IEEE-International Conference on Signal processing, Communications and Networking
Madras Institute of Technology, Anna University Chennai India, Jan 4-6, 2008. pp65-70

Prosodic Features For Language Identification

Leena Mary [1], B.Yegnanarayana [2]

[1] Asst. Professor, Rajiv Gandhi Institute of Technology (Government Engineering College),
Kottayam, Kerala – 686 501
[2] Professor, Indian Institute of Information Technology
Hyderabad - 500 032
leenmary@rediffmail.com, yegna@iiit.ac.in

Abstract: **In this paper, we examine the effectiveness of prosodic features for language identification. Prosodic differences among world languages include variations in intonation, rhythm, and stress. These variations are represented using features derived from fundamental frequency (F_0) contour, duration, and energy contour. For extracting the prosodic features, speech signal is segmented into syllable-like units by locating vowel-onset points (VOP) automatically. Various parameters are then derived to represent F_0 contour, duration, and energy contour characteristics for each syllable-like unit. The features obtained by concatenating the parameters derived from three consecutive syllable-like units are used to represent the prosodic characteristics of a language. The prosodic features thus derived from different languages are used to train a multilayer feedforward neural network (MLFFNN) classifier for language identification. The effectiveness of the proposed approach is verified using Oregon Graduate Institute (OGI) multi-language telephone speech corpus and National Institute of Science and Technology (NIST) 2003 language identification database.**

Index: **Terms: language identification, prosody, intonation.**

I. Introduction

Automatic language identification (LID) is the task of identifying the language of a given utterance of speech using a machine. Applications of LID fall in two main categories: Pre-processing for machines and pre-processing for human listeners [1]. A multilingual voice-controlled information retrieval system is an example of the first category. Language identification system used to route an incoming telephone call to a human operator at a switchboard, fluent in the corresponding language, is an example of the second category [1]. For such multilingual applications, the machine should be capable of distinguishing among languages.

Prosody refers to certain characteristics that lend naturalness to human speech. The variation of the pitch provides some melodic properties to speech, and this controlled modulation of pitch is referred as *intonation*. The duration of sound units are varied (shortened or lengthened) in accordance to some underlying pattern, giving *rhythm* to speech. Some syllables or words are made more prominent than others, resulting in *stress*. The information gleaned from the melody, timing, and stress in speech increases the intelligibility of spoken message, and also conveys information such as lexical tone, accent, and emotion. The characteristics that make us perceive these effects are collectively referred to as prosody. Prosody has a great deal to offer for effective human language identification.

Much of the LID research so far has placed its emphasis on spectral information, mainly using the acoustic features of sound units (referred as acoustic phonetics), and their alignment (referred as phonotactics). Such systems may perform well in similar acoustic conditions [2]. But their performance degrades due to noise and channel mismatch. Prosodic features derived from pitch contour, amplitude contour and duration are relatively less affected by channel variations and noise. This motivated us to explore prosodic features, for language identification. Though the systems based on spectral features outperform the prosody-based LID systems, their combined performance may provide the needed robustness. To make use of the language-specific information present in prosody, most of the existing LID systems use segment boundaries and text labels obtained using speech recognizers. In this paper, we propose an approach for LID using prosodic features derived directly from the acoustic speech signal, without the use of transcription of the signal. The rest of the paper is organized as follows: Section 2 describes the prosodic differences among languages. In Section 3, automatic extraction and representation of language-specific prosody is described. Section 4 explains the representation of prosodic features. In Section 5, we discuss the results based on our experimental studies. Section 6 summarizes the work.

II. Prosodic differences among languages

Languages can be broadly categorized as stress-timed and syllable-timed, based on their timing/rhythmic properties. In stress-timed languages like English and German, duration of the syllables are

978-1-4244-1923-4/08$25.00 ©2008 IEEE

mainly controlled by the presence of stressed syllables which may occur at random. In stress-timed languages, roughly constant separation (in terms of time) is maintained between two stressed syllables. Syllables that occur between two stressed syllables are shortened to accommodate this property. In syllable-timed languages such as French and Spanish, the durations of syllables remain almost constant. 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. There is yet another categorization of languages as tonal and nontonal, based on the tonal properties of a language. 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 [3]. There are many other languages which strictly do not follow the rules of a class, which means that these classifications are rather a continuum. Therefore languages may differ in terms of intonation, rhythm, and stress.

A. Intonation

Pitch is a perceptual attribute of sound which can be described as a sensation of the relative ``altitude" of sound. The physical correlate of pitch is the fundamental frequency (F_0). The difference in F_0 contour between languages is illustrated for the case of two languages, namely Farsi and Mandarin in Figure 1. It can be observed that in general Mandarin has large variations in F_0 values compared to Farsi, in spite of the variations in speaker characteristics.

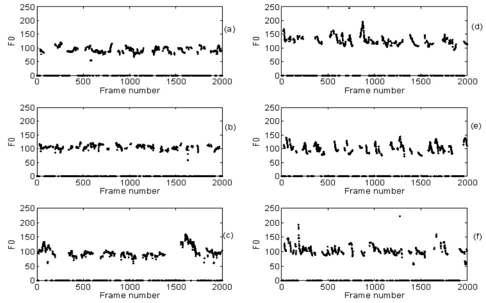

Figure 1: *Variation in dynamics of F_0 contour for utterances in Farsi and Mandarin, spoken by three male speakers each. (a), (b) and (c) correspond to Farsi (d), (e) and (f) correspond to Mandarin utterances (taken from OGI database).*

B. Rhythm

The ability to distinguish languages based on rhythm has been documented in infants as well as in adults [4]. Two (correlated) variables defined over an utterance, namely the proportion of vocalic intervals and the standard deviation of the duration of consonantal intervals, are identified as correlates of linguistic rhythm [5]. Both these measures will be directly influenced by segmental inventory and the phonotactic regularities of a specific language.

C. Stress

In all languages, some syllables are in some sense perceptually stronger than other syllables, and they are described as stressed syllables. In most of the languages, higher intensity, larger pitch variation and longer duration help to assign prominence to stressed syllables. But the position of stressed syllable in a word varies from language to language. English is a stress-timed language, where stressed syllables appear roughly at a constant rate, and unstressed syllables are shortened to accommodate this. In some languages, stress is always placed on a given syllable, as in French, where the words are always stressed in the last syllable. In English and French, a longer duration syllable carries more pitch movements. But such a correlation may not hold equally well for all languages. 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 languages [6].

III. AUTOMATIC EXTRACTION OF PROSODIC FEATURES

Approaches for extraction of prosodic features can be broadly categorized based on the use of automatic speech recognizer (ASR) as (1) ASR-based approach and (2) ASR-free approach. The ASR-based approach uses segment boundaries obtained from ASR, for extracting the prosodic features [7]. But for applications like language identification, the use of ASR may not be needed. In the second approach, inflection points and start or end of voicing of pitch are used for segmentation [8]. The pitch contour dynamics is then represented using parameters derived from linear stylized pitch segments [9]. This approach has the advantage that prosodic features can be derived directly from the speech signal. In both the approaches, the segmented trajectories are quantized and represented using a small set of labels that describe the dynamics of pitch and energy. The n-grams of these labels are formed to model the characteristics of a speaker or a language. In this work, we use a new technique for extraction and representation of prosodic features. The proposed method utilizes the location of vowel onset points (VOP) for identifying the syllable-like regions in continuous speech. This method combines the salient features of the existing approaches mentioned above, namely, the association with the syllabic pattern as in the first approach, and the extraction of features without using ASR as in the second approach.

For representing syllable-based rhythm, intonation, and stress, the speech signal should be segmented into syllables. 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, it is accomplished with the knowledge of VOPs as illustrated in Figure 2, where the 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]. The availability of pitch values helps in further reduction of spurious VOPs. For example, the absence of voicing between two VOPs numbered as `10' and `11' shown in Figure 2(b), helps to eliminate the spurious peak `10'.

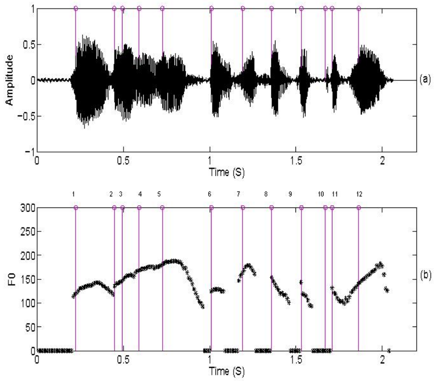

Figure 2: *(a) Segmentation of speech into syllable-like units using VOPs. (b) F0 contour associated with VOPs.*

IV. Representation of Prosodic Features

In this work, the locations of VOP are used for segmenting speech into syllable-like units. The locations of VOP are then associated with F_0 contour for extracting the prosodic features.

A. Representation of Intonation

The F0 contour between two consecutive VOPs (as shown in Figure 3 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. With reference to Figure 3, tilt parameters [11], namely amplitude tilt (A_t) and duration tilt (D_t) for a segment of F_0 contour are defined as follows:

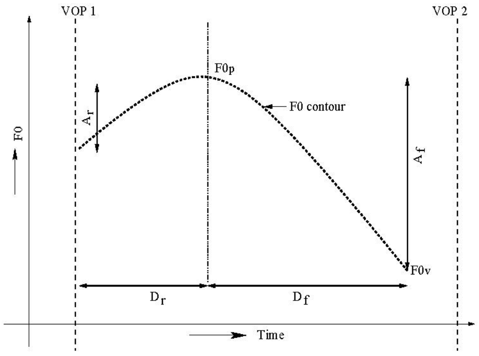

Figure 3: *A segment of F_0 contour. Tilt parameters A_t and D_t defined in terms of A_r, A_f, D_r, and D_f represent the dynamics of a segment of F_0 contour.*

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

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

where A_r and A_f represent the rise and fall in F_0 amplitude, respectively, with respect to peak value of fundamental frequency F_{0p}. Similarly D_r and D_f represent the duration taken for rise and fall respectively.

Studies have shown that, speakers can vary the prominence of pitch accents by varying the height of the fundamental frequency peak, to express different degrees of emphasis. Likewise, the listener's judgment of prominence reflects the role of F_0 variation in relation to variation in prominence [12]. To express the height of the F_0 peak, the difference between peak and valley fundamental frequency ($\Delta F_0 = F_{0p} - F_{0v}$) is used in this study. It has been observed that the length of the F_0 peak (length of onset) has a role in the perceptual prominence [12]. In this study, this is represented using the distance of F_0 peak location with respect to VOP (D_p).

In summary, the intonation features used for this language identification study are the following:

(a) Change in F_0 (ΔF_0)
(b) Distance of F_0 peak with respect to VOP (D_p)
(c) Amplitude tilt (A_t)
(d) Duration tilt (D_t)

B. Representation of Rhythm

In this work, we hypothesize that rhythm is perceived due to closing and opening of the vocal tract in the succession of syllables. The proportion of voiced intervals within each syllable region gives a measure of this transition. Segmenting continuous speech into syllable-like units enables representation of the rhythmic characteristics. We use the duration of syllable (D_s) (approximated to the distance between successive VOPs) and the duration of voiced region (D_v), to represent rhythm. The following features are used to represent rhythm:

(a) Syllable duration (D_s)
(b) Duration of voiced region (D_v) within each syllable

C. Representation of Stress

The syllable carrying stress is prominent with respect to the surrounding syllables, due to its loudness, large movement of F_0 and/or longer duration. Therefore along with F_0 and duration features mentioned above, we use change in log energy (ΔE) within voiced region to represent the stress.

D. Representation Of Prosody

It has been observed that tones of adjacent syllables influence both the shape and height of the F_0 contour of a particular syllable, and prominence of a syllable is estimated based on the pitch characteristics of the contour around it [12]. Similarly, rhythm is formed by a sequence of syllables, and a syllable in isolation cannot be associated with rhythm. Therefore temporal dynamics of these parameters are important while representing the prosodic variations among languages. The context of a syllable, i.e., the characteristics of preceding and succeeding syllable is used in this work to represent language-specific prosody. Since the specific interaction between pitch movements, energy, and duration play an important role in determining the prosody, these parameters together are used to form a prosodic feature vector.

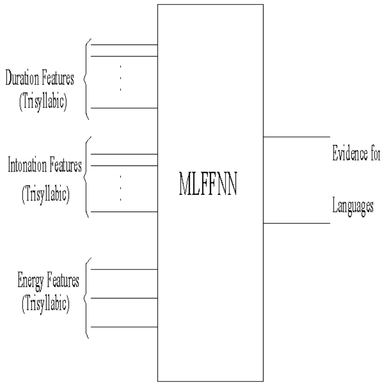

Figure 4: *Prosody-based neural network classifier for language identification.*

V. Results from Experimental Studies

To demonstrate the effectiveness of prosodic features mentioned above for language identification, a study was conducted using the OGI database. The OGI multi-language telephone-based speech (MLTS) corpus consists of 11 languages with an average of 90 speakers per language [13]. It consists of English (En), Farsi (Fa), French (Fr), German (Ge), Hindi (Hi), Japanese (Ja), Korean (Ko), Mandarin (Ma), Spanish (Sp), Tamil (Ta), and Vietnamese (Vi). This set includes representatives of the tonal (Mandarin, Vietnamese), and stress-timed (English, German) and syllable-timed (French, Spanish), as well as languages whose affiliations are less clear.For all the languages, 40 speech (unrestricted spontaneous) files (each with an average duration 45 sec) corresponding to 40 different speakers are used for training. An average of 20 speech files from different speakers are used for evaluating the proposed LID system. The training and testing speaker sets are different. Separate MLFFNN models are trained for each language pair in the OGI database. The structure of MLFFNN is *21L 64N 6N 2N*, 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 example, to build a model for discriminating English from Mandarin, an MLFFNN classifier is trained with examples from English and Mandarin, with output set to {+1,-1}, and {-1,+1), respectively as shown in Figure 4. During testing, for each prosodic vector in the test utterance, evidence of different languages at the output of MLFFNN classifier is noted. Then evidences obtained for all the prosodic feature vectors in the test utterance are averaged to obtain the confidence scores for each language.

The results of pair-wise language discrimination task on OGI database are given in Table 1. The results of two recent studies are given in square bracket for comparison [14, 15]. It is observed that prosodic features are more effective for discriminating languages that fall into different categories based on rhythm or tonal characteristics. For example, Japanese and Mandarin are discriminated very well from other languages, whereas, discrimination between them is somewhat poor. Due to the limited size of the speech data available in the OGI database, it was difficult to extend this study to multi-class LID problem, as noted by the other researchers [14, 15].

To demonstrate the effectiveness of the proposed prosodic features for a multiclass LID problem, 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 sec 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 multilayer feedforward neural network is trained for 500 epochs using prosodic feature vectors as shown in Figure 4. The structure of MLFFNN is *21L 64N 16N 12N*. Performance of the proposed language identification system is evaluated using NIST 2003 evaluation set is shown using the DET curve in Figure 5. The system results in an EER of 32 which is close to the results of other prosody-based system performance [8].

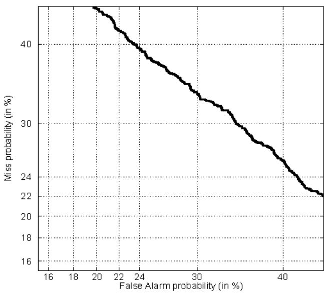

Figure 5: *DET curve showing the performance of prosody-based language identification for NIST 2003 language recognition evaluation database of 12 languages.*

VI. Summary

The effectiveness of prosodic features derived from the speech signal is demonstrated. The results on the OGI database show that prosodic features are effective for discriminating languages. The performance is better for languages that fall into different categories based on rhythm/tonal characteristics. We have also demonstrated the effectiveness of prosodic features for language identification in the case of NIST LRE 2003 task. Though the success was constrained by the limited speech data available for training, it clearly illustrates the potential of prosodic features for distinguishing languages.

Table 1. *Performance of pair-wise language discrimination task on OGI database. The entries from column 2 to 11 denote the percentage of test utterances identified correctly, for a model corresponding to the languages in first column and first row. For comparison, results of Rouas's [14] and Lin's [15] work are given in square brackets. Hindi was not included in their studies.*

Lang.	Fa	Fr	Ge	Hi	Ja	Ko	Ma	Sp	Ta	Vi
En	63 [76][62]	85 [52][54]	69 [60][56]	73 [-][-]	70 [68][84]	78 [79][75]	78 [75][76]	57 [68][53]	90 [77][64]	70 [68][80]
Fa	-	67 [69][87]	78 [72][73]	58 [-][-]	76 [67][85]	67 [75][70]	81 [76][82]	60 [67][73]	77 [70][71]	61 [67][69]
Fr	-	-	60 [56][42]	85 [-][-]	90 [56][65]	86 [55][54]	84 [61][69]	84 [64][57]	90 [60][44]	78 [58][76]
Ge	-	-	-	88 [-][-]	86 [66][77]	86 [71][65]	72 [62][84]	79 [59][49]	90 [70][59]	71 [66][69]
Hi	-	-	-	-	89 [-][-]	67 [-][-]	92 [-][-]	60 [-][-]	77 [-][-]	78 [-][-]
Ja	-	-	-	-	-	76 [66][75]	62 [50][78]	85 [63][81]	85 [59][79]	88 [69][89]
Ko	-	-	-	-	-	-	91 [74][80]	70 [76][59]	81 [62][58]	81 [56][73]
Ma	-	-	-	-	-	-	-	82 [81][71]	89 [74][69]	85 [50][79]
Sp	-	-	-	-	-	-	-	-	85 [65][48]	85 [62][61]
Ta	-	-	-	-	-	-	-	-	-	88 [71][77]

REFERENCES

[1] M. A. Zissman and K. M. Berkling, “Automatic language identification,” *Speech Communication*, vol. 35, pp. 115-124, 2001.

[2] M. A. Zissman, “Comparison of four approaches to automatic language identification of telephone speech,” *IEEE Trans. Speech and Audio Processing*, vol. 4, no. 1, pp. 31-44, Jan. 1996.

[3] F. Cummins, F. Gers, and J. Schmidhuber, “Language identification from prosody without using explicit features,” in *Proc. EUROSPEECH*, Budapest, Hungary, Sep. 1999, pp. 371-374.

[4] F. Ramus and J. Mehler, “Language identification with suprasegmental cues: A study based on speech resynthesis,” *J. Acoust. Soc. Am.*, vol. 105, no. 1, pp. 512-521, Jan. 1999.

[5] F. Ramus, M. Nespor, and J. Mehler, “Correlates of linguistic rhythm in speech signal,” Cognition, vol. 73, no. 3, pp. 265-292, 1999.

[6] A. Cutler and D. R. Ladd, *Prosody: models and measurements*, Springer-Verlag, Berlin, Heidelberg, New York, Tokyo, 1983.

[7] E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke, “Modeling prosody for speaker recognition“, *Speech Communication*, vol. 46, pp. 455-472, 2005.

[8] A. G. Adami and H. Hermansky, “Segmentation of speech for speaker and language recognition,” in *Proc. EUROSPEECH*, Geneva, Sep. 2003, pp. 841-844.

[9] A. G. Adami, R. Mihaescu, D. A. Reynolds, and J.J. Godfrey, “Modeling prosodic dynamics for speaker recognition,” in *Proc. Int. Conf. Acoust., Speech and Signal Processing*, Hong Kong, China, Apr. 2003, vol. 4, pp. 788-791.

[10] S. R. Mahadeva Prasanna, S. V. Gangashetty, and B. Yegnanarayana, “Significance of vowel onset points for speech analysis,” in *Proc. Int. Conf. Signal Processing and Communication*, Bangalore, India, Jul. 2001, vol. 1, pp. 81-86.

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

[12] 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, pp. 3009-3022, Nov. 1997.

[13] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika, “The OGI multi-language telephone speech corpus,” in *Proc. Int. Conf. Spoken Language Processing*, Banf, Alberta, Canada, Oct. 1992, vol. 2, pp. 895-898.

[14] J. L. Rouas, J. Farinas, F. Pellegrina, and R. A. Obrech, “Modeling prosody for language identification on read and spontaneous speech,” in *Proc. Int. Conf. Acoust., Speech and Signal Processing*, Hong Kong, China, May 2003, vol.1, pp. 40-43.

[15] C. Lin and H. Wang, “Language identification using pitch contour information,” in *Proc. Int. Conf. Acoust., Speech and Signal Processing*, Philadelphia, USA, Apr. 2005, vol. 1, pp. 601-604..