Prosodic Features for Speaker Verification

Leena Mary and B.Yegnanarayana

Speech and Vision Laboratory
Department of Computer Science and Engineering
Indian Institute of Technology Madras, Chennai - 600 036 India
{leena, yegna}@cs.iitm.ernet.in

Abstract

In this paper we study the effectiveness of prosodic features for speaker verification. We hypothesize that prosody is linked to linguistic units such as syllables and prosodic features can be better represented with reference to the syllabic sequence. For extracting prosodic features, speech is segmented into syllable-like regions using the knowledge of vowel onset points (VOP). We use a technique based on excitation source information to detect VOPs automatically. The location of VOPs serve as reference for extracting prosodic features directly from speech signal. Various parameters are used to represent the pitch and energy dynamics of the region between two consecutive VOPs. The effectiveness of the derived prosodic features for speaker verification is demonstrated on NIST SRE 2003 extended data. The complementary nature of prosodic features and spectral features help to improve the accuracy of the combined speaker verification system.

Index Terms: prosody, speaker verification, syllable, vowel onset point, F_0 contour.

1. Introduction

Speaker characteristics are manifested in speech signal as a result of anatomical differences inherent in speech production organs and differences in learned speaking habits of individuals [1]. Characteristics of a speaker can be represented using short term and long term features [2]. Short-time features are capable of reflecting the physiological difference among the speakers. The long term features mostly represent the habitual attributes of a speaker such as *prosody* and *idiolect*. *Prosody* is a term used for representing characteristics such as intonation, timing and stress in a collective manner.

Current text independent speaker verification systems rely mostly on spectral features derived through short term spectral analysis. This approach does not attempt to model the long-term speaker-specific characteristics present in the speech signal. The long-term features are relatively less affected by channel mismatch and noise. In order to incorporate long-term features, system generally require significantly more data for training. Hence in 2001, NIST introduced the extended data task which provides multiple conversation sides for speaker training [3]. This helps in the study of long-term features for speaker verification. A workshop was conducted at the John Hopkins University to explore a wide range of features for speaker verification using NIST 2001 extended data task as its testbed [4].

In general, the existing approaches followed for extraction of prosodic features can be broadly categorized into two. One approach uses the explicit segment boundaries obtained using automatic speech recognizer (ASR), for extracting various duration, pitch and energy features for each estimated syllables [5]. But for applications like speaker recognition, the use of ASR may not be needed. In the other approach, inflection points and start or end of voicing of pitch and energy trajectories are used to segment the speech signal and features are derived from linear stylized segments of pitch and energy contour [6]. This approach has the advantage that features are derived directly from the speech signal. In this paper, we propose a technique based on vowel onset points for extracting prosodic features directly from speech signal. This method combines the salient features of both the approaches mentioned above, namely, the association with the syllabic pattern as in first approach, and extraction of features without using ASR as in the second approach.

This paper is organized as follows: Section 2 discuss the speaker-specific aspect of prosody. In Section 3, automatic extraction, representation and modeling of prosodic features for speaker verification is described. The results of our experimental studies are discussed in Section 4. The final section summarizes the studies.

2. Speaker-specific Aspect of Prosody

It is not just the physiological aspects of speech production organs of a speaker that influence the way an utterance is spoken. It is also influenced by the habitual aspect of a particular speaker. The acquired speaking habits are characteristics learned over a period of time, mostly influenced by the social environment and also by the characteristics of the first/native language in the ‘critical period’ (lasting roughly from infancy until puberty) of learning. The prosodic characteristics as manifested in speech give important information regarding the speaking habit of a person.

Pitch is a perceptual attribute of sound. The physical correlate of pitch is the fundamental frequency (F_0) of vibration of vocal folds. It is speaker-specific due to differences in the physical structure of the vocal folds among speakers. The average value of F_0 is generally higher for children and females, due to smaller size of the vocal folds. Researchers have attempted to capture the global statistics of F_0 values of a speaker using appropriate distributions for speaker verification task [7]. The F_0 value is controlled either by varying the subglottal pressure or laryngeal tension or a combination of both [8], which is speaker-specific. The dynamics of F_0 contour is influenced by several factors such as the identity of the sound unit spoken, position with respect to phrase/words, context (the units that precede and follow), speaking style of a particular speaker, intonation rules of the language, type of sentence (interrogative or declarative) etc. But when the same text is repeated by the same speaker in the same context, F_0 contour characteristics are consistent for a particular speaker, but different across speakers as demonstrated in Fig 1. The speaker-specific information present in F_0 contour and energy variations are useful for modeling a speaker.

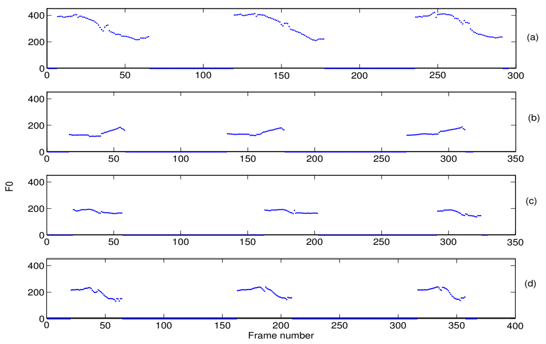

Figure 1: Variation in F_0 contour dynamics of four different speakers while repeating the same text $(Sunday, Sunday, Sunday)$. Figures show the F_0 contours of (a) a child, (b) and (c) two different males, and (d) a female speaker.

3. Prosodic Features for Speaker Verification

Prosody is linked to the underlying syllable sequence [9], and it is meaningful to associate the prosodic features to the syllabic sequence. The association of syllable sequence with F_0 contour, as used in this study is shown in Fig. 2. Segmentation into syllable-like regions is accomplished with the knowledge of vowel onset points (VOP) as illustrated in Fig. 3(a), where VOP refers to the instant at which the onset of vowel takes place in a syllable. A technique based on the excitation source information for extracting the VOPs from continuous speech is used in this study [10]. It uses the Hilbert envelope of linear prediction (LP) residual which represents the strength of excitation. The instant with maximum excitation within a pitch period corresponds to the instant of glottal closure. The places with significant change in the strength of excitation gives the evidence for the detection of VOPs. The strength of excitation at the instants of glottal closure for voiced sounds is generally higher compared to the strength at random instants present in the unvoiced sound. Also, the strength of excitation at the instants of glottal closure for vowels is higher compared to the strength of the voiced consonants. This change in strength of excitation is utilized for detecting VOP by convolving the Hilbert envelope of the LP residual with a Gabor window.

The locations of VOP are then associated with F_0 contour as in Fig. 3(b), for feature extraction. The continuous portion of the F_0 contour with nonzero values, located within the region of two consecutive VOPs, is treated as one segment of F_0 contour. The F_0 contour, located within the region of two consecutive VOPs is treated as one segment for feature extraction.

3.1. Representation of prosodic features

As shown in Fig. 1, the shape of the F_0 contour reflects certain speaking habits of a person. Therefore it is important to represent it using suitable parameters. We use tilt parameters [11] for representing the dynamics of F_0 contour. With reference to Fig. 4, the tilt parameters, namely the amplitude tilt (A_t), and the duration tilt (D_t) are defined as follows:

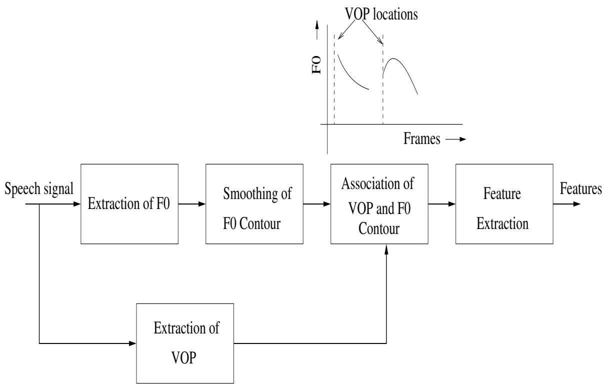

Figure 2: Association of F_0 contour with syllabic sequence using automatically detected VOPs, for prosodic feature extraction.

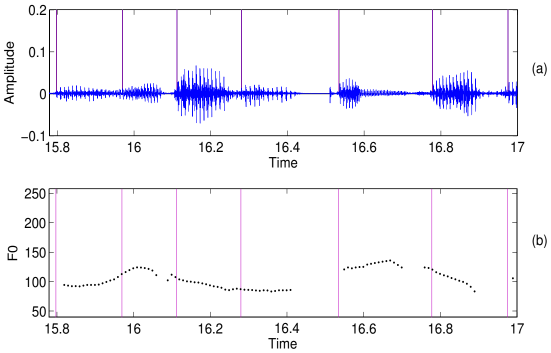

Figure 3: (a) Segmentation of speech into syllable-like units using automatically detected VOPs (b) F_0 contour associated with VOPs.

$$A_t = \frac{|A_r| - |A_f|}{|A_r| + |A_f|} \tag{1}$$

$$D_t = \frac{|D_r| - |D_f|}{|D_r| + |D_f|} \tag{2}$$

A_r and A_f represent the rise and fall in F_0 amplitude, respectively, with respect to F_{0_p}. Similarly D_r and D_f represent the duration taken for the rise and fall, respectively.

The role of articulatory constraints in shaping the F_0 contour in speech has been investigated by researchers [12]. The maximum speed of pitch change limits how fast the F_0 movements can be produced, and the coordination of laryngeal and supralaryngeal movements limits how syllables and tone can be aligned to each other [13]. Studies have also indicated that listeners are more sensitive to variations in fundamental frequency peak F_{0_p} than valley F_{0_v} [9]. Hence the height of F_0 peak ($\Delta F_0 = F_{0_p} - F_{0_v}$), distance of F_0 peak (D_p) and peak value of pitch (F_{0_p}) for each pitch segment may be useful for speaker verification. An increase in pitch may be obtained by increasing the vocal fold tension, by increasing the subglottal pressure, or a combination of them. Therefore F_0 peak (F_{0_p}) and F_0 mean(F_{0_μ}) obtained for each pitch segment reflects some

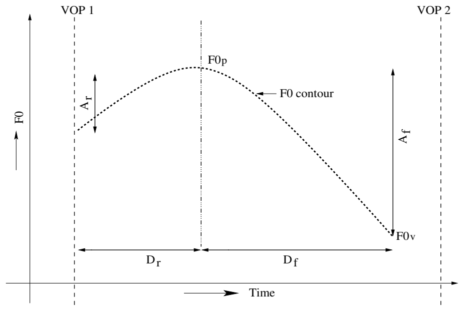

Figure 4: A segment of F_0 contour. Tilt parameters A_t and D_t along with height of F_0 peak and distance of F_0 peak represent the dynamics of F_0 contour segment.

physiological aspect of a speaker. The change in log energy (ΔE) in the voiced region along with F_0 change give a quantitative measure of stress, therefore may be specific to a particular speaker. The F_0 and energy related parameters used in this study for characterizing the speaker-specific aspect of prosody are the following:

(a) Mean value of pitch ($F0_\mu$)

(b) Peak fundamental frequency ($F0_p$)

(c) Change of F_0 (ΔF_0)

(d) Distance of F_0 peak with respect to VOP (D_p)

(e) Amplitude tilt (A_t)

(f) Duration tilt (D_t)

(g) Change of log energy (ΔE)

Each region between two consecutive VOPs is represented using the above mentioned parameters to form a 7-dimensional feature vector.

3.2. Modeling of speaker-specific prosody

We hypothesize that the distribution of syllable level prosodic feature vectors for a particular speaker form a unique cluster in the feature space. To capture the distribution of the feature vectors, autoassociative neural network (AANN) models or alternatively conventional Gaussian mixture models (GMM) can be used. The AANN is a feedforward neural network which tries to map an input vector onto itself, and hence the name autoassociation or identity mapping. It consists of an input layer, an output layer and one or more hidden layers. A typical structure of a five layer AANN is shown in Figure 5. The number of units in the input and output layers are equal to the size of the input vectors. The number of units in the middle hidden layer is less than the number of units in the input and output layers, and this layer is called the dimension compression hidden layer. The activation function of the units in the input and output layers are linear, whereas the activation function of the units in the hidden layers can be either linear or nonlinear. It has been demonstrated that the AANN has the ability to capture the distribution of input data [14].

To capture the distribution of the input feature vectors in the feature space, the feature vectors are extracted from the signal, and are presented in a random order to the AANN. The structure of the AANN model used for capturing the distribution of the speaker-specific prosodic features is *7L 28N 2N 28N 7L*, where *L* represent linear units, *N* represent nonlinear units, and the numerals represent the number of units in the layers. One AANN model is trained for each speaker using 500 epochs using backpropagation algorithm. During testing, for each prosodic feature vector (corresponding to each syllable) in the test utterance, the error between the output and the input of AANN is noted. This error is converted into confidence value using $C_i = exp(-E_i)$, where E_i is the squared error for the i^{th} frame. The average confidence is computed as $C = \frac{1}{N}\sum_{i=1}^{N} C_i$, where C_i is the confidence value for the i^{th} syllable, and N is the number of syllables in the test utterance.

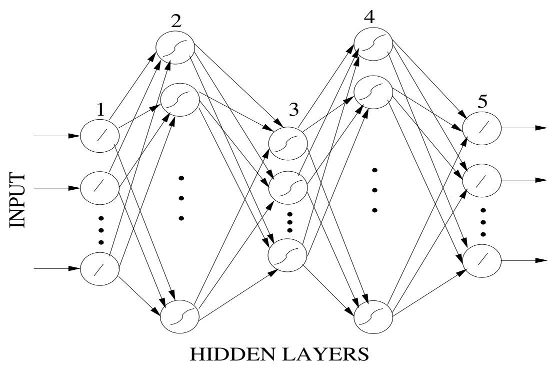

Figure 5: Structure of five layer AANN model.

4. Results from Experimental Study

To demonstrate the effectiveness of the prosodic features for speaker verification, we use the 16 side conversational data in the first subset of NIST 2003 extended data task. It provides 16 conversation sides (where each conversation side contains approximately 2.5 minutes of speech) for training the target speaker model. Test utterances are of duration 2.5 minutes approximately. Each target model is tested with a set of test utterances where the task is to find out whether the particular test utterance belongs to the target speaker or not. The subset chosen for our study consists of 137 speaker models and 1076 test utterances.

For each target speaker, a model is developed to capture the characteristic distribution of the prosodic features. A set of background models built from a known set of impostor speakers helps to fix a global threshold for verification, to decide whether the test utterance belongs to the target speaker or not. The background models consists of a set of male and female models as illustrated in Fig. 6. For each test utterance, gender decision is made based on the mean score of male/female background model set. The mean and standard deviation of scores of the appropriate male/female background model set is then used for test score normalization. Prosody based system resulted in an equal error rate (EER) of 12.4 for the particular data set as per the detection error tradeoff (DET) curve shown in Fig. 7.

As spectral features are vulnerable to channel mismatch and noise, the use of prosodic features can play important role in improving the robustness of the speaker verification system. The evidence about the speaker from different features may be combined in several ways to achieve better performance. One simple approach is the addition of evidences from different systems. Our baseline speaker verification system [14] use AANN model for capturing the distribution of spectral vectors represented using weighted linear cepstral coefficients (WLPCC). As single conversation side of 2.5 minutes is sufficient for building

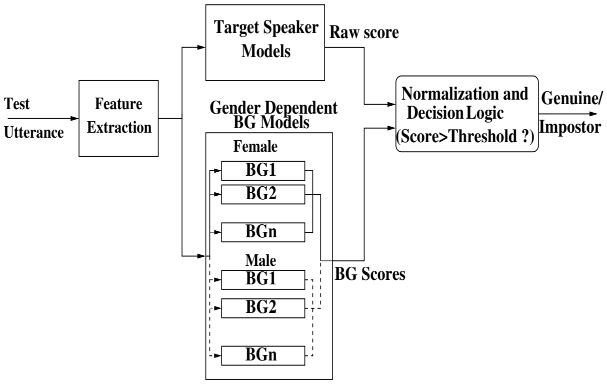

Figure 6: Block diagram showing testing of an unknown utterance against target and background models. The mean and standard deviation of gender dependent background model set is used for normalizing the raw score.

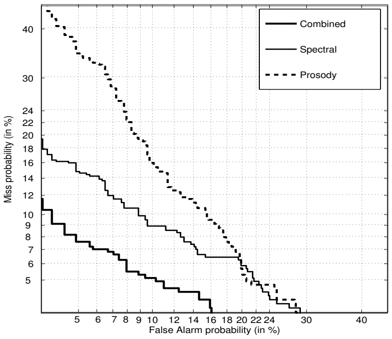

Figure 7: DET curves showing the performance of spectral based, prosody based, and combined system for 16 side conversational case.

a spectral-based model, 16 AANN models are built for a target speaker using 16 conversation sides. Average of N-best (N=8) confidence scores is taken as the score of the target speaker, and this score is normalized as shown in Fig. 6. This scoring approach is expected to reduce the effects of channel variability. This spectral-based system resulted in an EER of 9.5 for the same 16 side data set. Combining prosody-based evidences and spectral-based evidences by simple addition results in an EER of 6.8 as illustrated in Fig. 7, showing the presence of complementary information in these features.

5. Summary and Conclusions

The goal of this study was to demonstrate the usefulness of prosodic features, extracted directly from the speech signal, for speaker verification task. The feature extraction technique used in this study do not use automatic speech recognizer, but still gives a meaningful association of prosodic features with the corresponding syllable sequence. This is done using the location of VOPs detected automatically from the Hilbert envelope of the LP residual of the speech signal. The set of parameters derived from the F_0 and energy variation for the region between consecutive VOPs form the feature vector and distribution of these feature vectors are captured to model prosodic characteristics of the speaker. A study conducted on NIST SRE 2003 extended data demonstrated the potential of these prosodic features for speaker verification. The complementary nature of the prosodic and spectral features helps to improve the overall performance of speaker verification, while combining the evidences.

6. References

[1] J. P. Campbell, “Speaker recognition: A tutorial”, Proc. IEEE, Vol. 85, no. 9, p 1437–1462, Sep. 1997.

[2] L. P. Heck, “ Integrating high-level information for robust speaker recognition” in John Hopkins University workshop on SuperSID, Baltimore, Maryland, http: www.cslp.jhu.edu/ws2002/groups/supersid, July 2002.

[3] NIST 2001 speaker recognition evaluation website:, http://www.nist.gov/speech/tests/spk/2001.

[4] D. Reynolds, W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adami, Q. Jin, D. Klusacek, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones and B. Xiang, “The superSID project: Exploiting high-level information for high-accuracy speaker recognition”, in Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing, Hong Kong, China, Vol. 4, p 784–787, Apr. 2003.

[5] E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke, “Modeling prosodic feature sequences for speaker recognition”, Speech Communication, Vol. 46, p 455–472, 2005.

[6] A. G. Adami, R. Mihaescu, D. A. Reynolds, and J. J. Godfrey, “Modeling prosodic dynamics for speaker recognition”, in Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing, Hong Kong, China, Vol. 4, p 788–791, Apr. 2003.

[7] M. K. Sonmez, L. Heck, M. Weintraub, and E. Shriberg“A lognormal tied mixture model of pitch for prosody-based speaker recognition”, Proc. EUROSPEECH, Rhodes, Greece, Vol. 3, p 1391–1394, 1997.

[8] James E. Atkinson, “Correlation analysis of the physiological factors controlling fundamental voice frequency”, J. Acoust. Soc. Am., Vol. 63, no. 1, p 211–222, 1978.

[9] Y. Xu, “Consistency of tone-syllable alignment across different syllable structures and speaking rates”, Phonetica, Vol. 55, p 179–203, 1998.

[10] S. R. Mahadeva Prasanna, Suryakanth V. Gangashetty, and B. Yegnanarayana, “ Significance of vowel onset point for speech analysis” in Proc. Int. Conf. on signal processing and communication, Bangalore, India, Vol. 1, p 81–86, July 2001.

[11] P. Taylor, “Analysis and synthesis of intonation using the tilt model”, J. Acoust. Soc. Am., Vol. 107, no. 3, p 1697–1714, Mar. 2000.

[12] Y. Xu and S. Xuejing, “Maximum speed of pitch change and how it may relate to speech”, J. Acoust. Soc. Am., Vol. 111, no. 3, p 1399–1413, Mar. 2002.

[13] C. Gussenhoven, B. H. Reepp, A. Rietveld, H. H. Rump and J. Terken, “The perceptual prominence of fundamental frequency peaks”, J. Acoust. Soc. Am., Vol. 102, no. 5, p 3009–3022, Nov. 1997.

[14] B. Yegnanarayana and S. P. Kishore, “AANN-An alternative to GMM for pattern recognition”, Neural Networks, Vol. 15, p 459–469, 2002.