International Journal of Computer and Communication Technology

Volume 4 | Issue 1 Article 4

January 2013

Significance of Vowel Onset Point Information for Speaker Verification

Gayadhar Pradhan
gayadhar@iitg.ac.in

S. R. Mahadeva Prasanna
prasanna@iitg.ernet.in

Recommended Citation
Pradhan, Gayadhar and Prasanna, S. R. Mahadeva (2013) "Significance of Vowel Onset Point Information for Speaker Verification," *International Journal of Computer and Communication Technology*: Vol. 4 : Iss. 1 , Article 4.
DOI: 10.47893/IJCCT.2013.1164
Available at: https://www.interscience.in/ijcct/vol4/iss1/4

Significance of Vowel Onset Point Information for Speaker Verification

Gayadhar Pradhan, S. R. Mahadeva Prasanna
gayadhar@iitg.ac.in, prasanna@iitg.ernet.in

Abstract— This work demonstrates the significance of information about vowel onset points (VOPs) for speaker verification. VOP is defined as the instant at which the onset of vowel takes place. Vowel-like regions can be identified using VOPs. By production, vowel-like regions have impulse-like excitation and therefore impulse-response of vocal tract system is better manifested in them, and are relatively high signal to noise ratio (SNR) regions. Speaker information extracted from such regions may therefore be more discriminative. Due to this better speaker modeling and reliable testing may be possible using the features extracted from vowel-like regions. It is demonstrated in this work that for clean and matched conditions, relatively less number of frames from vowel-like regions are sufficient for speaker modeling and testing. Alternatively, for degraded and mismatched conditions, vowel-like regions provide better performance.

Keywords— VOP, vowel-like region, speaker information, speaker verification

I. INTRODUCTION

Speaker verification (SV) validates the identity claim of a person [1]. A SV system is expected to accept the claim only from genuine persons and reject the claim from impostors [2]. The performance of SV is therefore measured in terms of how many genuine trials are rejected, given by false rejection rate (FRR) and how many impostor trials are accepted, given by false acceptance rate (FAR). The value for which FRR is equal to FAR is meaningfully termed as equal error rate (EER). For a good SV system, both FRR and FAR should be low, indicated in terms of low EER [2]. The performance of a SV system, like any other pattern recognition task, depends on the quality of incoming speech signal, extracted features and modeling. For given conditions of collecting speech data, feature extraction and modeling, the performance of SV system can be further improved by selecting only those speech regions that are more speaker discriminative. This can be achieved using the knowledge of vowel onset point (VOP). VOP helps in identifying vowel-like regions that are high Signal-to-Noise Ratio (SNR) regions from the production perspective of different speech sounds. Hence they may be more speakers discriminative and exploring this aspect is the focus of this work.

VOP is defined as the instant at which the onset of vowel takes place [3]. The typical cases in which VOP occurs include isolated vowel, consonant vowel (CV) and consonant cluster vowel (C^nV, where n > 1). If we have a method for the detection of VOPs from the speech signal, then vowel regions can be identified. If the VOP detection method is not perfect (i.e., 100% performance), then the errors are manifested in terms missing and spurious VOPs, and also the resolution with which VOPs are detected [4]. Some of the vowel regions may not be detected due to the missing VOPs. Similarly, some of the non-vowel regions may be hypothesized as vowel regions due to the spurious VOPs. The poor resolution for VOP detection leads to either missing some initial portion of the vowel or hypothesizing some preceding region as vowel. All these errors may not be very critical in the context of speaker verification as compared to speech recognition. Thus a method for VOP detection that provides reasonably good performance, even though not perfect, may suffice for the speaker verification task. The VOPs detected from a VOP detection method having all the errors mentioned above can be used for identifying vowel-like regions and not always only vowel regions. Since majority of them are vowel regions, the observations made using vowel-like regions are also valid for the case of only vowel regions.

The major excitation that provides speaker characteristics to the speech signal is the vibration of vocal folds [5]. From the excitation source perspective, vowel-like regions are produced using the vocal folds vibration and hence may have relatively more speaker information compared to non-vowel-like regions. Vowel-like regions are produced by exciting the vocal tract system using impulse-like excitation due to the sudden closure of vocal folds. Due to impulse-like excitation, the impulse response of the vocal tract system may be better manifested and hence more speaker-specific. Vowel-like regions are produced by keeping the vocal tract in an open configuration which offers relatively less obstruction for the air flow and hence

Gayadhar Pradhan and S. R. Mahadeva prasanna are with the Indian Institute of Technology, Guwahati - 781039, India.
Corresponding author phone: 0361-2582545, e.mail: gayadhar@iitg.ernet.in

high energy or high SNR regions. Therefore if we have a method for detecting vowel-like regions and use speaker information from such regions, then better speaker modeling as well as reliable testing may be made. This may help in reducing the amount of data for training and testing, and also increasing the robustness in degraded and mismatched conditions.

In the existing speaker verification systems, speech regions are separated out from the silence regions based on energy threshold, and features from the speech regions are used for modeling and testing. In the proposed approach, vowel-like regions are separated out from the non-vowel-like- regions based on the knowledge of VOP, and features from the vowel-like regions are used for modeling and testing. Suppose if clean speech collected in matched condition is used, then the proposed approach may provide better performance in terms of requirement of data. That is, it may provide equal or better performance using relatively less amount of speech data from vowel-like regions. Alternatively, if degraded speech collected in mismatched condition is used, then the proposed approach may provide better performance in terms of EER.

The rest of the paper is organized as follows: A method for VOP detection is described in Section II. Speaker verification system using vowel-like regions is described in Section III. The experimental studies are described in Section IV. The experimental results are discussed in Section V. Summary of the present work and scope for future work are mentioned in Section VI.

II. VOP DETECTION

In the present work VOP refers to the instant at which the onset of vowel takes place [3]. There are several methods proposed in the literature for VOP detection [3]. The present work uses a recent method based on the excitation source information. The motivation behind this choice is its better discriminability at the VOP and hence better performance. This is because; most of the VOP detection methods are based on short term energy computed either in time or frequency domain. The VOPs are hypothesized as significant changes in energy values. Even though this is a good feature, there are several cases like nasal CV units where changes in the excitation source characteristics may be more crucial for detecting the VOP [6].

The VOP detection method using the excitation source information involves the following steps: The speech signal is processed in blocks of 20 ms with a shift of 10 ms. For each 20 ms block, 10th order LP analysis is performed to estimate the linear prediction coefficients (LPCs) [7]. The time-varying inverse filter is constructed using these LPCs. The speech signal is passed through the inverse filter to extract the LP residual signal. The time varying nature of excitation source characteristic is further enhanced by computing Hilbert envelope of the LP residual [8]. The Hilbert envelope $h_e(n)$ of the LP residual $e(n)$ is defined as $h_e(n) = \sqrt{e^2(n) + e_h^2(n)}$, where $e_h(n)$ is the Hilbert transform of e(n). For every 5 ms block with one sample shift, the maximum value of the Hilbert envelope of LP residual is noted to construct smoothed excitation contour. The change in the excitation characteristics at the VOP event is detected by convolving the smoothed excitation contour with a first order Gaussian differentiator (FOGD) of length 100 ms (800 for 8 kHz) and standard deviation as one sixth of the window length (134 for 8 kHz). This convolved output is termed as VOP evidence plot using excitation source information. The peaks in the convolved output represent the locations of the VOPs and are selected by finding the maximum value between two successive positive to negative zero crossing with some threshold to eliminate the spurious ones.

A. Performance of VOP detection algorithm

The VOP detection method is evaluated using 60 speakers data from the TIMIT database for the sentence Don't ask me to carry an oily rag like that. The starting instants of vowel phonemes from the manual markings available in the database are used as the reference VOPs. The performance measured in terms of average error rate (AER) is given in Table 1. For comparison the results of a short term energy (STE) based method is also given in the table. As it can be observed the performance of the excitation source information is better both in terms of performance and also resolution. The error of the VOP detection method is not zero in terms of AER and hence as mentioned earlier, the regions detected using VOPs from this method will be termed as vowel-like regions.

Table 1: Performance of VOP detection methods using Excitation source information.

Method	**HYPO VOPs**	**DET VOPs in % (within ms)**				**MISS VOPs**	**SPU VOPs**	**AER**
		±10	**±20**	**±30**	**±40**			
ESI	827	57.8	76.3	84.2	87.4	12.59	23.58	18.08

STE	751	47.2	66.1	73.4	78.8	21.16	24.10	22.63

III. SPEAKER VERIFICATION USING VOWEL-LIKE REGIONS

A. Database

We have used a subset of IITG multi-variability (MV) speaker recognition database [9] developed in-house for initial studies, and the complete NIST-2003 Speaker Recognition database for evaluation on a standard database [10]. IITG-MV database is collected in a set up having five different sensors, two different environments, two different languages and two different styles. The five different sensors include headphone microphone mounted close to the speaker, inbuilt tablet PC microphone, two mobile phones and one digital voice recorder. Except for headphone microphone, all the other four sensors are placed at a distance of about two-three feet from the speaker. Speech was recorded simultaneously over these sensors and sampled at 8 kHz and stored with 16 bits/sample resolution. The recording was done in two different environments, namely, office and hostel rooms. The recording was done in two languages, namely, English and favorite language of the speaker which happens to be one of the Indian languages like Hindi, Telugu, Kannada, Oriya, and so on.

B. Detection of Vowel-like regions

As described in the Section II, VOPs are determined using the excitation source information derived from the speech signal. Using each hypothesized VOP as the anchor point, 100 ms regions right to the VOPs are marked as vowel-like regions. In case of speaker verification using vowel-like regions, features derived only from these regions are used for training and testing. Alternatively, in case of speaker verification using conventional approach, regions identified based on energy threshold are used.

C. Feature Extraction

In the training and testing process, the speech signal is processed in frames of 20 ms duration at 10 ms frame rate. For each 20 ms Hamming windowed frame, mel-frequency cepstral coefficients (MFCC) are calculated using 22 logarithmically spaced filter banks [11]. The first 13 coefficients excluding zeroth coefficient value are used as a feature vector. Delta (Δ) and delta-delta ($\Delta\Delta$) of MFCC are computed using two preceding and two succeeding feature vectors from the current feature vector. Thus the feature vector will be of 39 dimensions with 13 MFCC, 13 Δ MFCC and 13 $\Delta\Delta$ MFCC.

D. Parameter normalization

The feature vectors are normalized to fit a zero mean and unit variance distribution. However, when there is not much variability in the recording sensor and environment, the blind deconvolution like cepstral mean subtraction (CMS) seem to reduce the performance [12]. Hence, in the present work we use only cepstral variance normalization (CVN) for sensor matching experiments of IITG-MV database and CMS with CVN for sensor mismatched experiments of IITG-MV database and NIST-2003 database.

E. Speaker Modeling

The main motivation of this work is to study the discriminative information present in vowel-like regions for speaker modeling and testing. Except for deriving frames from vowel-like regions, there is no difference in the steps of speaker verification system development. Hence, for speaker modeling the extensively used Gaussian Mixture Modeling (GMM) is employed [12].

IV. EXPERIMENTAL STUDIES

In order to compare the performance obtained using vowel-like regions, we have developed another speaker verification system based on energy threshold (0.06 × average energy) which is termed as baseline system. The only difference between baseline system and proposed system lies in the selection of speech frames during training and testing process. In the baseline system the speech frames are selected by using an energy threshold and in the proposed case using vowel-like regions.

For the present work, we consider 100 speakers set of IITG MV database, which include 75 male speakers and 25 female speakers. The initial 2 minutes of speech data recorded in the first session is used for building the models. For each speaker, 10 speech segments between 30-45 sec duration from the second session are taken as test utterances. Therefore for 100 speakers set there are in total 1000 test trials. In the testing process, each test segment is tested against 11 models; out of which one is genuine model and rest are impostor models. Out of the five sensors, speech recorded over digital voice recorders (D01), due to its high sensitivity, is worst affected by environmental noise like air conditioner, fan sound and room reverberation. The speech recorded in the headphone microphone (H01) is more clean

compared to other sensors. Accordingly, the speech recorded in D01 is considered as noisy speech and speech recorded in H01 is considered as clean speech. Keeping the language as English and conversational style, three experiments are conducted on IITG-MV database as follows:

1. ***Clean and sensor matched condition*:** Speech recorded over sensor H01 is used for training and testing.
2. ***Noisy and sensor matched condition*:** Speech recorded over sensor D01 is used for training and testing.
3. ***Noisy and sensor mismatched condition*:** Speech recorded over H01 is used for training and speech recorded over D01 is used for testing.

Finally, the performance of the system is also evaluated on complete NIST-2003 speaker recognition database.

V. RESULTS AND DISCUSSIONS

Table 2: Number of frames used for training and testing in baseline system and speaker verification system using vowel-like regions.

Data set	Baseline			Vowel-like regions		
	Avg	Max	Min	Avg	Max	Min
H01 Train	7210	4474	9112	3517	1990	4977
H01 Test	2725	1197	3847	1269	440	1887
D01 Train	10277	8301	10791	3940	1052	5619
D01 Test	3953	2805	4191	1385	400	2038
Nist-2003 Train	6070	2085	8151	3307	897	4803
Nist-2003 Test	1621	144	2984	896	84	1845

Table 3: Performance of speaker verification system using vowel-like regions.

Speech data	Equal error rate (EER)			
	16	32	64	128
Clean & sensor matched	8.1	7.8	7.4	7.5
Noisy & sensor matched	18.2	18	18.6	19.2
Noisy & sensor mismatched	29.4	29.4	30.3	32.2
NIST-2003	19.2	18.7	19.7	20.3
NST-2003 (fixed no. of frames)	-	18.95	-	-

Table 4: Performance of baseline speaker verification system.

Speech data	Equal error rate (EER)			
	16	32	64	128
Clean & sensor matched	9.5	8.7	7.5	7.2
Noisy & sensor matched	21.5	20.8	20.2	19.9
Noisy & sensor mismatched	33.7	32.8	32.7	32.1
NIST-2003	19.8	19.5	18.7	18.6
NST-2003 (fixed no. of frames)	-	22.54	-	-

For each set of data, the average number of frames used for training and testing of baseline system and speaker verification system using vowel-like regions are given in Table 2. The table also contains the minimum and maximum number of frames used for training and testing. The average number of frames used for training and testing in baseline system is around two times more than vowel-like regions. The GMM is a statistical classifier; it not only depends on the qualitative speech feature, but also on the number of feature

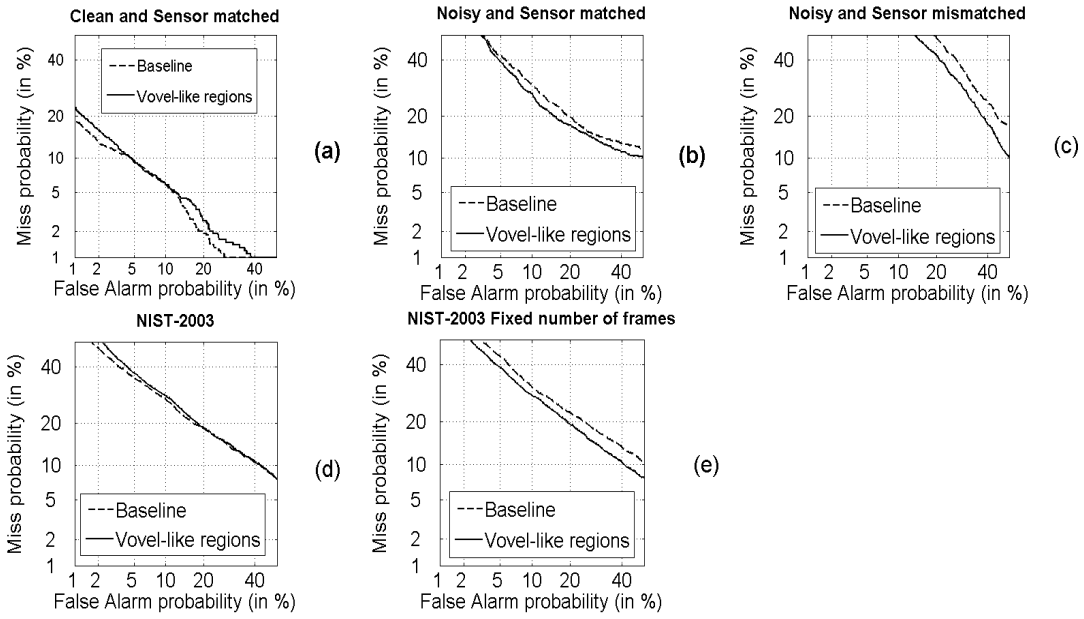

Figure 1. DET curves for different experimental conditions on IITG-MV and NIST-2003 databases.

vectors used in the training and testing process. Since, the number of feature vectors selected in vowel-like regions is less, it may give better performance for smaller mixture size, which may not be true for baseline system. So fixing a particular GMM size makes unfair comparison. Keeping this factor in mind, we evaluate the performance of systems for 16, 32, 64 and 128 GMM component densities. Table 3 and Table 4 show the performance of speaker verification using vowel-like region and baseline system, respectively, for different component densities.

A. Clean and sensor matched condition:

For a clean speech, the cepstral feature derived by short term analysis mostly contains speaker information along with the recording sensor information. The training and testing data used for this experiment are collected through the same sensor. Therefore, the sensors have almost no effect on the verification performance. To demonstrate the discriminating speaker information present in vowel-like regions, we conducted one experiment on this set of data. The DET plot in Figure 1(a) shows that for clean and sensors matched condition both vowel-like regions and speech regions provide same performance [13]. The important observation is, vowel- like regions need only about half the data of the baseline case and hence improved computationally efficiency.

B. Noisy and sensor matched condition:

The Table 2 shows that, for noisy speech (D01), the number of frames selected by the baseline system is relatively large compared to clean data (H01). In the noisy environment, separation of speech frames from silence region is a difficult task. A high threshold will eliminate most of the speech frames and low threshold takes the non-speech frames as speech frames. Also the low SNR regions are almost corrupted by noise. Alternatively, the effect of noise is not that much in case of vowel-like frames. The slight degradation may be due to the missing or spurious VOPs. Also the vowel-like regions are high SNR regions by production and hence less affected by noise. The DET plot in Figure 1(b) shows a better performance for vowel-like regions indicating that under noisy condition, speaker information can be modeled better by selecting vowel-like regions.

C. Noisy and sensor mismatched condition:

This experiment is conducted to verify the significance of vowel-like regions in a more practical situation, where the models are trained with clean data and testing data may come from other sensor or environment. The DET plot in Figure 1(c) shows that even with less number of frames, the performance using vowel-like region is better compared to the baseline system. This infers that if data is not a constraint, a better speaker verification system can be developed by using vowel-like regions under degraded and sensor mismatched conditions.

D. NIST-2003 Speaker recognition database:

In NIST-2003 database, speech data is collected through different communication channels and sensors. The DET plot in Figure 1(d) shows that the performances of both the systems are almost same in terms of EER. From the Table 2, it can be observed that, the number of frames used by the baseline system is more than double compared to vowel-like frames. For some speaker, the vowel-like frames are very small to model the speaker. As discussed earlier the performance of the system along with other factors depends on the number of training and testing feature vectors. To illustrate this point, we conduct another set of experiments by limiting the number feature vectors for training and testing. In this experiment fixing the mixture size as 32, for both the systems out of the selected frames, initial 3000 frames are used for building the models and initial 600 frames used for testing. If the number of frames are less for any speaker, for such speakers experiment is conducted with available frames. The 3000 silence removed frames for baseline system may come from one minute of speech. It is assumed that within this span of time the speaker covers all acoustic classes. Similarly 600 silence removed frames may come from about 20 sec of speech. The DET plot in Figure 1(e) shows that, the performance of vowel-like regions degraded by 0.25% compared to the performance obtained by using all vowel-like frames, where as the performance of baseline degrades around 3%. So, the baseline system is getting added advantage for more number of feature vectors. Thus if we have enough data for vowel-like regions, then in this case also vowel-like frames may show better performance.

VI. SUMMARY

In this work we introduced a new analysis technique to find the qualitative speech frames for speaker verification. This work shows that for clean speech, small number of vowel-like frames are sufficient for

speaker verification. Alternatively, for degraded and mismatched conditions, vowel-like regions provide better performance. In the practical scenarios, where a long duration speech can be made available, a robust speaker verification system can be built by selecting the vowel-like regions. The future work may focus on the detection of VOPs with better resolution and accuracy, and developing some algorithm to separate vowel region from other regions of speech. Evaluation may also be carried out on a database having enough vowel-like frames like NIST-2004.

ACKNOWLEDGMENT

This work is part of the project titled ”Development of Person Authentication System Based on Speaker Verification in Uncontrolled Environment ” supported by the Department of Information Technology (DIT), New Delhi, India.

REFERENCES

[1] J. P. Campbell, “Speaker Recognition: A Tutorial,” Proc. IEEE, vol. 85, no. 9, pp. 1437–1462, September 1997.

[2] S. Furui, “Cepstral analysis technique for automatic speaker verification,” IEEE Trans. on Acoust, Speech and Signal Processing, vol. ASSP-29, no. 2, pp. 254–272, April 1981.

[3] A. N. Khan and B.Yegnanarayana, “Vowel onset point based variable frame rate analysis for speech recognition,” in Pro. Int. Conf. Intelligent Sensing and Information Processing, January 2005, pp. 392–394.

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

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

[6] S. R. M. Prasanna and B. Yegnanarayana, “Detection of vowel onset point events using excitation source information,” in Proc. INTERSPEECH, Lisbon, Portugal, September 2005, pp. 1133– 1136.

[7] J. Makhoul, “Linear prediction:a tutorial review,” Proc. IEEE, vol. 63, no. 04, pp. 561–580, April 1975.

[8] B. Yegnanarayana, S. R. M. Prasanna, J. M. Zachariah, and S. Gupta, “Combining evidence from source, suprasegmental and spectral features for a fixed text speaker verification system,” IEEE Trans. speech and audio Processing, vol. 13, no. 4, pp. 575 – 582, July 2005.

[9] B. C. Haris, G. Pradhan, A. Misra, S. Shukla, R. Sinha, and S. R. M. Prasanna, “Multi-variability speech database for robust speaker recognition,” in Proc. National Conf. on Communication (NCC), Bangalore, India, January 2011.

[10] NIST, “Nist-speaker recognition evaluations.” Available :http://www.nist.gov/speech/tests/spk. [Online].

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

[12] D. A. Reynolds, “Speaker identification and verification using Gaussian mixture speaker models,” Speech Communication, vol. 17, pp. 91–108, March 1995.

[13] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, “The det curve in asssessment of detection task performance,” in Proc. Eur. Conf. Speech Communication Technology, Rhodes, Greece, 1997, pp. 1895–1898.

Gayadhar Pradhan was born in India in 1976. He received the B.E. degree in Electronics and Communication Engineering from Orissa Eengineering College, Bhubaneswar, Utkal University, India, in 2001 and the M.Tech. degree in Electronics and Communication Engineering from Indian Institute of Technology Guwahati, India, in 2007. He is currently pursuing the Ph.D. degree in Electronics and Electrical Engineering at Indian Institute of Technology Guwahati, India.

His research interests are in speech and speaker recognition.

S. R. Mahadeva Prasanna was born in India in 1971. He received the B.E. degree in Electronics Engineering from Sri Siddartha Institute of Technology, Bangalore University, Bangalore, India, in 1994, the M.Tech. degree in Industrial Electronics from the National Institute of Technology, Surathkal, India, in 1997, and the Ph.D. degree in Computer Science and Engineering from the Indian Institute of Technology Madras, Chennai, India, in 2004. He is currently an Associate Professor in the Department of Electronics and Electrical Engineering, Indian Institute of Technology, Guwahati.

His research interests are in speech and signal processing, application of AI tools for pattern recognition tasks in speech, and signal processing.