Int. Symposium on Frontiers of Research on Speech and Music 2008 (FRSM 2008), Kolkata, India, pp. 323-327.

AUTOMATED DETECTION OF SPEECH LANDMARKS USING GAUSSIAN MIXTURE MODELING

A. R. Jayan and P. C. Pandey

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

Abstract- Landmarks in speech signal are regions with abrupt spectral variations. Automated detection of these regions is important for several applications in speech processing. Performance of landmark detection using parameters extracted from predefined spectral bands generally gets limited by speaker related spectral variability. This paper presents a landmark detection technique which adapts to the acoustic properties of speech. Parameters are extracted from Gaussian mixture modeling (GMM) of smoothed spectral envelope. A single rate of rise function, obtained from the set of GMM parameters, is used for locating landmark regions. The method was evaluated using manually labeled VCV syllables and sentences. It was possible to detect 85 % of stop release bursts in VCV syllables and 82 % in sentences, with an accuracy of 5 ms, compared to the manually located landmarks.

Keywords- Landmark detection, release burst, centroid frequency, Gaussian mixture model (GMM).

1. Introduction

Landmarks are information rich areas in speech waveform, with concentration of acoustic cues for phoneme identification. Acoustically abrupt landmarks are produced by movement of a primary articulator or by sudden changes in sound by glottal or velo-pharyngeal activity [1]. They generally coincide with regions of major spectral changes. In TIMIT database [2], about 68 % of the total landmarks are acoustically abrupt, 29 % are vocalic, and 3 % are non-abrupt as in the case of semi-vowels and vowel-to-vowel transitions.

Automated landmark detection has several applications in speech processing. Identification of landmarks can improve the accuracy and speed of segment based speech recognition systems, by reducing the search space required. Schutte and Glass [3] reported a robust landmark detection technique for sonorants using mel-frequency cepstral coefficients and support vector machines (SVMs). Sainath and Hazen [4] reported a landmark detection method based on sinusoidal model, for improving segment based speech recognition, using short-time energy and signal harmonicity as parameters. Liu [1] proposed a landmark detection algorithm for distinctive feature based speech recognition, based on energy variations in 6 spectral bands (0-0.4, 0.8-1.5, 1.2-2.0, 2.0-3.5, 3.5-5.0, and 5.0-8.0 kHz). Time derivative of maximum energy in these bands, taken every 1 ms, was used for locating stop closures and releases, fricatives, nasals, and the onsets and offsets of glottal vibration. The temporal accuracy of the landmarks detected was evaluated by comparing with manually annotated speech material from TIMIT database. Out of the total landmarks, 44 % were detected within 5 ms, 73 % within 10 ms, 83 % within 20 ms, and 88 % within 30 ms of the manual labels.

Enhancing the landmark regions by natural or synthetic methods may improve speech intelligibility. This involves precise identification of locations of transition segments, stop closure, release burst, voicing onset etc, and performing modifications specific to these sub-segments. In automatic intelligibility enhancement methods [5], [6], [7], the regions for modification are identified by a landmark detection stage. The properties of speech spectrum in the initial 10-20 ms of stop consonants contain important cues for the identification of place of articulation [8]. For modification of the burst spectrum and temporal parameters like voice onset time (VOT) and formant transition duration to disambiguate stop consonants, we need landmark detection with high temporal resolution, as the acoustic events like VOT have short duration (≈ 30 ms).

The detection rates and accuracy of landmark detectors depend on the extent to which the selected parameters represent the acoustic variations, the smoothening performed during the extraction of

parameters, and the measure used for locating the landmarks. In an earlier investigation [7] for improving identification of stop consonants, we used a modified form of Liu's landmark detection algorithm [1] for locating the boundaries of vowel-consonant and consonant-vowel transition segments. The spectrum was divided into 5 non-overlapping bands (0–0.4, 0.4–1.2, 1.2–2.0, 2.0–3.5, 3.5–5.0 kHz). Speech was sampled at 10 k samples/s, and short-time spectral analysis was performed using 512- point DFT on 6 ms Hanning windowed segments. The maximum energy in each spectral band and the centroid frequency, evaluated every 1 ms, were used as parameters for landmark detection. Rate of rise (ROR) contours of energy and centroid frequencies were computed and their geometric mean averaged across bands was used as an indicator of the overall spectral variation. Using this method it was possible to detect 60 % of release bursts in VCV syllables within 5 ms of manual labels, 80 % within 10 ms, and 94 % within 30 ms.

Maximum energy and centroid frequency derived from spectral bands with fixed boundaries represent the resonance peaks and frequencies of the vocal tract in an approximate way. Further, the fixed bands may not be able to capture spectral prominences due to speaker related spectral variability. A better approximation of resonance peaks and frequencies may improve the detection rates and temporal resolution of the landmark detector. Gaussian mixture model (GMM) provides a reliable parametric representation of smoothed spectral envelope with the effect of excitation removed, and can be used for extracting formant like features. Zolfaghari and Robinson [9] developed a formant extraction technique using GMM of cepstrally smoothened short-time spectrum. Formant tracks agreed with the tracks estimated using an LPC based formant tracker, but the bandwidths were found to be slightly broader. Using this analysis technique, they developed a formant vocoder [10]. Stuttle and Gales [11] reported performance improvement in speech recognition by combining GMM and MFCC parameters, particularly in noisy environments. GMM parameters were found to carry information complementary to MFCC parameters and were capable of improving word recognition rates by about 7 %. Omar *et al*. [12] reported improvement in phoneme recognition accuracy by about 3.5 %, by using a GMM feature set for capturing the dynamics of the speech signal at phoneme boundaries, in an HMM based speech recognition system. Lindblom and Samuelsson reported a bounded support expectation maximization algorithm (EMBS) optimized for Gaussian modeling of speech source spectrum [13].

We have investigated landmark detection using variations in GMM parameters, modeling the speech spectra. The GMM means, variances, and amplitudes may be considered to be related to formant frequencies, bandwidths, and amplitudes respectively in a parallel-formant model of speech production. The articulatory movements result in changes in the acoustic signal, and this gets captured by the GMM parameters. As the GMM fitting adapts itself to the spectral properties, the parameter tracks are smooth, giving a better representation of the spectral variations, compared to the maximum energy and centroid frequency used in [7].

2. Gaussian mixture modeling

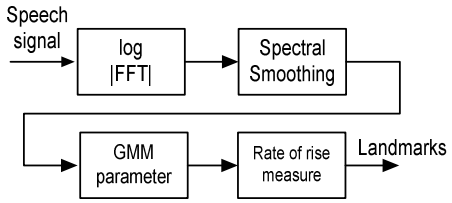

Fig. 1. Landmark detection using GMM parameters.

As shown in Fig. 1, a rate-of-rise measure taken on the GMM parameter set, modeling the smoothed spectrum, is used for landmark detection. Speech signal is sampled at 10 k samples/s and magnitude spectrum is computed using 512-point DFT on 6 ms Hanning windowed frames. The short duration window suppresses the effect of pitch harmonics in the spectrum. Frames are taken every 1 ms to track acoustic variations [1]. The magnitude spectrum is converted to dB scale, with a dynamic range of 100 dB. The harmonic structure in the spectrum is smoothed by a low pass filter with impulse response in the form of a raised cosine window [11]. Assuming the pitch to be lower than 200 Hz, a 20-point filter is used. The smoothened log magnitude spectrum $S_x(n,k)$ is approximated by a weighted sum of *M* Gaussian components, given as

$$\hat{S}_x(n,k) = \sum_{m=1}^{M} A_m(n) G\left(\mu_m(n), \sigma_m(n)\right) \quad (1)$$

where the m^{th} Gaussian component is given as

$$G(\mu_m(n),\sigma_m(n)) = 1/\sqrt{2\pi\sigma_m{}^2}\exp\left(-\frac{(x-\mu_m)^2}{2\sigma_m{}^2}\right) \quad (2)$$

with amplitude $A_m(n)$, mean $\mu_m(n)$, and standard deviation $\sigma_m(n)$ at time index *n*. The GMM parameters were computed using expectation maximization (EM) algorithm [9], [11]. Means were initialized with equal spacing along the *k* axis as $\mu_m = (m-0.5)N/(2M)$, for *N*-points DFT. Mixture weights and standard deviations were initialized with equal values of 1/*M*, and *N*/(2*M*) respectively, for all the Gaussian components.

An examination of the various GMM parameters showed that the amplitude variations were most consistent during vowel, consonant, and silence segments, and hence were used for landmark detection. The first Gaussian component models the energy variation in the lower frequency range, and was found not to be consistently related to landmarks and hence only the amplitudes of higher components were used. The parameter tracks were normalized to the range 0 to 1, and were smoothed by a 10-point median filter to remove the frame-to-frame discontinuities, without affecting the major transitions corresponding to the landmarks. Square root operation on the amplitudes, providing expansion in the lower range and compression in the upper range, was found to help in better localization of bursts and voice onsets. Normalization, median filtering, and square root operations on each Gaussian amplitude $A_m(n)$ resulted in parameter track $\alpha_m(n)$. An overall rate of rise function was calculated as the rms of the first difference of each of these tracks (excluding the first Gaussian)

$$r(n) = \left[\sum_{m=2}^{M}\left(\alpha_m(n+s)-\alpha_m(n-s)\right)^2\right]^{0.5} \quad (3)$$

For first differences, *s* = 15, corresponding to a time step of 3 ms, was used. For landmark location, the rate of rise *r*(*n*) was compared with an empirically determined threshold.

3. Results and discussion

This section presents the selection of optimum number of Gaussians required to model the smoothed spectral envelope, evaluation results using VCV syllables and sentences, and comparison with the method using maximum energy and centroid frequencies in spectral bands with fixed boundaries [7].

3.1. Estimation of number of Gaussians

The number of Gaussian components needed to model the spectrum was decided by computing the mean squared error estimates between the smoothed spectrum $S_x(n,k)$ and the Gaussian modeled spectrum $\hat{S}(n,k)$ as

$$e(n) = \sum_{k=1}^{N/2}\left(S_x(n,k)-S_x(\hat{n},k)\right)^2 \Big/ \sum_{k=1}^{N} S_x(n,k)^2 \quad (5)$$

The error was computed and averaged for 10 frames. Table 1 lists the errors for vowels (/*a*/, /*i*/, /*u*/), and fricatives (/*v*/, /*z*/, /*f*/, /*s*/). Increasing the number of Gaussian components from 1 to 2 significantly reduced the error, but not much reduction is observed for further increase in the number of components. It is also observed that GMM provides better spectral approximation for voiced sounds. As short-time spectral envelope of speech can be considered to have four significant resonances, we have used 4 components in GMM for further analysis. Figure 2 shows the windowed signal, actual, smoothed, and the Gaussian modeled spectrum using 4 Gaussians, for a 6 ms segment of vowel /a/. The waveform, wideband and Gaussian modeled spectrograms of VCV syllable /*apa*/ are shown in Fig. 3 (a), (b), and (c). Variations of Gaussian amplitudes, the rate-of-rise, and the located burst landmark for /*apa*/ are shown in Fig. 3 (d), and (e).

3.2. Evaluation using VCV syllables and sentences

VCV syllables recorded from 6 speakers (3 male and 3 female), consisting of 6 stops (/*b*/, /*d*/, /*g*/, /*p*/, /*t*/, /*k*/), in the context of 3 vowels (/*a*/, /*i*/, /*u*/) were used for evaluation. There were a total of 108 utterances (6 speakers × 6 vowel contexts × 6 stops). The locations of automatically detected stop release bursts were compared with the locations obtained manually by inspection of the waveforms and spectrograms. A comparison was made in terms of detection rates and temporal accuracy of the

presented method (M1), with the method using maximum energy and centroid frequency in fixed spectral bands (M2). The detection rates of release bursts versus the temporal accuracy in detection for both these methods are shown in Fig. 5 (a). It is observed that 85 % of the burst landmarks are detected within 5 ms, using M1, compared to 57 %, using M2. The differences between the detection rates for the two methods decreased as the time limit was relaxed. For 30 ms limit, M2 performed slightly better, there being 4 deletions in M1, and 2 deletions in M2.

Table 1 Normalized mean squared error in GMM based spectrum approximation

Phone-me	No. of Gaussian components				
	1	2	3	4	5
/a/	0.22	0.08	0.06	0.05	0.04
/i/	0.45	0.08	0.05	0.05	0.05
/u/	0.35	0.12	0.08	0.07	0.05
/v/	0.18	0.05	0.04	0.04	0.03
/z/	0.49	0.10	0.01	0.01	0.01
/f/	0.43	0.28	0.20	0.19	0.18
/s/	0.77	0.16	0.13	0.11	0.13

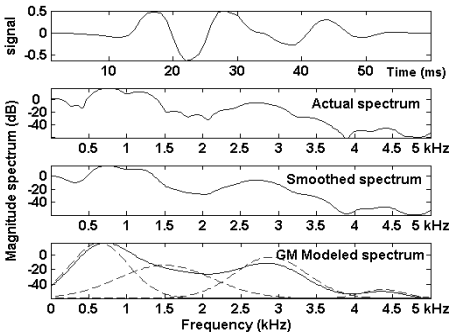

Fig. 2. Signal, actual, smoothed, and Gaussian modeled spectrum for /a/.

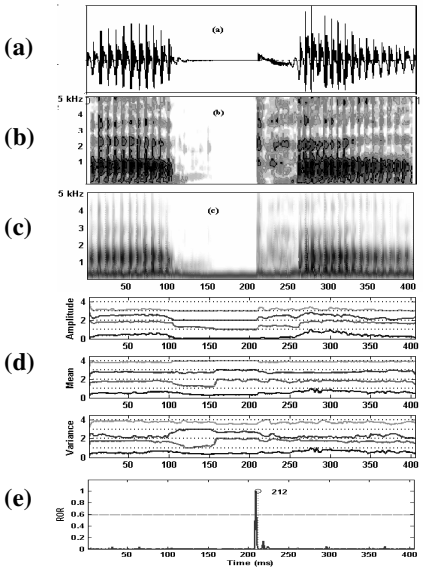

Fig. 3. (a) Signal, (b) wideband spectrogram (c) GMM spectrogram. (d) variation of parameters (e) ROR and landmark.

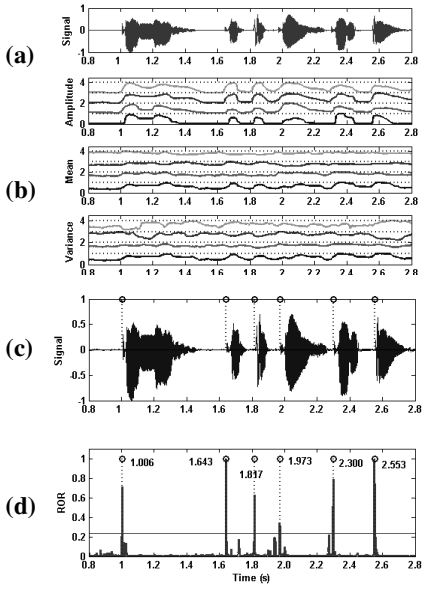

Fig. 4. (a) Sentence, (b) GMM parameter tracks (c) signal with automatically detected bursts, (d) ROR and landmarks.

The method was also evaluated using 15 Marathi sentences with 98 manually marked stop release bursts, uttered by a male speaker. Waveform of a portion of a sample sentence, Gaussian parameter tracks, ROR and located burst landmarks are shown in Fig. 4. The detection rates versus temporal accuracy for the two methods is shown in Fig. 5(b). It was observed that out of the 98 release bursts, M1 detected 81 within 5 ms of manual labels, compared to 31 detected in the same range by M2. The performance difference decreased as the time limit was relaxed. At 30 ms, M2 performed slightly better, as there were 10 deletions in M1 and 8 deletions in M2. This indicates that Gaussian amplitudes capture spectral prominences more effectively, with good temporal resolution, compared to parameters from spectral bands with fixed boundaries.

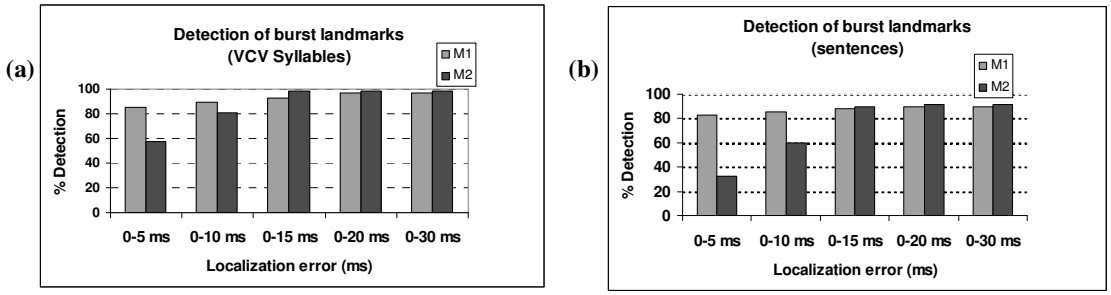

Fig. 5. Detection rates and localization errors for (a) VCV syllables (b) sentences.

4. Summary and conclusion

A landmark detection technique is presented using Gaussian parameters used for modeling the speech spectrum. The parameter extraction process involves smoothening in the spectral domain, and no smoothening is needed in the temporal domain. This makes the temporal resolution of detected landmarks high compared to the method using maximum energy and centroid frequency in fixed spectral bands. The method adapts to speaker related spectral variations. However, this method is computation intensive and further investigations are needed for adapting it for real-time applications.

References

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

[2] 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.

[3] Ken Stutte and James Glass, "Robust detection of sonorant landmarks," in *Proc. Interspeech 2005*, Lisbon, Portugal, 2005, pp. 1005-1008.

[4] T. N. Sainath, and T. J. Hazen., "A sinusoidal model approach to acoustic landmark detection and segmentation for robust segment based speech recognition," *in Proc ICASSP 2006*, Toulose, France, 2006, pp. I-525-528.

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

[6] M. D. Skowronski and J. G. Harris, "Applied principles of clear and Lombard speech for automated intelligibility enhancement in noisy environments," *Speech Communication*, 48, no.5, pp. 549-55, 2006.

[7] 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.

[8] D. O. Shaughnessy, *Speech Communication*: *Human and Machine*. New York: Addison Wesley, 1987.

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

[10] P. Zolfaghari and T. Robinson, "A formant vocoder based on mixtures of Gaussians," *in Proc. IEEE ICASSP 1997*, vol. 2, pp.1575-1578, 1997.

[11] 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, Sep. 2002.

[12] M. K. Omar., M. H. Johnson, S. Levinson, "Gaussian mixture models of phonetic boundaries for speech recognition," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, Dec. 2001, pp. 33-36.

[13] J. Lindblom and J. Samuelsson, "Bounded support Gaussian mixture modeling of speech spectra," *IEEE Trans. Speech Audio Processing*, vol. 11, no. 1, pp. 88-98, Jan. 2003.