INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Excitation Source Features for Improving the Detection of Vowel Onset and Offset Points in a Speech Sequence

Gayadhar Pradhan, Avinash Kumar and S Shahnawazuddin

Department of Electronics and Communication Engineering
National Institute of Technology Patna, India.
(gdp, k.avinash, s.syed)@nitp.ac.in

Abstract

The task of detecting the vowel regions in a given speech signal is a challenging problem. Over the years, several works on accurate detection of vowel regions and the corresponding vowel onset points (VOPs) and vowel end points (VEPs) have been reported. A novel front-end feature extraction technique exploiting the temporal and spectral characteristics of the excitation source information in the speech signal is proposed in this paper to improve the detection of vowel regions, VOPs and VEPs. To do the same, a three-class classifiers (vowel, non-vowel and silence) is developed on the TIMIT database using the proposed features as well as mel-frequency cepstral coefficients (MFCC). Statistical modeling based on deep neural network has been employed for learning the parameters. Using the developed three-class classifier, a given speech sample is then forced aligned against the trained acoustic models to detect the vowel regions. The use of proposed feature results in detection of vowel regions quite different from those obtained through the MFCC. Exploiting the differences in the evidences obtained by using the two kinds of features, a technique to combine the evidences is also proposed in order to get a better estimate of the VOPs and VEPs.

Index Terms: vowel recognition system, vowel onset point, vowel end point.

1. Introduction

The instants of starting and ending of a vowel region in a speech sequence are referred to as the vowel onset point (VOP) and end point (VEP), respectively [1, 2, 3, 4]. Due to their larger amplitude, periodicity and longer duration [5], the vowels happen to be the prominent regions in a speech signal. The changes in the excitation source and the vocal tract system are reflected at these instants. These aspects of speech production have been exploited in the existing methods for detecting the vowel regions and their corresponding VOPs and VEPs [2, 3, 4, 6]. The accurate detection of the vowel regions, the VOPs and VEPs are employed in extracting features that are robust to the environmental degradation. Such features are preferred in the development of various speech-based applications [7, 8, 9, 10, 3].

It is well known that the transition characteristics of the vowels vary with the context of the spoken utterance [5, 11]. For example, the transition from a fricative to vowel is completely different from that for a semivowel to vowel transition. Due to the similarities in the production characteristics of the vowels and semivowels, accurate detection of semivowels, VOPs and VEPs for the semivowel-vowel clusters and the diphthongs become challenging. The existing methods based on the transition characteristics are generally threshold dependent. In general, the VOPs and VEPs are detected by convolving the features characterizing the temporal variations with a first order Gaussian difference (FOGD) operator within a region that is 100 ms in duration [2, 3, 4, 6]. Next, the convolved output is employed as the evidence for the detection of the VOPs and VEPs. Since the convolved output mainly depends on the 100 ms regions under consideration, most of the weak transitions are smoothed out. On the other hand, convolution in a smaller region leads to spurious detections. To address this shortcoming, a threshold independent vowel detection system should be developed by statistically modeling the vocal tract system and excitation source and their transient behaviors. Motivated by this, an excitation-based feature is proposed in this work to extract the temporal and spectral characteristics of the source information.

In order to accurately detect the vowel regions, acoustic modeling based on deep neural network (DNN) [12] is employed in this study. For detecting vowels, a three-class classifier (vowel, non-vowel and silence) is developed. The speech sound units excluding the vowels are termed as non-vowels in this study. Separate classifiers are developed on the TIMIT database [13] using the proposed excitation features and the conventional mel-frequency cepstral coefficients (MFCC) [14]. The given test speech sample is forced aligned against the corresponding acoustic models to detect the vowel regions. To determine the impact of semivowel and nasal sound units on the detection of vowel regions, the correctly detected and spurious vowel regions are also analyzed in detail. During our experimental evaluations, the vowel-regions detected using the MFCC and the proposed features were observed to be quite different. Motivated by that, a scheme to combine the obtained evidences is also proposed in this work. Combining the evidences is found to significantly improve the accuracy with which the vowel regions and their corresponding VOPs and VEPs are detected.

The rest of the paper is organized as follows: The proposed excitation source features is discussed in Section 2. The experimental evaluations and a detail analysis of the detected vowel regions using the VOPs and VEPs is presented in Section 3. Finally, the paper is concluded in Section 4.

2. Excitation source features

Vowels in speech signal are produced by the vibration of the vocal folds [5]. Due to a sudden closure of the vocal folds during the production of vowels, the excitation is observed to be impulse like. The strength of excitation in these regions is relatively higher when compared to other consonants. In the existing approaches based on excitation source, the linear prediction (LP) residual signal is processed only in the temporal domain. The variation of the excitation strength in different frequency bands is completely neglected. It is to note that, the temporal and spectral characteristics of the excitation source vary in different frequency bands [15]. This fact is exploited to extract

http://dx.doi.org/10.21437/Interspeech.2017-135

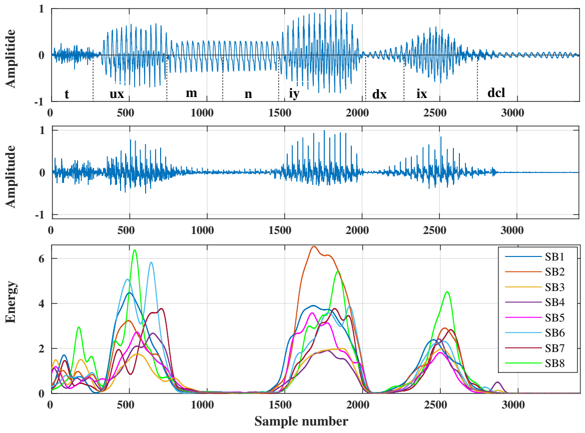

Figure 1: *Top panel shows a speech segment along with the reference marking for different sound units (dotted lines). Middle panel shows the corresponding LP residual signal while bottom panel shows the spectral energy in the residual signal corresponding to the* 8 *frequency bands.*

excitation features in this study. The frequency bands considered in this work are derived by splitting the analysis range of 0-4 kHz (since speech data used in this work is sampled at 8 kHz rate) into 8 non-overlapping bands of bandwidth 500 Hz each. This is done by filtering the LP residual signal through a bank of band-pass filters each having a bandwidth of 500 Hz. Narrowing the bandwidth does not provide more discrimination since there is not much variation in the dynamic range of the LP residual spectrum. At the same time, increasing the number of filters results in an increase in the number of coefficients in the feature vector. This, in turn, increases the complexity of the classifier. On the other hand, increasing the bandwidth results in a degradation of the discriminative property of the features. The choice of 500 Hz is found to be more suitable through preliminary studies performed on a development data.

The variation of spectral energy in the considered frequency bands is shown in Figure 1 (bottom panel). It is to note that, variation in the spectral energies for the considered frequency bands is greater in the case of vowels (e.g. /ux/, /iy/ and /ix/). On the other hand, variation is lesser for the non-vowel regions (e.g. /t/). The sub-band energies for the nasal units (e.g. /m/ and /n/) are observed to be much smaller in comparison to those for the vowels. This is mainly due to the nature of the LP residual signal (middle panel). The features for modeling the excitation source information within vowels may be obtained by considering these variations in different sub-bands. As reported in earlier works [15, 16], the energy in the LP residual signal is mostly concentrated around the instants of glottal closure. As the nature of the excitation in the vowel regions is different from that for the non-vowel regions, another set of features can be derived by processing the 2 ms portions around the significant excitation in the residual signal. The sequence of steps involved in the extraction of the proposed features are described next.

First, short-time analysis of speech signal is done considering a frame size of 20 ms with a frame-shift of 10 ms. Next, the speech signal is processed through the following sequence of steps for estimating the temporal and spectral characteristics of the excitation source information in the vowel regions:

- The instants of significant excitation or the glottal closure instants (GCIs) are detected using zero frequency filtering (ZFF) [17, 3].
- For each 20 ms frame of speech, 10$^{\text{th}}$-order LP analysis is performed to estimate the linear prediction coefficients (LPCs). A time-varying inverse filter is constructed using the LPCs and the speech signal is processed by the inverse filter to derive the LP residual signal.
- Next, the LP residual signal is split into 8 sub-bands by filtering through a bank of 8 non-overlapping filters each having a bandwidth of 500 Hz. Feature vectors are then extracted by computing the spectral energy in each sub-bands considering a frame size of 20 ms with 50% overlap. This results in an 8-dimensional feature vector per frame capturing the spectral variations in the LP residual.
- Similarly, for each block of 20 ms, the GCI locations within that frame are identified using ZFF. If one or more GCIs are found within an analysis frame, for all the sub-bands of the LP residual, those regions that are 2 ms to the right of each GCI are identified by anchoring the GCIs. The temporal energies are then computed within those regions. To do so, Hamming windowed regions that are 2 ms in duration are considered. Finally, the average energy for all the GCIs within that frame is computed. If no GCIs are found within the analysis frame under consideration, the temporal energy for all the sub-bands are computed by considering the central portions of the analysis frame with a duration of 2 ms. This results in another 8-dimensional feature vector per frame capturing the temporal variations in the LP residual .
- Logarithm of the spectral and temporal energies are taken to reduce the dynamic range. Finally, the derived set of features are concatenated to obtain a 16-dimensional base feature vector.

Since the analysis frames considered for the computation of the spectral and the temporal energies are quite different in duration, the derived feature vectors turn out to be different.

3. Experimental evaluation

The Kaldi speech recognition toolkit [18] was used to develop the vowel-non-vowel detection system employing DNN-based acoustic modeling. The system development and evaluation was done on the TIMIT corpus [13]. The speech corpus was split into orthogonal sets following the standard Kaldi recipe. Speech data from 462 speakers comprising of 3696 utterances was used to train the acoustic model parameters. The training/test transcription was modified to represent the possible vowels in the database as a single class. The non-vowels were grouped together to represent the second class. The silence, short-pause and other non-speech units (fillers) were grouped together to represent the third class (silence). The test set comprised of 192 utterances from 24 speakers. A development set consisting of 400 utterances from 50 speakers was used for optimizing the tunable parameters. All the experiments reported in this paper were performed on 8 kHz re-sampled data to simulate telephone-based speech interface.

For each short-time frame of speech, 13-dimensional base MFFC features ($C_0 - C_{12}$) were computed employing 23-channel mel-filterbank. Time-splicing of the base MFCC, considering a context size of 9 (± 4), was done making the total feature dimension equal to 117. The dimensionality of the derived time-spliced features was then reduced to 40 using linear discriminant analysis. This was followed by further de-correlation using maximum likelihood linear transform. Speaker normalization using feature-space maximum likelihood linear regression (fMLLR) [19] was employed next to further improve the

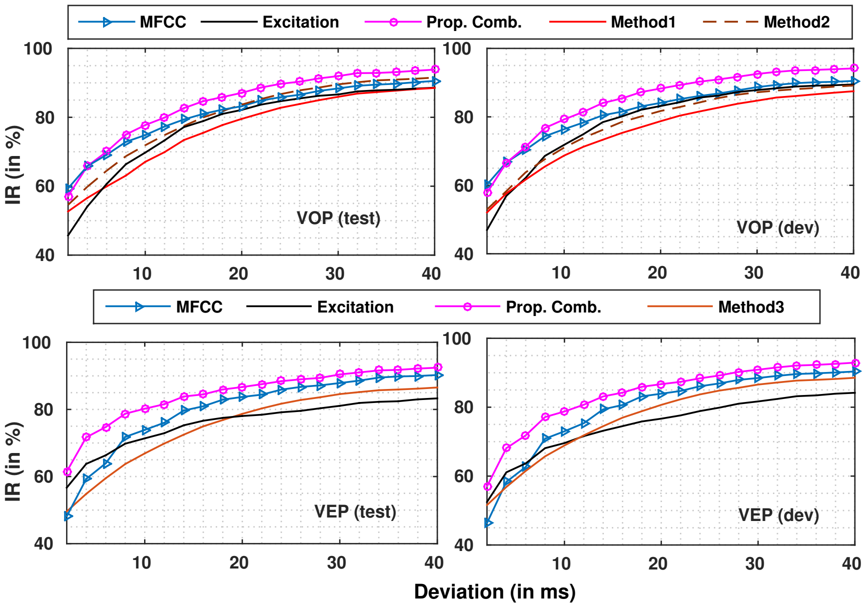

Figure 2: *The IR profiles for the VOPs/VEPs with respect to DNN-HMM systems developed using the MFCC and excitation features, respectively. The predefined deviation is varied from* 2 *ms to* 40 *ms in steps of* 2 *ms. Also shown is IR profile for the proposed approach of combining the evidences. The IR profiles for the existing techniques are also shown.*

Table 1: *Classification error rates for the DNN-based three-class classifiers developed using the conventional MFCC and the proposed excitation features.*

Data	Error (in %)	
set	MFFC	Excitation
Dev.	11.48	15.27
Test	12.11	16.22

performance as suggested in [20]. Similarly, in the case of the proposed features, the 16-dimensional base features were spliced in time and 40-dimensional fMLLR-normalized features were derived. For learning the DNN parameters, the 40-dimensional time-spliced features with fMLLR-based normalization were further spliced over 5 frames to the left and right of the central frame. Six layers of restricted Boltzmann machine (RBM) consisting of 1024 hidden nodes were used in deep belief (DBN) network pre-training. The final DNN-HMM system was developed using sequential discriminative training employing minimum phone error (MPE) criterion.

The recognition performances for the three-class classifier developed on the MFCC features are given in Table 1. The error rates were computed in the same way as the word error rates with the possible words being vowel, non-vowel and silence. The recognition performances for the proposed features are also enlisted in Table 1. From the presented error rates, it can be concluded that the proposed excitation features can be used for developing the classification system even though such a classifier will be inferior to that developed using the MFFC features. Motivated by these results, the classification systems developed using the two kinds of features were employed for the detection of vowel regions in a given speech signal. From the studies presented in the following sub-section, it will be noted that the two kinds of features result in different and, at times, non-overlapping vowel regions for the same speech sequence. Consequently, the evidences obtained from these two features are combined to get a better estimate of the vowel regions.

3.1. Detection of vowel onset and end points

The test data was forced-aligned with respect to the trained acoustic models under the constraints of the first-pass hypothesis to generate the frame-level alignments required to detect the vowel regions. The starting and the ending points of the detected vowel regions are marked as the VOPs and the VEPs, respectively. Using the manual markings given in the database as the reference, the performances of the detected VOPs and VEPs are measured using the following metrics:

Identification rate (IR**):** The number of times the reference VOPs/VEPs match with the detected VOPs/VEPs within the pre-defined deviation (in ms) measured in percentage.

Spurious rate (SR**):** The percentage of detected VOPs/VEPs, which are detected outside the vowel regions.

The IR profiles for VOP and VEP detection with respect to the DNN-HMM-based three-class classifier developed using the MFCC as well as the proposed features are shown in Figure 2. For both the test as well as development (dev.) sets, the use of MFCC is noted to be better.

3.2. Combining the evidences

In order to improve the IR and SR, we explored the possibility of combining the vowel evidences obtained by using MFFC and excitation features. Furthermore, since the IR for the MFCC are found to be better than the excitation features, a higher weighting should be given to the evidence obtained using the MFCC. To achieve this, a method is proposed for combining the evidences and is discussed next.

First, the detected evidences are classified into two broad categories, i.e., the overlapping and non-overlapping ones. If there happens to be a minimum overlap of 70% between the evidences obtained by using the MFCC and the excitation features, then those are considered as overlapping evidences. At the same time, if the overlap is less than 70% then those are referred to as non-overlapping. The starting and ending points

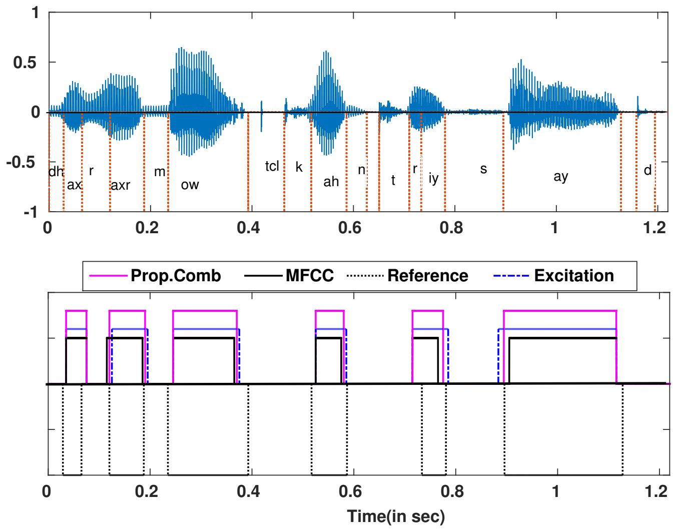

Figure 3: *Illustrations depicting the effectiveness of the proposed method for combining the evidences. A segment of speech,* ***"the remote countryside"****, with reference markings for the sound units is shown in top panel. In the bottom panel, the detected vowel evidences obtained by using the MFCC, the excitation features and the proposed approach for combining the evidences are shown.*

of the combined evidence in the case of overlapping evidences are decided as follows:

- If the starting point of an evidence detected by using the excitation feature falls within three analysis frames (240 samples) of the starting point of the evidence detected by using MFCC, then the staring point of final evidence is considered as the mean of those two locations. Otherwise, the starting point of the vowel evidence detected by the MFCC is considered as the starting point.
- Similar steps are also followed for deciding the end points in the case of overlapping evidences.

For both kinds of features, the non-overlapping evidences that are a minimum of 100 ms in duration are identified and preserved in the final evidences without any modification. Those non-overlapping evidences that are less than 100 ms in duration are treated as spurious detections and are eliminated.

The evidences for the vowel regions obtained by using the MFCC, the excitation features and the proposed combination with respect to the trained acoustic models are shown in Figure 3. The differences in the obtained evidences may probably be attributed to the fact that the excitation features and MFCC model the few frames near vowel transitions quite differently. On comparing the detected vowel evidences with the references, the proposed method of combining the evidences helps in detecting the vowel regions more accurately in contrast to those detected using each of the individual features. The same is depicted in terms of IR profiles shown in Figure 2.

Finally, we have compared the proposed approach with some of the existing techniques. In this regard, two different state-of-the-art methods for VOP detection are considered and are referred to as Method 1 [2] and Method 2 [3], respectively. In the case of VEP detection, the existing approach reported in [4] is considered for comparison and is referred to as Method 3. For proper comparison, the parameters for the computation of the features and the evidences in the case of existing techniques are chosen to be the same as described in those original works [2, 3, 4]. The IR profiles for VOPs and VEPs detected using the existing approaches are shown in Figure 2. For both the cases (VOPs/VEPs), the proposed approach for combining the evidences is noted to be much superior to the existing techniques. The SR obtained by using the MFCC, excitation features, proposed combination of evidences and the existing techniques are given in Table 2. It is to note that the proposed technique for combining the evidences is superior to the existing methods in terms of SR as well.

Table 2: *Comparison of SR for VOP and VEP detection obtaining by using the MFCC, excitation features, proposed technique for combing the evidences and the existing methods.*

VOP	SR (in %)			
	MFCC	Excitation	Method 2	Proposed
Test	5.9804	5.1260	6.60	5.2114
Dev	5.0020	4.3742	5.76	4.5362
VEP	MFCC	Excitation	Method 3	Proposed
Test	5.9804	9.3550	7.57	6.2367
Dev	5.3868	9.7205	6.87	5.9336

4. Summary and Conclusions

In this paper, a front-end feature extraction method is proposed to extract the temporal and the spectral characteristics of the excitation source. The excitation features are used to develop a DNN-based three-class classifier for detecting the vowel regions, VOPs and VEPs. Another three-class classifier is developed using the conventional MFCC as well. Finally, a novel method is proposed to combine the evidences obtained using the MFCC and the excitation feature vectors to enhance the detection of vowel regions, VOPs and VEPs. The proposed combination scheme is observed to provide better IR with an overall reduction in SR.

5. Acknowledgements

The authors wish to thank EMST Lab, Department of EEE, IIT Guwahati for sharing the speech database used in this study.

6. References

[1] D. J. Hermes, “Vowel onset detection,” *Journal of the Acoustical Society of America*, vol. 87, no. 2, pp. 866–873, February 1990.

[2] S. R. M. Prasanna, B. V. S. Reddy, and P. Krishnamoorthy, “Vowel onset point detection using source, spectral peaks, and modulation spectrum energies,” *IEEE Transactions on audio, speech, and language processing*, vol. 17, no. 4, pp. 556–565, May 2009.

[3] A. Vuppala, J. Yadav, S. Chakrabarti, and K. S. Rao, “Vowel onset point detection for low bit rate coded speech,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1894–1903, August 2012.

[4] J. Yadav and K. S. Rao, “Detection of vowel offset point from speech signal,” *IEEE Signal Processing Letters*, vol. 20, no. 4, pp. 299–302, April 2013.

[5] K. N. Stevens, *Acoustic Phonetics*. The MIT Press Cambridge, Massachusetts, London, England, 2000.

[6] K. Vuppala, K. S. Rao, and S. Chakrabarti, “Improved vowel onset point detection using epoch intervals,” *AEU-International Journal of Electronics and Communications*, vol. 66, no. 8, pp. 697–700, August 2012.

[7] N. Fakotakis and J. Sirigos, “A high performance text independent speaker recognition system based on vowel spotting and neural nets,” in *International Conference on Acoustics Speech and Signal Processing*, vol. 2, May 1996, pp. 661–664.

[8] S. R. M. Prasanna, S. V. Gangashetty, and B. Yegnanarayana, “Significance of vowel onset point for speech analysis,” in *International Conference on Signal Processing and Communications*, July 2001, pp. 81–88.

[9] L. Mary and B. Yegnanarayana, “Extraction and representation of prosodic features for language and speaker recognition,” *Speech communication*, vol. 50, no. 10, pp. 782 –796, April 2008.

[10] G. Pradhan and S. R. M. Prasanna, “Speaker verification by vowel and nonvowel like segmentation,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 4, pp. 854–867, April 2013.

[11] S. R. M. Prasanna and G. Pradhan, “Significance of vowel-like regions for speaker verification under degraded condition,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2552–2565, May 2011.

[12] G. E. Hinton, L. Deng, D. Yu, G. Dahl, A. R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, “Deep neural networks for acoustic modeling in speech recognition,” *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, November 2012.

[13] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, *TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1*. Linguistic Data Consortium, December 1993, vol. 33.

[14] S. B. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE Trans. Acoustics, Speech and Signal Process.*, vol. ASSP-28, no. 4, pp. 357–366, August 1980.

[15] B. Yegnanarayana, C. Avendano, H. Hermansky, and P. S. Murthy, “Speech enhancement using linear prediction residual,” *Speech Communication*, vol. 28, no. 1, pp. 25–42, May 1999.

[16] B. Yegnanarayana and P. S. Murthy, “Enhancement of reverberant speech using LP residual signal,” *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 3, pp. 267–281, May 2000.

[17] K. S. R. Murthy and B. Yegnanarayana, “Epoch extraction from speech signals,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1602–1613, November 2008.

[18] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, “The kaldi speech recognition toolkit,” in *Workshop on automatic speech recognition and understanding*, December 2011.

[19] V. Digalakis, D. Rtischev, and L. Neumeyer, “Speaker adaptation using constrained estimation of gaussian mixtures,” *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 5, pp. 357–366, September 1995.

[20] S. P. Rath, D. Povey, K. Vesel, and J. Cernock, “Improved feature process. for deep neural networks.” in *INTERSPEECH*, August 2013, pp. 109–113.