs ensures that the strong peaks occur at the point of sharp spectral variation. These peaks indicate the possible location of release burst onsets.

A voicing detector based on the spectral peak variation in the band 0-400 Hz is used for detecting voicing onsets and offsets. Peaks taken every 1 ms from the squared magnitude spectrum, and smoothed by 20-frame moving average, form the contour $E(n)$. A first difference operation on the log of $E(n)$ with a time-step of 26 ms is used to get a rate-of-rise function $r_e(n)$. Positive and negative peaks above and below thresholds of +9 and -9 dB respectively are taken as voicing onsets (+g) and offsets (-g) respectively. Another validation of the located landmarks is carried out using spectral flatness measure (SFM) calculated as

$$\mathrm{SFM}(n)=\left(\prod_{k=1}^{N/2}|X_n(k)|\right)^{2/N} \Big/ \left((2/N)\sum_{k=1}^{N/2}|X_n(k)|\right), \qquad (10)$$

where N is the number of points in the DFT computation. It is low for voiced frames with peaky spectra and close to 1 for frication with flat spectra [9]. It is computed for 20 ms Hanning windowed frames every 1 ms.

The detection of +g and -g peaks is used for locating a burst in a preceding segment using the peaks in the ROC of GMM parameters. If a +g peak is preceded by a -g peak, the segment extends backwards to an instant 50 ms before the -g peak, otherwise it extends backwards by 50 ms. A burst is located on the detection of a -g peak only if it is preceded by another -g peak. In such a case, the segment extends backwards 50 ms before the preceding -g peak. The segment boundaries get slightly extended to the voiced segments, at least on one side, by giving an offset, and the need for setting a threshold for the peaks in the ROC is eliminated. The detected burst onset is validated by checking for the presence of a preceding closure interval. The amplitude tracks for the higher three Gaussians, normalized by the respective peaks in the segment, should remain below 0.5 in the preceding 10 ms interval. The first component is not used because of the possibility of the voice bars preceding the burst. This validation eliminates the peaks in ROC corresponding to the unvoiced fricatives. The stop release burst is declared valid only if the spectral flatness measure is above 0.5 for at least one frame within 15 frames on either side. For a valid release burst, the preceding -g point is taken as the closure onset. The following +g point, with value of spectral flatness measure below 0.5, is taken as the voicing onset.

4. TEST RESULTS

The technique was evaluated using VCV utterances and sentences from the TIMIT database. The detection rates were tabulated for different temporal accuracies, i.e. misalignment of the detected landmarks with respect to the manually located ones.

4.1. Detection of stop landmarks in VCV utterances

The technique was applied for detection of stop landmarks in the VCV utterances consisting of 6 stops (/*b*/, /*d*/, /*g*/, /*p*/, /*t*/, /*k*/) in the context of 3 vowels (/*a*/, /*i*/, /*u*/). These were recorded from 10 speakers (5 female and 5 male). There were a total of 180 utterances, with an average duration of 400 ms. Figure 2 shows the median smoothed Gaussian parameter tracks for /*apa*/. Wideband spectrogram, Gaussian modeled spectrogram, and the ROC contour

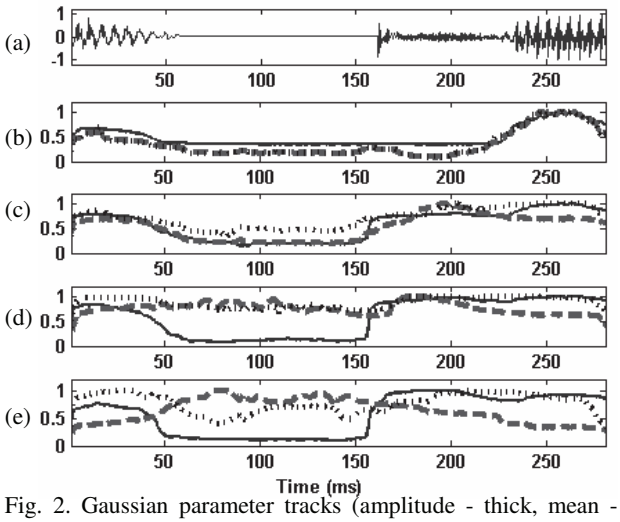

Fig. 2. Gaussian parameter tracks (amplitude - thick, mean - dashed, and variance - dotted): (a) waveform of /apa/,: (b) Gaussian 1, (c) Gaussian 2, (d) Gaussian 3, and (e) Gaussian 4.

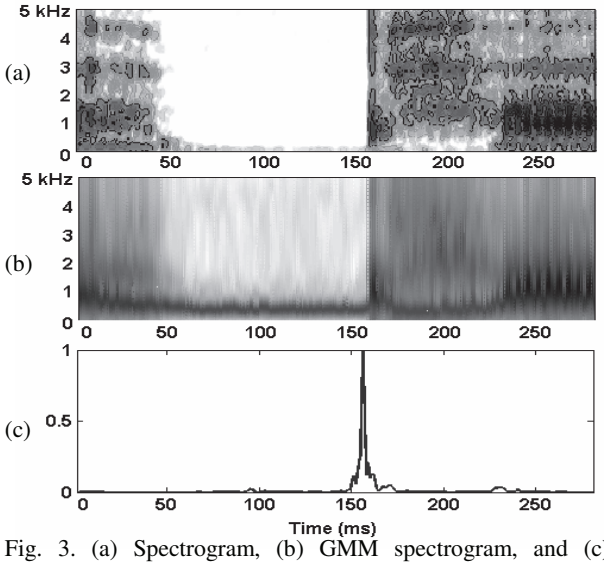

Fig. 3. (a) Spectrogram, (b) GMM spectrogram, and (c) Gaussian ROC contour, for /apa/.

are shown in Fig. 3. There were no insertion errors. Stop landmark detection rates for different temporal accuracies are given in Table 1. Out of the total 180 bursts, 90 % were detected within 5 ms of the manual landmarks. Detection rates were 90, 92, 93, 96, and 98 % for temporal accuracies of 5, 10, 15, 20, and 30 ms respectively. The detection rates for the voicing onsets were almost similar. The closure onsets were detected with less temporal accuracy as compared to the voicing onsets.

4.2. Detection of stop landmarks in TIMIT sentences

The technique was used for detection of the stop landmarks in a set of 50 conversational style sentences (3 female and 2 male speakers × 10 sentences each) from the TIMIT database. The detected landmarks were compared with the corresponding manual transcription in the database. Performance of the detection process for different types of landmarks is summarized in Table 2. Out of the total 306 stops, the method was able to detect 223 stops (~73 %) within 5 ms of the TIMIT transcriptions. An analysis of

alignment errors for burst detection showed a mean bias of -1.4 ms, and a standard deviation of 5.8 ms. Closure onsets were evaluated on 270 tokens (marked as *bcl, dcl, gcl, pcl, tcl, kcl* in the TIMIT transcription) having preceding voiced segments. Voicing onset detection was evaluated on 232 stop release bursts followed by voiced segments. There were a total of 39 (~13 %) insertions, which are described by the phoneme transitions listed in Table 3. Affricate detections were not counted as insertions. The insertions were mainly due to burst like clicks preceded by low energy segments and abrupt spectral transitions caused by glottal stops (marked as *q* in the TIMIT transcription).

The techniques reported earlier generally give excellent detection rates at temporal accuracies of 30 and 20 ms, with the rates falling sharply at 10 and 5 ms. For burst detection, the technique reported here gave detection rates of 98, 97, 95, 90, 73 % at temporal accuracies of 30, 20, 15, 10, 5 ms, respectively.

Table 1
Detection rates of stop landmarks in the VCV utterances

Landmark (no. of tokens)	Temporal accuracy (ms)				
	5	10	15	20	30
Burst (180)	90	92	93	96	98
Closure (180)	36	64	73	83	93
Voicing onset (180)	76	93	98	98	99

Table 2
Detection rates of stop landmarks in the TIMIT sentences

Landmark (no. of tokens)	Temporal accuracy (ms)				
	5	10	15	20	30
Burst (306)	73	90	95	97	98
Closure (270)	19	40	63	80	90
Voicing onset (232)	45	71	82	91	96

Table 3
Insertions rates in the TIMIT sentences

Type of transition	Insertion rate (%)
Clicks, glottal stops	8
Vowel-semivowel	4
Stop to /l/, /r/	1

5. CONCLUSION

A stop landmark detection technique based on Gaussian mixture modeling of speech spectrum was investigated. A total of 4 Gaussian functions with appropriately initialized parameters were used. The landmarks were detected using a rate-of-change function on the parameters derived from the GMM approximated spectrum, along with a voicing onset-offset detector and spectral flatness measure. The performance of the landmark detector needs to be evaluated under noisy conditions. Fine tuning of the modeling process and investigations with other rate-of-rise measures may further improve the detection rates. Application of the technique in speech enhancement under adverse listening conditions needs to be evaluated.

6. REFERENCES

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

[2] A. Salomon, C. Y. Espy-Wilson, and O. Deshmukh, "Detection of speech landmarks: Use of temporal information," J. Acoust. Soc. Am. 115 (3), pp. 1296-1305, 2005.

[3] D. O'Shaughnessy, Speech Communications: Human and Machine, Reading, Mass.: Addison-Wesley, 1987.

[4] M. A. Stone and B. C. J. Moore, "Spectral feature enhancement for people with sensorineural hearing impairment: Effects on speech intelligibility and quality," J. Rehab. Research., 29(2), pp. 39-56, 1992.

[5] 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 (1), pp. 96-103, 1985.

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

[7] R. W. Guelke, "Consonant burst enhancement: A possible means to improve intelligibility for the hard of hearing," J. Rehab. Res. Develop., 24 (4), pp. 217-220, 1987.

[8] V. Colotte and Y. Laprie, "Automatic enhancement of speech intelligibility," in Proc. ICASSP 2000, pp. 1057–1060.

[9] M. D. Skowronski and J. G. Harris, "Applied principles of clear and Lombard speech for automated intelligibility enhancement in noisy environments," Speech Comm., 48 (5), pp. 549-55, 2006.

[10] 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 2007, paper no. CAS-03-006.

[11] S. A. Liu, "Landmark detection for distinctive feature based speech recognition," Ph.D dissertation, MIT, Cambridge, Mass, 1995.

[12] P. Niyogi and M. M. Sondhi, "Detecting stop consonants in continuous speech," J. Acoust. Soc. Am., 111 (2), pp. 1063-1076, 2002.

[13] P. Zolfaghari and T. Robinson, "Formant analysis using mixtures of Gaussians," in Proc. ICSLP 1996, vol.2, pp. 1229-1232.

[14] M. N. Stuttle and M. J. F. Gales, "Combining a Gaussian mixture model front end with MFCC parameters," in Proc. ICSLP 2002, pp. 1565-1568.

[15] M. N. Stuttle, "A Gaussian mixture model spectral representation for speech recognition," Ph.D dissertation, University of Cambridge, 2003.

[16] B. P. Nguyen and Akagi, M., "A flexible spectral modification method on temporal decomposition and Gaussian mixture model," in Proc. Interspeech 2007, pp. 538-541.

[17] P. Zolfaghari, H. Kato, Y. Minami, A. Nakamura, and S. Katagiri, "Dynamic assignment of Gaussian components in modeling speech spectra," J. VLSI Signal Processing, pp. 45, 7-19, 2006.

[18] A. R. Jayan and P. C. Pandey, "Automated detection of speech landmarks using Gaussian mixture modeling," in Proc. Int. Symposium on Frontiers of Research on Speech and Music (FRSM) 2008, Kolkata, India, pp. 323-327.

[19] R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification, Singapore: John Wiley, 2004.

[20] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, Discrete-Time Processing of Speech Signals, New York: John Wiley, 2000.