Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 50 (2008) 782–796
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Extraction and representation of prosodic features for language and speaker recognition

Leena Mary [a,*], B. Yegnanarayana [b]

[a] Speech and Vision Laboratory, Department of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai 600 036, India
[b] Department of Computer Science and Engineering, International Institute of Information Technology, Hyderabad 500 032, India

Received 6 April 2006; received in revised form 20 February 2008; accepted 24 April 2008

Abstract

In this paper, we propose a new approach for extracting and representing prosodic features directly from the speech signal. We hypothesize that prosody is linked to linguistic units such as syllables, and it is manifested in terms of changes in measurable parameters such as fundamental frequency (F_0), duration and energy. In this work, syllable-like unit is chosen as the basic unit for representing the prosodic characteristics. Approximate segmentation of continuous speech into syllable-like units is obtained by locating the vowel onset points (VOP) automatically. The knowledge of the VOPs serve as reference for extracting prosodic features from the speech signal. Quantitative parameters are used to represent F_0 and energy contour in each region between two consecutive VOPs. Prosodic features extracted using this approach may be useful in applications such as recognition of language or speaker, where explicit phoneme/syllable boundaries are not easily available. The effectiveness of the derived prosodic features for language and speaker recognition is evaluated in the case of NIST language recognition evaluation 2003 and the extended data task of NIST speaker recognition evaluation 2003, respectively.

Keywords: Prosody; Vowel onset point; Intonation; Stress; Rhythm; Language recognition; Speaker recognition; Multilayer feedforward neural network; Autoassociative neural network

1. Introduction

Speech is primarily intended to convey some message. It is conveyed through a sequence of legal sound units in a language. But speech cannot be merely characterized as a sequence of sound units. There are some characteristics that lends naturalness to speech. The variation of pitch provides some recognizable melodic properties to speech. This controlled modulation of pitch is referred as *intonation*. The sound units are shortened or lengthened in accordance to some underlying pattern giving rhythmic properties to speech. Some syllables or words may be made more prominent than others, resulting in linguistic stress. The information gleaned from melody, timing and stress in speech increases the intelligibility of spoken message, enabling the listener to segment continuous speech into phrases and words with ease (Shriberg et al., 2000). It is also capable of conveying many more lexical and nonlexical information such as lexical tone, prominence, accent and emotion. The characteristics that make us perceive these effects are collectively referred to as prosody. Prosodic cues include stress, rhythm and intonation. Each cue is a complex perceptual entity, expressed primarily using three acoustic parameters: pitch, energy and duration.

Prosodic characteristics such as rhythm, stress and intonation in speech conveys some important information regarding the identity of the spoken language. Results of perception studies on human language identification,

* Corresponding author. Tel.: +91 44 22575382; fax: +91 44 22574352.
E-mail addresses: leenmary@rediffmail.com (L. Mary), yegna@cs.iiit.ac.in (B. Yegnanarayana).

0167-6393/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2008.04.010

confirm that prosodic information, specifically pitch and intensity, are used for language identification under conditions where the acoustics of sound units and phonotactics are degraded (Mori et al., 1999; Kometsu et al., 2001). A study using resynthesis has revealed the importance of rhythm and intonation for language discrimination (Ramus and Mehler, 1999). Since each speaker has unique physiological characteristics of speech production and speaking style, speaker-specific characteristics are also reflected in prosody. It is generally recognized that human listeners can better recognize those speakers who are familiar to them, than those who are relatively less familiar. This increased ability is due to speaker-specific prosody and idiosyncrasies that are recognized by the listener, either consciously or otherwise (Doddington, 2001). But it is very difficult even for a listener to describe the nature of language-specific and speaker-specific prosodic features that he/she will be using for recognition. Distinguishing the language-specific and speaker-specific aspect of prosody using acoustic parameters is even more difficult. Therefore, it is a challenging task to extract and represent prosodic features for recognizing a language or a speaker.

In general, there are two broad approaches for extracting prosodic features from speech. The first approach uses the explicit subword boundaries obtained using automatic speech recognizer (ASR) for extracting the prosodic features (Shriberg et al., 2005). But for applications like language and speaker recognition, the use of ASR may not be needed. In most of the ASR-free approaches, pitch contour dynamics are represented using parameters derived from linear stylized pitch segments (Sonmez et al., 1998; Adami et al., 2003; Reynolds et al., 2003; Peskin et al., 2003), which has the advantage that features are derived directly from the speech signal. In this paper, we propose a new technique for extraction and representation of prosodic features. The proposed technique combines salient features of both approaches mentioned above, namely, association with the syllabic pattern as in the ASR-based approach, and extraction of features without explicit speech recognition as in the ASR-free approach.

In this paper, we address the issues related to *language* and *speaker recognition*, focusing on prosodic features extracted from the speech signal. The remaining part of this paper is organized as follows: In Sections 2 and 3, we describe language-specific and speaker-specific aspect of prosody, respectively. In Section 5, automatic extraction and representation of prosodic features employed in the proposed approach is described. Section 6 describes the prosodic features we use for language recognition, and discusses the results of experimental studies on NIST language recognition evaluation (LRE) 2003 task. The features used for capturing the speaker-specific prosodic characteristics are described in Section 7, along with the results of experimental studies on the extended data task of NIST speaker recognition evaluation (SRE) 2003. The final section summarizes the studies.

2. Language-specific aspect of prosody

There are a number of striking acoustic similarities in the suprasegmental aspects of neutral sentences in different languages. This is mostly due to identical constraints imposed by the production and perception apparatus. Pitch is a perceptual attribute of sound. The physical correlate of pitch is the fundamental frequency (F_0) of vibration of vocal folds. The functions of intonation are mostly defined as attitudinal, accentual, discourse and grammatical (Roach, 1983). A comparison shows that languages differ greatly in this respect (Hirst and Di Cristo, 1998; Fox, 2000). Some functions that are performed by intonation in one language may be expressed lexically and/or syntactically in others (Hirst and Di Cristo, 1998). As an illustration, samples of the F_0 contours of Farsi and Mandarin are shown in Fig. 1. It can be observed that in spite of speaker differences, Mandarin has more F_0 variations compared to Farsi.

Languages can be broadly categorized as stress-timed, syllable-timed and mora-timed, based on their timing/rhythmic properties. In stress-timed languages like English and German, duration of the syllables are mainly controlled by the presence of stressed syllables which is directly encoded in the lexicon. But intervals between two stresses are said to be near-equal (Abercrombie, 1967). Syllables that occur in between two stressed syllables are shortened to accommodate this property. In syllable-timed languages such as French and Spanish, successive syllables are said to be of near-equal duration (Abercrombie, 1967). The mora-timing is exemplified by Japanese. Morae are subunits of syllables consisting of one short vowel and any preceding onset consonants. In mora-timing, successive morae are said to be near-equal duration (Grabe and Low, 2002). Languages are also classified as stress-accented and pitch-accented, based on the realization of prominence. In pitch-accented languages like Japanese, prominence of a syllable is achieved through pitch variations, whereas in stress-accented language, pitch variation is only one factor that helps to assign prominence. Languages are also categorized as tonal and nontonal. Modern Standard Chinese uses pitch patterns to distinguish one word from another (Ashby and Maidment, 2005). A lexical tone language is one in which an indication of pitch enters into the lexical realization of at least some morphemes (Hyman, 2005). We can identify languages which employ lexical tone such as Mandarin Chinese or Zulu (tonal languages), those which use lexically based pitch accents like Swedish or Japanese (pitch-accented languages), and stress-accented languages such as English or German (Cummins et al., 1999). There are many other languages which strictly do not follow the rules of a class, which means that these classifications are rather a continuum (Grabe and Low, 2002).

In most of the languages, higher intensity, larger pitch variation and longer duration help to assign prominence to stressed syllables. In languages like English and French, a longer duration syllable carries more pitch movements.

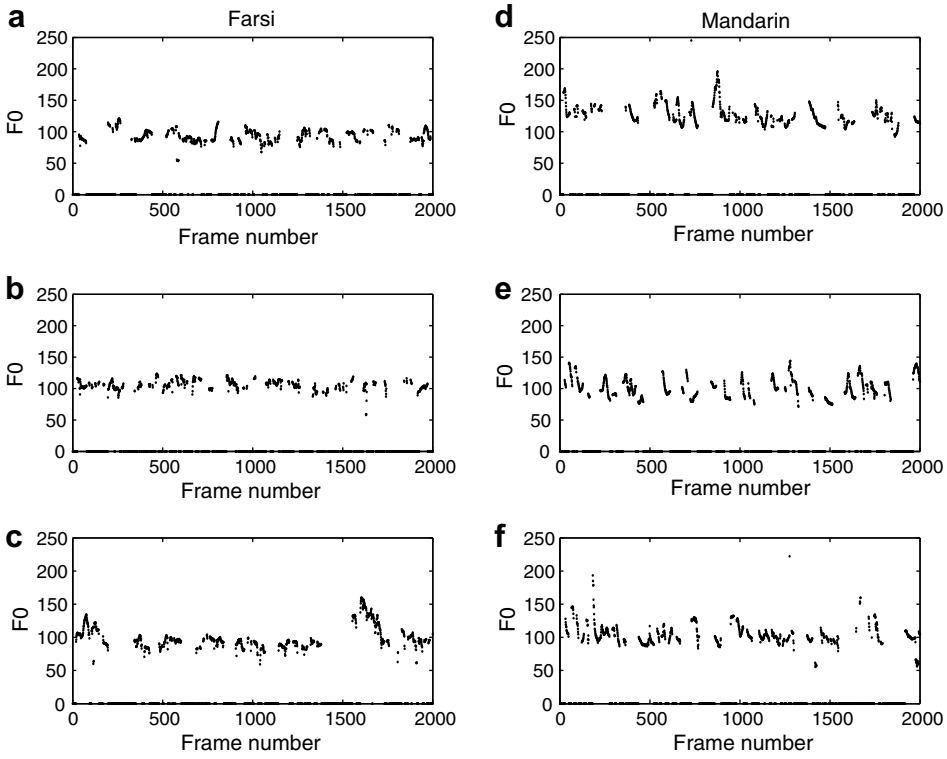

Fig. 1. Variation in dynamics of F_0 contour for utterances in Farsi and Mandarin, spoken by three male speakers each. (a), (b) and (c) correspond to Farsi (d), (e) and (f) correspond to Mandarin utterances (taken from Oregon Graduate Institute (OGI) multi-language telephone speech corpus).

But such a correlation may not hold equally well for all languages. Stress in some languages is what defines the rhythm of speech (Ashby and Maidment, 2005). Therefore it is possible that, the specific interaction between the suprasegmental features, and relation between suprasegmental and segmental aspects are the most salient characteristics that differentiate between languages (Cutler and Ladd, 1983).

3. Speaker-specific aspect of prosody

Speaker characteristics vary due to difference in (a) physiological characteristics of speech production organs and (b) acquired/learned habits. Physiological difference in the shape and size of oral tract, nasal tract, vocal folds and trachea can lead to differences in vocal tract dynamics and excitation characteristics. The distribution of fundamental frequency (F_0) values varies among speakers due to differences in the physical structure of the vocal folds as illustrated in Fig. 2.

It is not just the physiological aspects of speech production organs of a speaker that influence the way an utterance is spoken. Speaker characteristics are also influenced by the speaking habits of a particular speaker. The acquired speaking habits are 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. Dynamics of F_0 contour corresponding to a sound unit is influenced by several factors such as identity of the sound unit, its position with respect to the phrase/word, its context (the units that precede and follow), the speaking style of the speaker, intonation rules of the language, type of the sentence (interrogative or declarative), etc. The dynamics of F_0 contour and energy contour can be different among speakers due to different speaking style and accent. The dynamics of F_0 contour will be different for two speakers, even when they utter the same text as illustrated in Fig. 3. But when the same text is repeated by a given speaker, the characteristics of F_0 contour are consistent and this is true across speakers as illustrated in Fig. 4. The presence of speaker-specific information in temporal dynamics of F_0 contour may be used for characterizing a speaker. This property is used in text-dependent speaker verification, by comparing F_0 contours using dynamic time warping (DTW). It has been shown that the dynamics of F_0 contour can also contribute to text-independent speaker verification task (Sonmez et al., 1998; Adami et al., 2003).

4. Robustness of prosodic features

Most of the current speaker/language recognition systems rely on the spectral features derived through short-time spectral analysis of the speech signal. The magnitude of the short-time spectrum encodes information about vocal tract shape of the speaker (Makhoul, 1975; Furui,

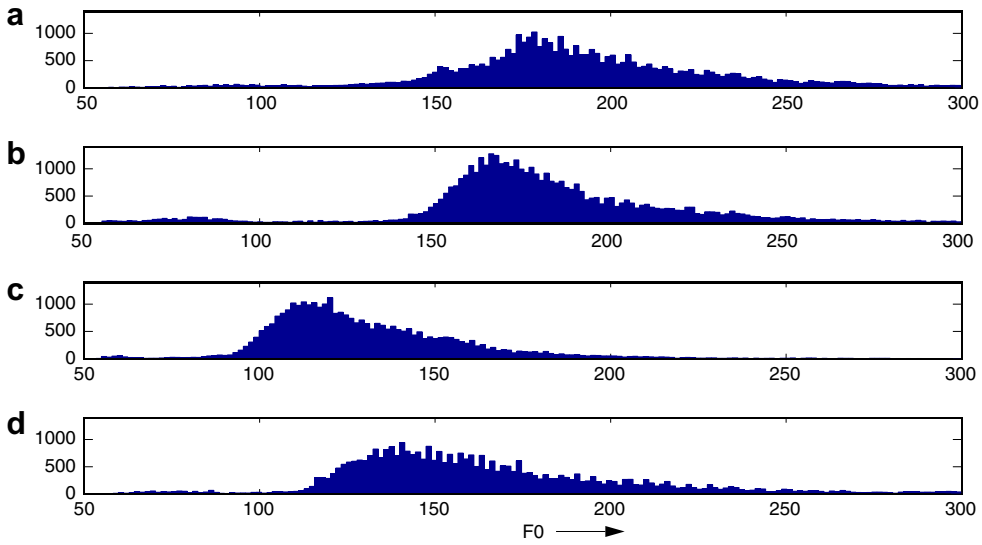

Fig. 2. Variation in histogram of F_0 for (a), (b) Two female and (c), (d) Two male speakers.

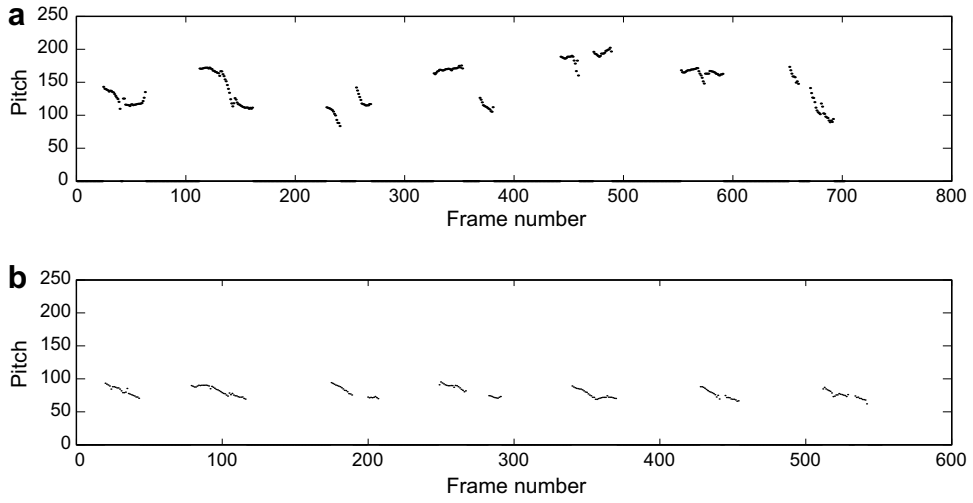

Fig. 3. Variation in dynamics of F_0 contour of two different male speakers while uttering *Sunday, Monday, Tuesday, Wednesday, Thursday, Friday, Saturday (taken from OGI multi-language telephone speech corpus)*.

1981; Reynolds and Rose, 1995). Therefore, spectral features are widely used for speaker modeling. The spoken language recognition research has its main focus on spectral information, using the acoustic property of sound units (referred as acoustic-phonetics) and their sequencing (referred as phonotactics) (Zissman, 1996; Navratil, 2001). Language and speaker recognition systems based on spectral features perform well in favorable acoustic conditions, but their performance degrade due to noise and unmatched acoustic conditions (Reynolds, 1996). Prosodic features derived from pitch, energy and duration are relatively less affected by channel variations and noise (Thyme-Gobbel and Hutchins, 1996). Though the systems based on spectral features outperform the prosody-based systems, their combined performance may provide the needed robustness to recognition systems.

The effect of channel variations on spectral feature vectors and F_0 contour are illustrated in Figs. 5 and 6, respectively. The same utterance *Don't carry an oily rag like that* recorded through three different channels, available in Texas Instruments and Massachusetts Institute of Technology (TIMIT) database, is used for comparing the effect of channel variations. Channels correspond to TIMIT, NTIMIT and CTIMIT represent speech collected over close-speaking microphone, noisy channel and cellular environment, respectively. Fig. 5 shows the difference in Euclidean distance of LPCC features due to variability in the channel characteristics, whereas Fig. 6 illustrates the robustness of F_0 contour characteristics against channel variations. In Fig. 6, the F_0 contours remain the same for all the cases except some durational variation of voiced region in (b) and (c) compared to (a).

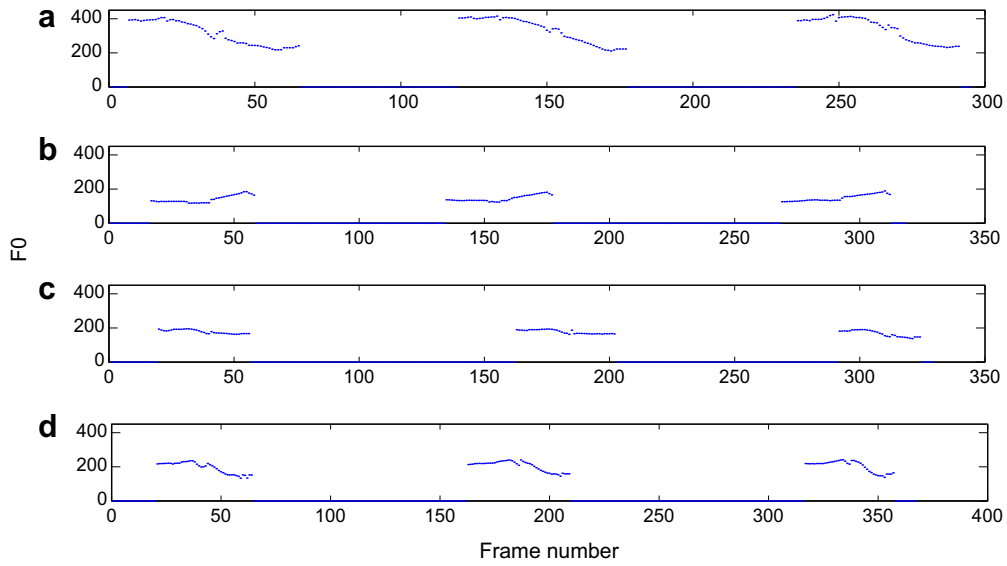

Fig. 4. Variation in F_0 contour dynamics of four different speakers: (a) Child voice. (b), (c) Two different male voices. (d) Female voice. All repeating the same text *Sunday, Sunday, Sunday*.

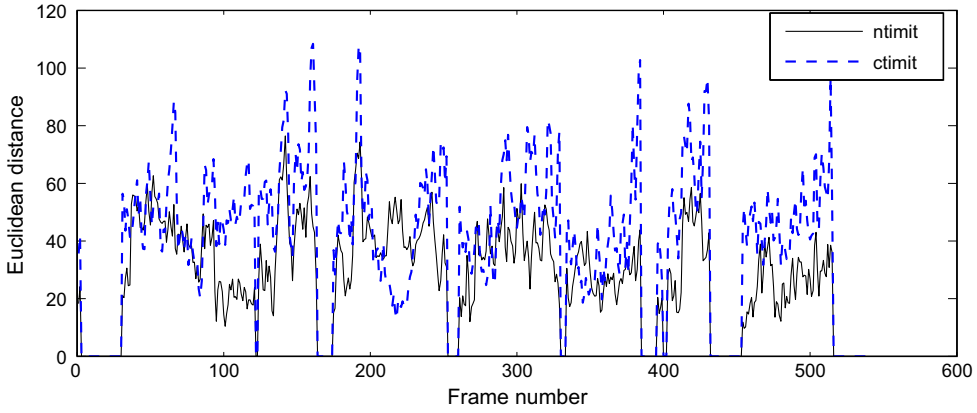

Fig. 5. Euclidean distance of LPCC feature vectors on a frame to frame basis for the same speaker and text *Don't carry an oily rag like that*. The solid line corresponds to the distance of NTIMIT data and dashed line corresponds to CTIMIT data with respect to TIMIT data.

5. Extraction of prosodic features from speech signal

Approaches used for prosodic feature extraction can be broadly categorized into two, based on the use of automatic speech recognizer (ASR). In the ASR-based approach, syllabic/phone boundaries are obtained with the help of ASR (Shriberg et al., 2005). In the ASR-free approach, segment boundaries are estimated using cues derived from the speech signal. In one approach, inflection points and start or end of voicing are used to segment the speech signal (Adami and Hermansky, 2003). The segmented trajectories are then quantized and labeled into a small set of classes that describe the dynamics of F_0 contour and energy contour. N-grams of these labels are used to model the characteristics of a speaker or a language (Shriberg et al., 2005; Adami and Hermansky, 2003). Recent approaches to automatic syllable-like segmentation include the use of vowel detection (Rouas et al., 2005) and group delay function of minimum phase signal (Nagarajan and Murthy, 2006). In the present study, we use locations of vowel onset points (VOP) for identifying the syllable-like regions in continuous speech.

5.1. Choice of syllable as the basic unit

All spoken utterances can be considered as sequence of syllables which constitute a continual rhythmic alternation between opening and closing of mouth while speaking (MacNeilage, 1998). Syllable of CV type provides an articulatory pattern beginning with a tight restriction and ending with an open vocal tract, resulting some rhythm that is especially suited both to the production and the perception mechanisms (Krakow, 1999). It is demonstrated that the tonal events are aligned to the segmental events such as onset and/or offset of a syllable (Atterer and Ladd,

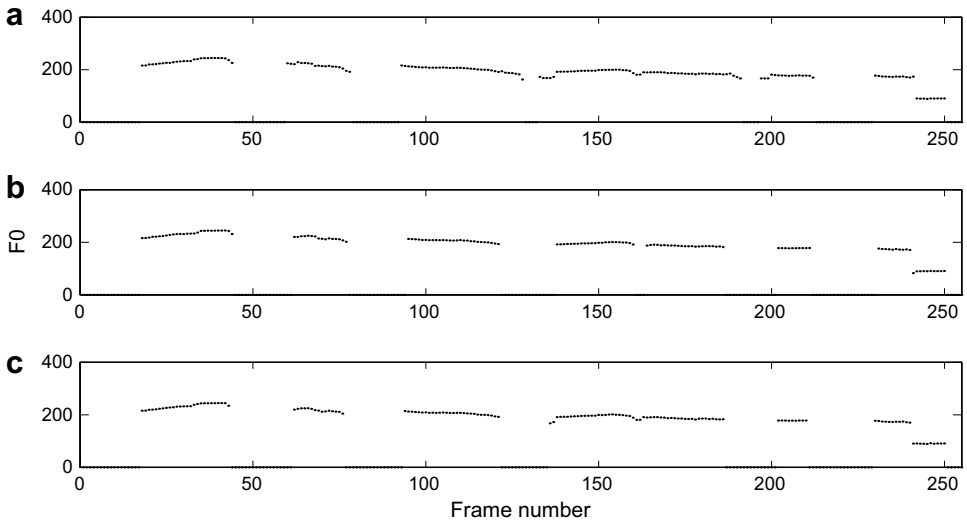

Fig. 6. F_0 contours of (a) TIMIT; (b) NTIMIT and (c) CTIMIT sentence of the same speaker for the same sentence *Don't carry an oily rag like that.*

2004). Therefore, syllable appears to be a natural choice for the basic unit for representing prosody.

5.2. Association of prosody with sequence of syllable-like units

For representing syllable-based rhythm, intonation, and stress, the speech signal should be segmented. Segmenting speech into syllables is typically a language-specific mechanism, and thus it is difficult to develop a language independent algorithm for this. In this work, segmentation into syllable-like units is accomplished with the knowledge of VOPs as illustrated in Fig. 7a, where VOP refers to the instant at which the onset of vowel takes place in a syllable. There may be limitations in this approach, but since it provides a language-independent solution to the segmentation problem, it is adopted in the present study. The F_0 contour of speech signal is then associated with the locations of VOPs as shown in Fig. 7b, for feature extraction.

5.3. Detection of vowel onset points

Vowel onset point is an important event in speech production, which may be described in terms of changes in the vocal tract and excitation source characteristics. For extracting the VOPs from continuous speech, a technique which relies on excitation source information is used in this study (Mahadeva Prasanna et al., 2001). It uses the Hilbert

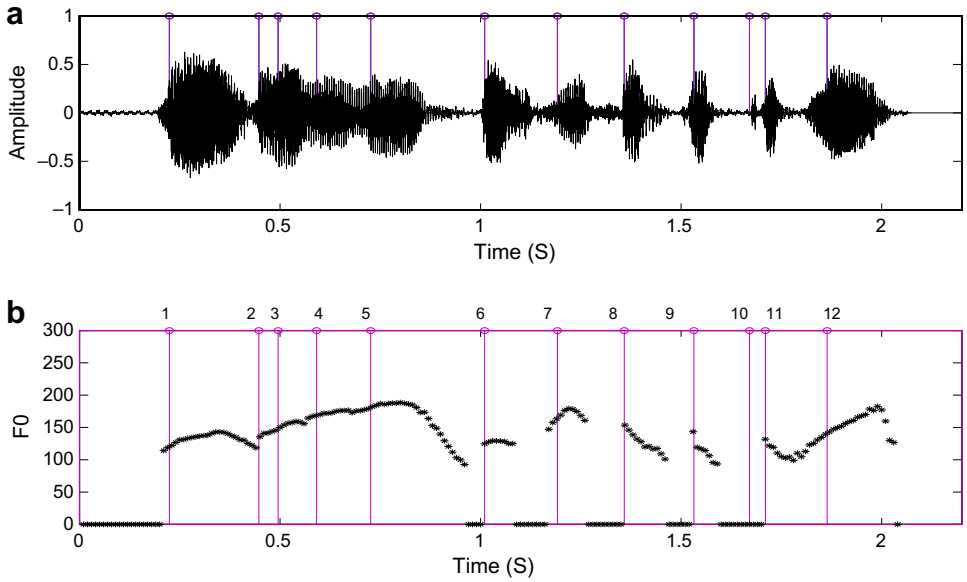

Fig. 7. (a) Segmentation of speech into syllable-like units using automatically detected VOPs and (b) F_0 contour associated with VOPs (marked as '1' to '12').

envelope of linear prediction (LP) residual, which is defined as

$$h_e(n) = \sqrt{r^2(n) + r_h^2(n)} \tag{1}$$

where $r(n)$ is the LP residual of the speech signal, and $r_h(n)$ is the Hilbert transform of $r(n)$, where Hilbert transform is defined as

$$r_h(n) = \mathrm{IFT}[R_h(\omega)] \tag{2}$$

where

$$R_h(\omega) = \begin{cases} \mathrm{j}R(\omega), & -\pi \leqslant \omega < 0 \\ -\mathrm{j}R(\omega), & 0 \leqslant \omega < \pi \end{cases} \tag{3}$$

where IFT denotes the inverse Fourier transform, and $R(\omega)$ is the Fourier transform of $r(n)$. As shown in Fig. 8b, the Hilbert envelope approximately represents the strength of excitation. 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. Therefore, the strength of excitation represented by the Hilbert envelope shows a significant change at the transition from consonant to vowel, and hence can be used as a cue for detecting VOP.

Fig. 8 shows the speech waveform with manual marked VOPs, the Hilbert envelope of the LP residual, the VOP evidence, output of peak picking algorithm, and the hypothesized VOPs. The VOP evidence is obtained from the Hilbert envelope of the LP residual by multiplying it with the Gabor filter, and taking the sum of the product for every sample shift. From the VOP evidence plot as shown in Fig. 8c, the peaks are located using a peak picking algorithm. Spurious peaks are eliminated as shown in Fig. 8d and e, using the characteristics of the VOP evidence plot, namely, between two true VOP events, there exists a negative region of sufficient strength due to vowel region. The procedure for eliminating the spurious peaks in VOP evidence (Fig. 7c) is as follows: Take a window of VOP evidence, find the maximum and minimum evidence. Set a threshold for positive region (positive threshold = scale factor * max evidence) and negative region (negative threshold = scale factor * min evidence). If the evidence value is greater than the positive threshold, then it is detected as a VOP candidate. This will give rise to spurious VOPs. Therefore the negative threshold is used to eliminate them. If the VOP evidence value between two successive detected peaks do not fall below the negative threshold, the first peak is eliminated. It is possible to further reduce some of the spurious VOPs using the F_0 information.

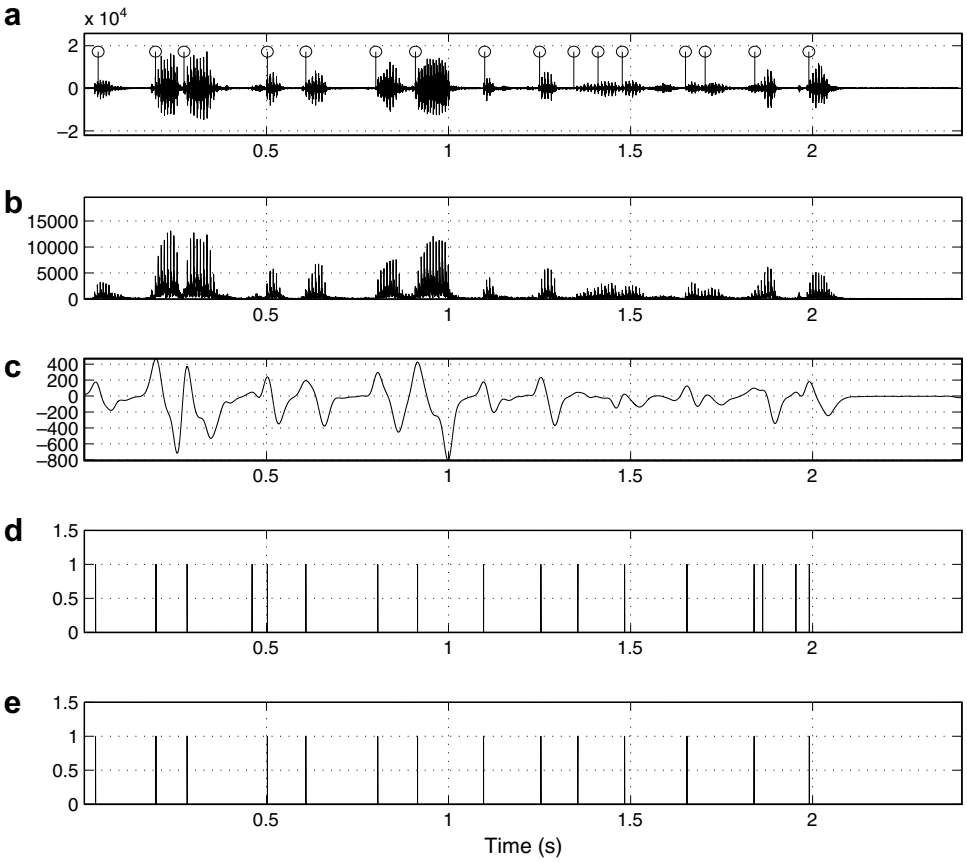

Fig. 8. (a) Speech waveform with manual marked VOPs, (b) Hilbert envelope of LP residual, (c) VOP evidence plot, (d) output of peak picking algorithm and (e) hypothesized VOP after eliminating few spurious peaks.

For example, the absence of voicing between VOP '10' and '11' shown in Fig. 7b helps to eliminate the spurious VOP '10'.

5.4. Feature parameterization

The association of syllable-like sequences with F_0 contour, as used in this study is shown in Fig. 9. The F_0 contour located within the region of two consecutive VOPs, is treated as one segment of F_0 contour. A set of parameters derived from F_0 movements, energy and duration are used for representing each segment.

The F_0 contour between two consecutive VOPs as shown in Fig. 10 corresponds to the F_0 movement in a syllable-like region, and it is treated as a segment of F_0 contour. The nature of F_0 variations for such a segment may be a rise, a fall, or a rise followed by a fall in most of the cases. We assume that more complex F_0 variations are unlikely within a segment. To represent the dynamics of the F_0 contour segment, we use tilt parameters (Taylor, 2000).

With reference to Fig. 10, tilt parameters, namely amplitude tilt (A_t), and duration tilt (D_t) are defined as follows:

$$A_t = \frac{|A_r| - |A_f|}{|A_r| + |A_f|} \tag{4}$$

$$D_t = \frac{|D_r| - |D_f|}{|D_r| + |D_f|} \tag{5}$$

where 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 rise and fall, respectively. Fig. 11a–f shows F_0 contours with different values of tilt. The use of tilt parameters help to represent F_0 patterns quantitatively, instead of quantizing and labeling of F_0 patterns as in other approaches (Shriberg et al., 2005; Adami and Hermansky, 2003).

Studies have shown that speakers can vary the prominence of pitch accents by varying the height of the F_0 maxima, to express different degrees of emphasis. Likewise, the listener's judgment of prominence reflect the role of F_0 variation in relation to prominence variation (Gussenhoven et al., 1997). From Fig. 11e and f, it is clear that the tilt parameters do not represent the height of F_0 peak. To express this, the difference between F_0 peak and F_0 valley ($\Delta F_0 = F_{0_p} - F_{0_v}$) is used in this study. The position of the F_0 peak (position of onset) makes difference in the perceptual prominence (Gussenhoven et al., 1997), and this is represented using the distance of F_0 peak (D_p) with reference to VOP.

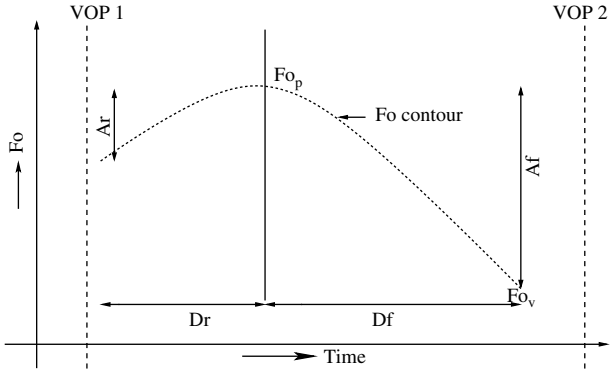

Fig. 10. A segment of F_0 contour. Tilt parameters A_t, D_t defined in terms of A_r, A_f, D_r, and D_f represent the dynamics of a segment of F_0 contour.

6. Prosodic features for language recognition

Languages differ in intonation, rhythm and stress characteristics. These differences are represented using parameters derived from F_0 contour, duration and energy.

6.1. Intonation

The direction of F_0 change, either rising or falling, is determined by the phonological patterns of the constituent words, which are language-specific. Our goal is to represent F_0 contour with suitable parameters to bring out the lan-

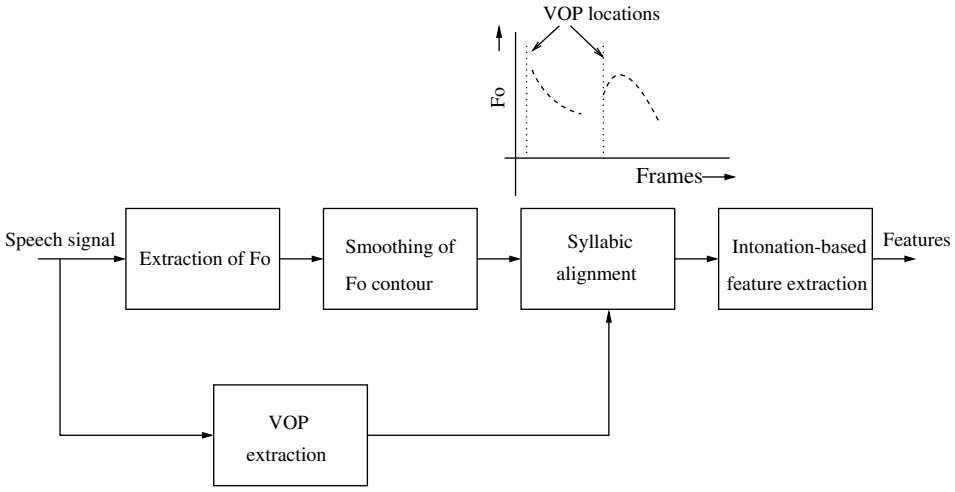

Fig. 9. Association of F_0 contour with locations of VOP for prosodic feature extraction.

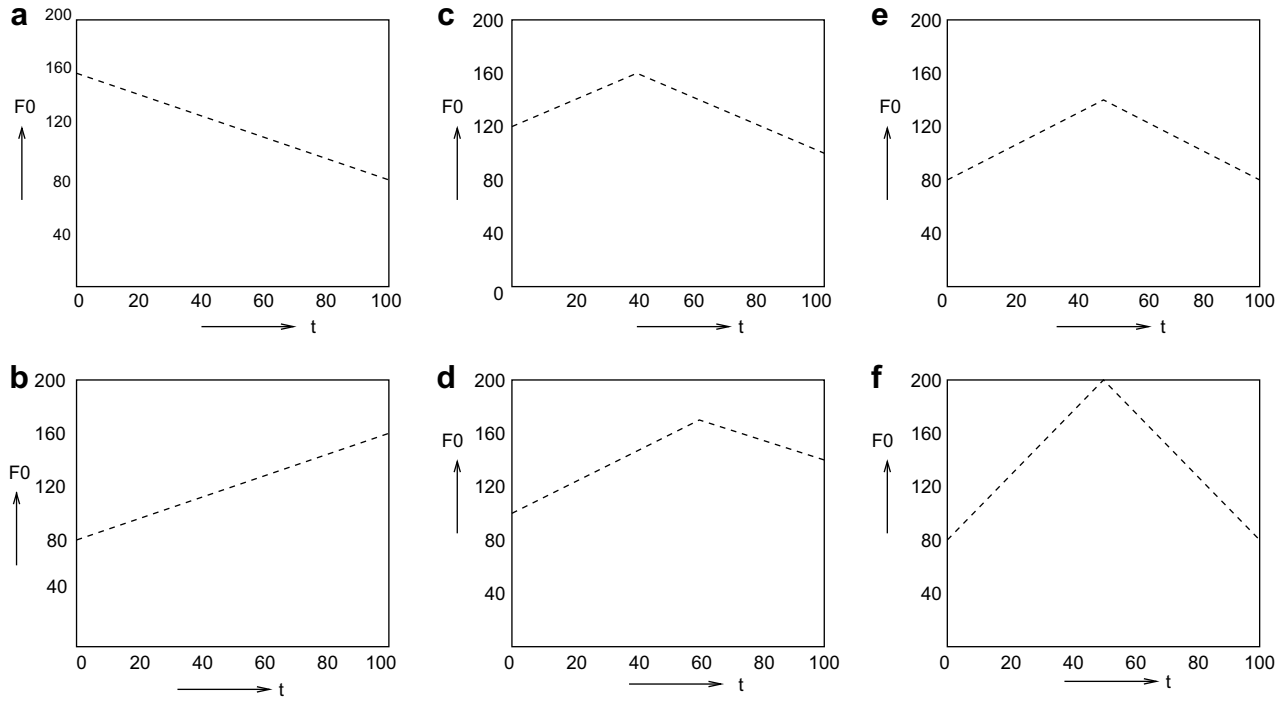

Fig. 11. Illustration of F_0 contours which are represented using various tilt parameters. (a) $A_t = -1$, $D_t = -1$; (b) $A_t = 1$, $D_t = 1$; (c) $A_t = -0.2$, $D_t = -0.2$; (d) $A_t = 0.4$, $D_t = 0.2$; (e) $A_t = 0$, $D_t = 0$ and (f) $A_t = 0$, $D_t = 0$.

guage-specific characteristics present in it. It was demonstrated that the tonal events are aligned to the segmental events such as the onset and/or offset of a syllable in German (Atterer and Ladd, 2004). In Mandarin, peaks of F_0 are found to be consistently aligned with the offset of the tone-bearing syllable in certain situations (Xu, 1998). In this work, F_0 contour is represented using the following measures:

1. Change in F_0 (ΔF_0).
2. Distance of F_0 peak with respect to VOP (D_p).
3. Amplitude tilt (A_t).
4. Duration tilt (D_t).

Absolute values of the frame level F_0 are more dependent on the physiological constraints, and hence are more speaker-dependent. Therefore, absolute F_0 values are not included in the feature set for language recognition studies.

6.2. Rhythm

The ability to distinguish among languages based on a signal which preserves low frequency information has been documented in infants as well as in adults (Ramus and Mehler, 1999). Two (correlated) variables, namely the proportion of vocalic intervals and the standard deviation of the duration of consonantal intervals, are identified as correlates of linguistic rhythm (Ramus et al., 1999). Both these measures will be directly influenced by segmental inventory and the phonotactic regularities of a specific language. A comparison of variability in parameters such as duration of vowels and duration of intervals between vowels has shown that durational variability is greater in stress-timed languages than syllable-timed (Grabe and Low, 2002).

In this work, we hypothesize that rhythm is perceived due to succession of syllables. Segmenting into syllable-like units enables representation of the rhythmic characteristics. We use the distance between successive VOPs (D_t) and the duration of voiced region (D_v) within each syllable-like region, to represent syllabic rhythm. A technique based on excitation information is used in the present study to detect the voiced regions (Chaitanya, 2005). The periodicity of the glottal closure instants in the excitation source information is explored to detect the voiced regions. The excitation information represented in the Hilbert envelope of the LP residual (Ananthapadmanabha and Yegnanarayana, 1979) as shown in Fig. 12b is used to detect the periodicity. The strength of the first major peak (after the center peak) in the normalized autocorrelation sequence of the Hilbert envelope is an indication of voicing in the segment. The normalized strength of the first peak is computed for every frame of the Hilbert envelope with a frame shift of one sample. The normalized peak strength values are used to decide whether a frame is voiced or not.

6.3. Stress

In all languages, some syllables are in some sense perceptually stronger than other syllables, and they are described as stressed syllables. The way stress manifests itself in the speech stream is highly language-dependent. Difference between strong and weak syllables is of some

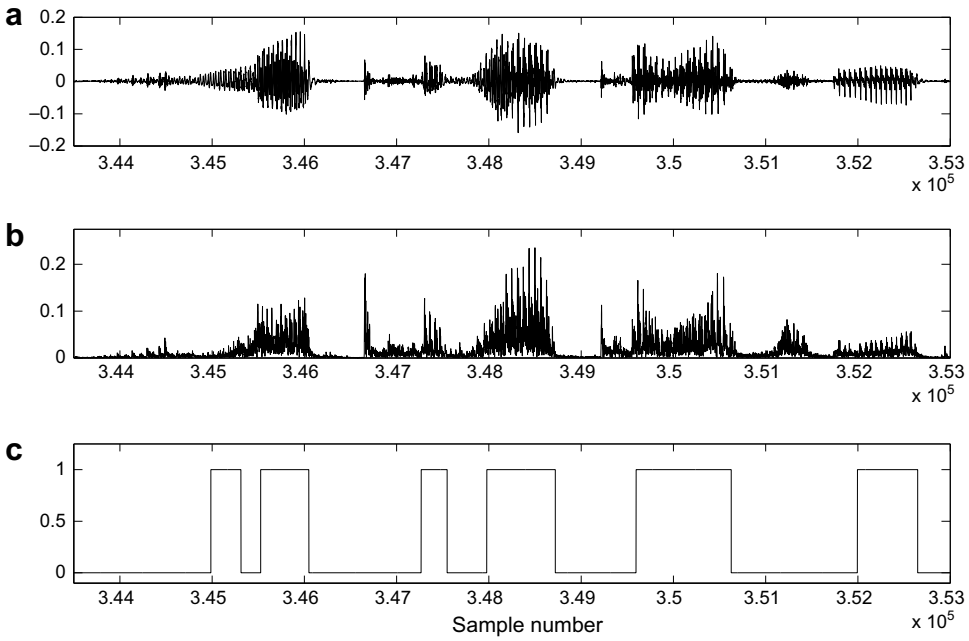

Fig. 12. Detection of voiced regions in speech using the strength and periodicity of excitation at the instants of glottal closure showing (a) speech signal (b) Hilbert envelope and (c) binary waveform having unity amplitude corresponding to the voiced regions.

linguistic importance in every language. However, languages differ in the linguistic function of such differences. Stress in many languages is what defines the rhythm of speech (Ashby and Maidment, 2005). In most of the languages, higher energy, larger F_0 movement and longer duration help to assign prominence to the stressed syllable relative to the surrounding syllable. We use change ΔE in log energy corresponding to the voiced regions of a syllable, along with the F_0 contour and the duration features mentioned above, to represent the stress.

6.4. Representation of prosodic features for language recognition

It has been observed that tones of adjacent syllables influence both the shape and height of the pitch contour of a particular syllable (Xu, 1999), and prominence of a syllable is estimated based on characteristics of F_0 contour around it (Gussenhoven et al., 1997). A syllable in isolation cannot be associated with rhythm, and hence a sequence of syllables is used to represent rhythm. Temporal dynamics of prosodic parameters are important to represent the prosodic variations among languages. As an approximation, we use context of a syllable, *i.e.*, characteristics of preceding and succeeding syllable along with that of present syllable to represent the language-specific prosody. This representation takes care of pitch variations which extend to the nearby syllables. When duration (approximated to distance between two successive VOPs) of a syllable exceeds certain threshold, it is hypothesized as probable word/phrase boundary or caused due to a long pause, and such syllables are not used. Since the specific interaction between pitch variations, intensity and duration play an important role in determining the prosody, the parameters representing F_0 contour, duration and energy are combined together to represent prosody.

6.5. Results from experimental study

To demonstrate the effectiveness of the proposed prosodic features for language recognition, an experimental study was conducted using NIST 2003 language recognition evaluation (LRE) database (website, http://www.nist.gov/speech/tests/lang/2003/). The task is to detect the presence of a hypothesized language, given a segment of conversational speech recorded over the telephone channel. The target languages include Arabic, English, Farsi, French, German, Hindi, Japanese, Korean, Mandarin, Spanish, Tamil and Vietnamese from the CallFriend Corpus. Both development data and evaluation data of NIST LRE 1996 are used as the development data for NIST LRE 2003 task. The NIST LRE 2003 evaluation set used in this experimental study contains 80 speech files from each of the 12 target languages, each of 30 s duration from the CallFriend Corpus. In addition to this, there are four sets from other conversational speech sources, namely, 80 Russian segments from the CallFriend Corpus, 80 Japanese segments from CallHome Corpus, 80 English segments from Switchboard-I Corpus, and 80 English segments from Switchboard Cellular Corpus. Equal error rate (EER) and detection error tradeoff (DET) curves are used as measures for evaluating the performance of the system.

A 21-dimension feature vector is formed by concatenating prosodic features derived from three consecutive syllable-like units. The seven parameters from each syllable-like unit consists of distance between successive VOPs (D_t),

voiced duration (D_v), change in F_0 (ΔF_0), distance of F_0 peak with reference to VOP (D_p), amplitude tilt (A_t), duration tilt (D_t) and change in log energy (ΔE). A multilayer feedforward neural network (MLFFNN) classifier is trained for 500 epochs using prosodic feature vectors as shown in Fig. 13. During training, one of the MLFFNN outputs is set to one (this particular output denotes the language identity of the training vector), while all others are set to zero. The structure of MLFFNN used is *21L 64N 16N 12N*, where *L* represent units with linear activation function, *N* represent units with nonlinear activation function, and the numerals represent the number of units in the layers.

For testing, similar 21-dimension feature vectors are derived from the test utterance. These test vectors are applied (one by one) to the input of already trained MLFFNN classifier, and evidence of different languages at the output are noted. The evidences obtained for all the feature vectors in the test utterance are averaged to obtain the confidence scores for each language. Performance of the proposed language recognition system evaluated using NIST 2003 evaluation set is shown using the DET curve in Fig. 14. The system results in an EER of 32% which is close to the results of other prosody-based system performance (Adami and Hermansky, 2003).

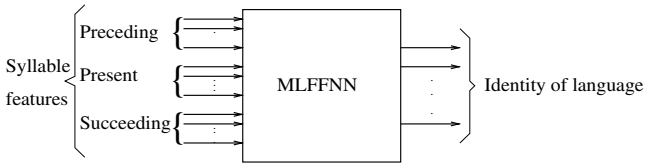

Fig. 13. Prosody-based neural network classifier for language recognition. Input feature vector consists of prosodic features derived from three consecutive syllables.

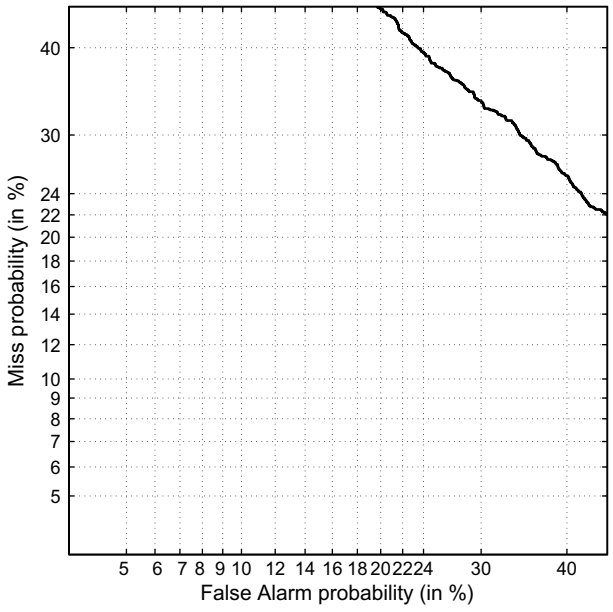

Fig. 14. DET curve showing the performance of prosody-based language identification system for NIST 2003 language recognition evaluation database.

7. Prosodic features for speaker recognition

Human beings use several levels of perceptual cues for speaker recognition ranging from high-level cues such as semantics, pronunciations, idiosyncrasies and prosody, to low-level cues such as acoustic aspect of speech (Heck, 2002). Prosodic cues such as pitch gestures, accent and stress characteristics reflect the physiological as well as habitual aspect of a speaker.

7.1. Speaker-specific prosodic features

The coordination of laryngeal and supralaryngeal movements limits how syllables and tone can be aligned to each other (Xu et al., 2002). Studies have indicated that listeners are more sensitive to variations in F_{0_p} than F_{0_v} (Gussenhoven et al., 1997). Hence change in F_0 (ΔF_0), distance of F_0 peak with reference to VOP (D_p) and peak value of F_0 (F_{0_p}) for each segment of F_0 contour may be useful for speaker recognition. An increase in F_0 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 segment of F_0 contour may reflect some physiological as well as habitual aspect of a speaker. The change in log energy (ΔE) along with F_0 change gives a quantitative measure of stress characteristics, therefore may be specific to a particular speaker. The F_0, duration and energy related parameters used in this study for characterizing the speaker-specific aspect of prosody are the following:

(a) F_0 mean (F_{0_μ}).
(b) F_0 peak (F_{0_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 syllable-like region between two consecutive VOPs is represented using the above mentioned parameters to form a seven-dimension feature vector.

7.2. Results from experimental studies

In order to use long-term features such as prosody for speaker recognition, 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 (website, http://www.nist.gov/speech/tests/spk/2001/). The effectiveness of the proposed prosodic features is demonstrated using the first subset of NIST 2003 extended data task (website, http://www.nist.gov/speech/tests/spk/2003/). Unlike the

traditional speaker recognition tasks, the extended data task provides more speech data for training (4-side/8-side/16-side, where each conversation side contains approximately 2.5 min of speech). Each target model is tested with a set of 1-side test utterances, where the task is to find out whether the particular test utterance belongs to the target speaker or not. We have chosen the first split in NIST 2003 extended data task for this study. It consists of 137, 54 and 74 speaker models for the 16-side, 8-side, and 4-side cases, respectively. The models are evaluated using 1076, 1238 and 1258 test utterances for the 16-side, 8-side and 4-side cases, respectively.

We hypothesize that the distribution of syllable level prosodic feature vectors are speaker-specific. 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 (Haykin, 1999; Yegnanarayana, 1999). It consists of an input layer, an output layer and one or more hidden layers. To capture the distribution of the feature vectors, examples are presented in a random order to the AANN and the network is trained using standard backpropagation algorithm (Yegnanarayana, 1999; Haykin, 1999). It has been demonstrated that the AANN has the ability to capture the distribution of input data (Yegnanarayana and Kishore, 2002).

To illustrate the speaker-specific distribution, the seven-dimensional prosodic feature vectors derived from speech corresponding to two male speakers in NIST 2003 database is compressed and then plotted. Fig. 15 shows the distribution of nonlinearly compressed prosodic feature vectors. Here the nonlinear compression is obtained using autoassociative neural network (AANN) model with a structure *7L 14N 3N 14N 7L* where the compressed feature vector is obtained from the dimension-compression hidden (middle) layer having three units.

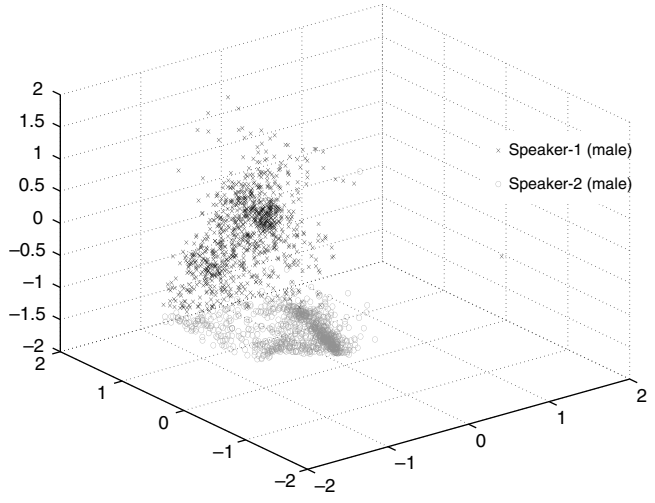

Fig. 15. Compressed prosodic feature vectors for two male speakers (taken from NIST 2003 extended data).

During testing, for each prosodic feature vector 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 ith 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 ith syllable, and N is the number of syllables in the test utterance.

For each target speaker, one AANN model is trained for 500 epochs to capture the distribution of prosodic feature vectors. 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 units with linear activation function, N represent units with nonlinear activation function, and the numerals represent the number of units in the layers. A set of background models built from a known set of impostor speakers (taken from another split of the same database) 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. 16.

Score normalization is used for scaling the likelihood scores, which helps to find a global speaker independent threshold for the decision making process. Feature vectors obtained from the test utterance is presented to the target speaker model as well as to a set of background models as shown in Fig. 16. For each test utterance, the decision on the gender is made based on the average score of male/female background model set. The raw score obtained from target speaker model is test normalized (Auckenthaler et al., 2000) using scores of the background (BG) models. The normalized score C_n is computed from raw score C as

$$C_n = (C - \mu_g)/\sigma_g \tag{6}$$

where μ_g and σ_g represent mean and standard deviation of BG scores corresponding to the hypothesized gender of test utterance.

Our prosody-based speaker verification system resulted in an EER of 12.4%, 15% and 23% for 16-side, 8-side and 4-side conversational cases of the particular data set, respectively. Performance is shown using the DET curves in Fig. 17. Better performance in case of 16-side and 8-side training cases show that more training speech is required for capturing the prosodic characteristics well. Performance of our prosody-based system is close to the results reported for NIST extended task 2001 using features derived without using ASR (Adami and Hermansky, 2003; Adami et al., 2003; Peskin et al., 2003).

7.3. Combining evidence from prosody and spectral-based systems

As spectral features are vulnerable to channel mismatch and noise, the use of prosodic features less affected by these factors, can play important role in improving the

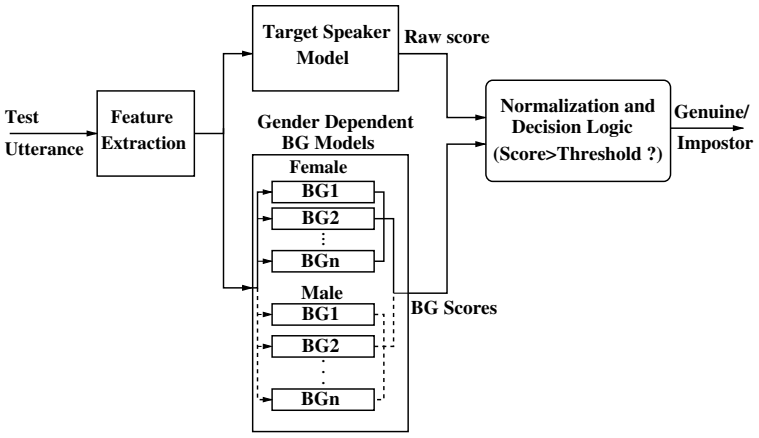

Fig. 16. Block diagram of prosody-based speaker verification system, showing the testing of an unknown utterance against target speaker model and a set of background models.

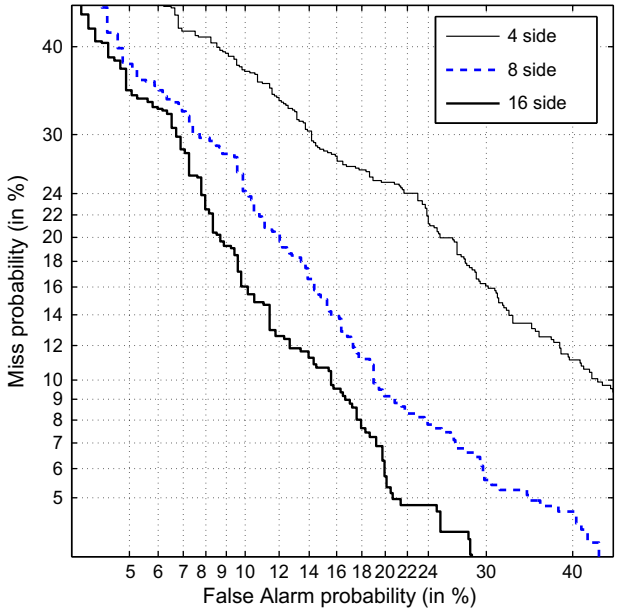

Fig. 17. DET curve showing the performance of prosody-based system for 16-side, 8-side and 4-side conversational cases.

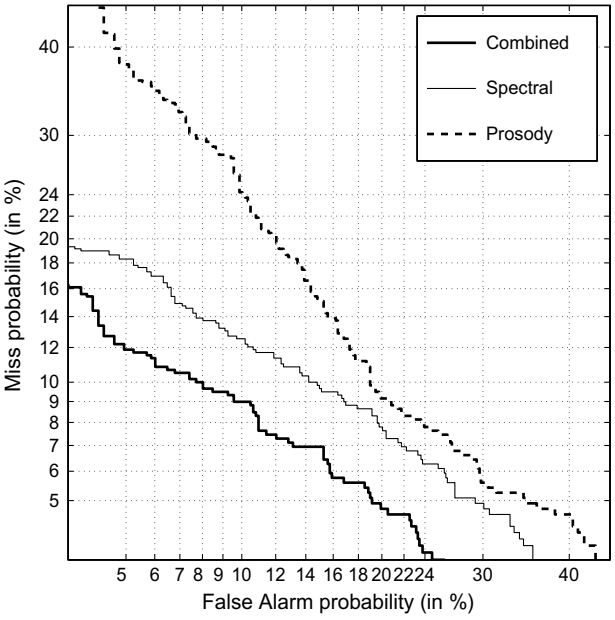

Fig. 18. DET curve showing the performance of spectral-based system, prosody-based system and combined system for 8-side conversational case.

robustness of the speaker recognition system. The evidence about the speaker from different features may be combined in several ways to achieve better performance. One simple approach is the weighted addition of evidences from different systems. Our spectral-based (WLPCC) baseline speaker verification system (Yegnanarayana and Kishore, 2002) gives an EER of 11.8% for the same 8-side data set. Prosody-based evidence provide complementary information while combining with the spectral-based evidence. Combining by simple addition results in an EER of 9.3%, showing the presence of complementary information in these features as illustrated in Fig. 18.

8. Summary and conclusions

In this paper, we have presented a new method for extracting prosodic features from the speech signal, useful for applications such as language and speaker recognition. This approach eliminates the requirement of automatic speech recognizer for prosodic feature extraction, but still gives a meaningful association of prosodic features with the corresponding syllable sequence. This is done with the knowledge of VOPs, detected automatically from the Hilbert envelope of the LP residual of the speech signal. The region between two successive VOPs is considered as a syllable-like region, and parameters are derived to represent duration, dynamics of F_0 contour and energy variations corresponding to each region.

We have evaluated the effectiveness of prosodic features extracted using the proposed approach for language recognition in the case of NIST LRE 2003 task. Though the success of language recognition was constrained by the limited

speech data available for training, it clearly illustrates the potential of prosodic features for distinguishing languages.

For evaluating the potential of prosodic features for speaker verification, a study was conducted using NIST SRE 2003 extended data. The performance seems to be significant, especially for cases where more speech data was available for training the models. The complementary nature of the prosodic features and spectral features helps to improve the overall performance of speaker verification, while combining evidence from prosody-based and spectral-based systems.

References

Abercrombie, D., 1967. Elements of General Phonetics. Edinburgh University Press, Edinburgh.

Adami, A.G., Hermansky, H., 2003. Segmentation of speech for speaker and language recognition. In: Proc. EUROSPEECH, Geneva. pp. 841–844.

Adami, A.G., Mihaescu, R., Reynolds, D.A., Godfrey, J.J., 2003. Modeling prosodic dynamics for speaker recognition. In: Proc. IEEE Int. Conf. Acoust., Speech Signal Process., Vol. 4, Hong Kong, China. pp. 788–791.

Ananthapadmanabha, T.V., Yegnanarayana, B., 1979. Epoch extraction from linear prediction residual for identification of closed glottis interval. IEEE Trans. Speech Audio Process., 309–319.

Ashby, M., Maidment, J., 2005. Introducing Phonetic Science. Cambridge University Press, Cambridge.

Atterer, M., Ladd, D.R., 2004. On the phonetics and phonology of "segmental anchoring of F_0: evidence from German. J. Phonetics 32, 177–197.

Auckenthaler, R., Carey, M., Thomas, H.L., 2000. Score normalization for text-independent speaker verification systems. Digit. Signal Process. 10, 42–54.

Chaitanya, M., 2005. Single channel speech enhancement. MS Thesis, Department of Computer Science and Engineering, Indian Institute of Technology Madras.

Cummins, F., Gers, F., Schmidhuber, J., 1999. Comparing prosody across languages. I.D.S.I.A. Technical Report IDSIA-07-99, Istituto Molle di Studie sull'Intelligenza Artificiale, CH6900 Lugano, Switzerland.

Cutler, A., Ladd, D.R., 1983. Prosody: Models and Measurements. Springer-Verlag, Berlin, Heidelberg, New York, Tokyo.

Doddington, G., 2001. Speaker recognition based on idiolectic differences between speakers. In: Proc. EUROSPEECH, Aalborg, Denmark. pp. 2521–2524.

Fox, A., 2000. Prosodic Features and Prosodic Structure. Oxford University Press.

Furui, S., 1981. Cepstral analysis technique for automatic speaker verification. IEEE Trans. Speech Audio Process. 29, 254–272.

Grabe, E., Low, E.L., 2002. Durational variability in speech and the rhythm class hypothesis. In: Papers Laboratory Phonology, Vol. 7. pp. 515–546.

Gussenhoven, C., Reepp, B.H., Rietveld, A., Rump, H.H., Terken, J., 1997. The perceptual prominence of fundamental frequency peaks. J. Acoust. Soc. Amer. 102 (5), 3009–3022.

Haykin, S., 1999. Neural Networks: A Comprehensive Foundation. Prentice-Hall International, New Jersey.

Heck, L.P., 2002. 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>.

Hirst, D., Di Cristo, A., 1998. Intonation Systems: A Survey of Twenty Languages. Cambridge University Press, Cambridge.

Hyman, L.M., 2005. Word-prosodic typology. US Berkeley Phonology Lab Annual Report. pp. 164–184.

Kometsu, M., Mori, K., Arai, T., Murahara, Y., 2001. Human language identification with reduced segmental information: comparison between Monolinguals and Bilinguals. In: Proc. EUROSPEECH, Vol. 1, Scandinavia. pp. 149–152.

Krakow, R.A., 1999. Physiological organization of syllables: a review. J. Phonetics 27, 23–54.

MacNeilage, P.F., 1998. The frame/content theory of evolution of speech production. Behav. Brain Sci. 21, 499–546.

Mahadeva Prasanna, S.R., Gangashetty, S.V., Yegnanarayana, B., 2001. Significance of vowel onset point for speech analysis. In: Proc. Int. Conf. Signal Process. Comm., Vol. 1, Bangalore, India. pp. 81–86.

Makhoul, J., 1975. Linear prediction: a tutorial review. Proc. IEEE 63, 561–580.

Mori, K., Toba, N., Harada, T., Arai, T., Kometsu, M., Aoyagi, M., Murahara, Y., 1999. Human language identification with reduced spectral information. In: Proc. EUROSPEECH, Vol. 1, Budapest, Hungary. pp. 391–394.

Nagarajan, T., Murthy, H.A., 2006. Language identification using acoustic log-likelihoods of syllable-like units. Speech Comm. 48, 913–926.

Navratil, J., 2001. Spoken language recognition – a step toward multilinguality in speech processing. IEEE Trans. Speech Audio Process. 9 (6), 678–685.

Peskin, B., Navratil, J., Abramson, J., Jones, D., Klusacek, D., Reynolds, D., Xiang, B., 2003. Using prosodic and conversational features for high-performance speaker recognition: report from JHU WS'02. In: Proc. IEEE Int. Conf. Acoust. Speech Signal Process. Vol. 4, Hong Kong, China. pp. 792–795.

Ramus, F., Mehler, J., 1999. Language identification with suprasegmental cues: A study based on speech resynthesis. J. Acoust. Soc. Amer. 105 (1), 512–521.

Ramus, F., Nespor, M., Mehler, J., 1999. Correlates of linguistic rhythm in speech signal. Cognition 73 (3), 265–292.

Reynolds, D.A., 1996. The effect of handset variability on speaker recognition performance: experiments on the Switchboard corpus. In: Proc. IEEE Int. Conf. Acoust. Speech Signal Process., Vol. 1, Atlanta, GA, USA. pp. 113–116.

Reynolds, D.A., Rose, R., 1995. Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Trans. Speech Audio Process. 3, 72–83.

Reynolds, D., Andrews, W., Campbell, J., Navratil, J., Peskin, B., Adami, A., Jin, Q., Klusacek, D., Abramson, J., Mihaescu, R., Godfrey, J., Jones, D., Xiang, B., 2003. The SuperSID project: exploiting high-level information for high-accuracy speaker recognition. In: Proc. IEEE Int. Conf. Acoust. Speech Signal Process., Vol. 4, Hong Kong, China. pp. 784–787.

Roach, P., 1983. English Phonetics and Phonology. Cambridge University Press, Cambridge.

Rouas, J., Farinas, J., Pellegrino, F., Andre-Obrecht, R., 2005. Rhythmic unit extraction and modelling for automatic language identification. Speech Comm. 47, 436–456.

Shriberg, E., Stolcke, A., Hakkani-Tur, D., Tur, G., 2000. Prosody-based automatic segmentation of speech into sentences and topics. Speech Comm. 32, 127–154.

Shriberg, E., Ferrer, L., Kajarekar, S., Venkataraman, A., Stolcke, A., 2005. Modeling prosody for speaker recognition. Speech Comm. 46, 455–472.

Sonmez, M.K., Shriberg, E., Heck, L., Weintraub, M., 1998. Modeling dynamic prosodic variation for speaker variation. In: Proc. Int. Conf. Spoken Language Process., Vol. 7, Sydney, Australia. pp. 3189–3192.

Taylor, P., 2000. Analysis and synthesis of intonation using the tilt model. J. Acoust. Soc. Amer. 107 (3), 1697–1714.

Thyme-Gobbel, A.E., Hutchins, S.E., 1996. On using prosodic cues in automatic language identification. In: Proc. Int. Conf. Spoken Language Process., Vol. 3, Philadelphia, PA, USA. pp. 1768–1772.

Xu, Y., 1998. Consistency of tone-syllable alignment across different syllable structures and speaking rates. Phonetica 55, 179–203.

Xu, Y., 1999. Effects of tone and focus on the formation and alignment of f_0 contours. J. Phonetics 27, 55–105.

Xu, Y., Xuejing, S., 2002. Maximum speed of pitch change and how it may relate to speech. J. Acoust. Soc. Amer. 111 (3), 1399–1413.

Yegnanarayana, B., 1999. Artificial Neural Networks. Prentice-Hall of India, New Delhi.

Yegnanarayana, B., Kishore, S.P., 2002. AANN – An alternative to GMM for pattern recognition. Neural Networks 15 (3), 459–469.

Zissman, M.A., 1996. Comparison of four approaches to automatic language identification of telephone speech. IEEE Trans. Speech Audio Process. 4 (1), 31–44.