Proceedings of
ICISIP - 2005

Neural Network Classifiers for Language Identification using Phonotactic and Prosodic Features

Leena Mary, K. Srinivasa Rao and B. Yegnanarayana
Speech and Vision Laboratory
Department of Computer Science and Engineering
Indian Institute of Technology Madras, Chennai - 600 036 India
{leena,ksr,yegna}@cs.iitm.ernet.in

Abstract

In this paper, we explore phonotactic and prosodic features derived from the speech signal and its transcription for identification of a language. The characteristics of languages represented by phonotactic and prosodic features at the trisyllabic level are used to train Feedforward Neural Network (FFNN) classifiers to discriminate among languages. We demonstrate that these features indeed contain language-specific information. We also show that phonotactic features in terms of broad phonetic categories are sufficient to represent the phonotactic regularities/constraints of languages. The performance of the FFNN classifier based on these features is evaluated for three Indian languages.

1. INTRODUCTION

Multilingual interoperability is an important issue for many applications of speech technology. Automatic language identification (LID) is the task of identifying the language of speech by a machine. An LID system can be connected as a front end device for a multilingual speech recognizer or a language translation system, which will enable the loading of speech recognizer designed for that language [1, 2].

An understanding of the characteristics of spoken language is essential for the development of an LID system. Each language has a finite set of syllables. As the vocal apparatus used in the production of languages is universal, there is significant overlap in the syllable sets, and the total number of syllables is finite. But there can be differences in the way the same syllable is pronounced in different languages.

The frequency of occurrence of syllables in languages differ significantly. Phonotactic rules governing the way different phonemes are combined to form syllables also differ. The sequence of allowable syllables are different from language to language. Certain phoneme/syllable clusters common in certain language may be rare in some other language.

The word roots and lexicons are usually different from language to language. Each language has its own vocabulary, and its own manner of forming words. Even when two languages share a word, the set of words that may precede or follow the word may be different. At higher levels, the sentence pattern are different among languages.

The sources of information that are relevant for the task of automatic language identification are summarized as follows:

1. Acoustic-phonetics
2. Prosody
3. Statistics of subword units such as phonemes or syllables
4. Vocabulary
5. Grammatical and lexical structure

It has been observed that human beings often can identify the language of an utterance even when they have no strong linguistic knowledge of that language. In the absence of higher level knowledge of a language, a listener presumably relies on lower level constraints such as acoustic-phonetics, phonotactics and prosody. Automatic LID can make use of any of the above information or a combination of them.

In the earlier work, we have used the acoustic-phonetics represented by the spectral feature vectors for language identification [3, 4]. In this work, we focus on the phonotactic and prosodic aspects of languages. This paper is organized as follows: In Section 2, the LID is formulated in a probabilistic framework. In Sections 3 and 4 we describe briefly the phonotactic and prosodic features used for language identification task respectively, and discuss how they are approximated to a trisyllabic level. In Section 5, the results of language identification experiments using phonotactic and prosodic features on three Indian languages is discussed. Section 6 gives conclusions of this study and issues to be addressed.

2. PROBABILISTIC FORMULATION OF LANGUAGE IDENTIFICATION PROBLEM

Let $S = \{s_1, s_2, s_3, ..., s_n\}$ represent a string of phonemes/syllables corresponding to any of the languages in the set $L = \{L_1, L_2, L_3, ..., L_M\}$. The task is to find out the most likely language L^* of the input speech consisting

of n phonemes/syllables. The problem can be expressed as:

$$L^* = \arg\max_i P(L_i|S) \quad (1)$$

where $Pr(L_i|S)$ is the posterior probability of language L_i. Let us assume that the input vector S belongs to one of M classes $L_i, 1 \leq i \leq M$. The main objective in pattern classification is to decide that to which class the given vector S belongs. According to Baye's rule, the problem reduces to maximizing the joint density $P(S, L_i) = P(S|L_i)P(L_i)$. Literature abounds with methods to estimate the likelihoods $P(S|L_i)$. According to the rule given in (1), the objective is to choose the class L^* for which the posterior probability $P(L_i|S)$ is maximum for a given S. This can also be implemented using

$$L^* = \arg\max_i P(S|L_i)P(L_i) \quad (2)$$

where $P(S|L_i)$ represents the likelihood probability of S corresponding to language L_i and $P(L_i)$ denotes the a priori language probability which can be assumed to be uniform for all languages, hence ignored. Therefore the problem is simplified to

$$L^* = \arg\max_i P(S|L_i) \quad (3)$$

Thus the LID task becomes the estimation of the posterior probability as per (1) or the likelihood probability as per (3).

3. PHONOTACTIC FEATURES FOR LANGUAGE IDENTIFICATION

Each language has its own set of syllables and they possess certain language dependent characteristics. Though languages have most of the syllables in common, they will differ in

1. Frequency of occurrence of certain syllables
2. Possible co-occurrence of syllables
3. Syllables which are unique to a language
4. Pronunciation variations in the case of the same syllable

Among the language-specific features, the phonotactic constraints are shown to be the most powerful feature for LID [5, 6]. LID researchers make use this property by building separate statistical language models for each of the language at phoneme level to represent these constraints.

3.1 Neural network classifiers based on phonotactic features

The syllable based properties of the languages can be exploited for language identification using neural network classifiers. The FFNN based classifier outputs estimates the posterior probability of different languages, when features derived from the test is applied at its input. The FFNN model shown in Figure 1 uses sequential constraints at the subword level (how phonemes are combined to form syllables and syllables are combined to form words) as the property for classifying languages. For syllable coding, each syllable is assumed to have four constituents, and each constituent is given a unique code. This will give rise to a collection of four codes to represent each syllable. Each word boundary is represented by the absence of syllable. The codes corresponding to each syllable along with preceding and following syllable can be normalized between 0 to +1, and can be used as input data for training the neural network based classifier along with language identity as output as shown in Figure 1. The output corresponding to the language of the training speech is set to +1 and other outputs to 0. The network can be trained using backpropagation algorithm. While testing, the trisyllabic code derived from the test data can be used as the input to the classifier, and the corresponding output score can be accumulated for continuous stream of n syllables. The identity of the language can be decided based on the highest accumulated score at the output.

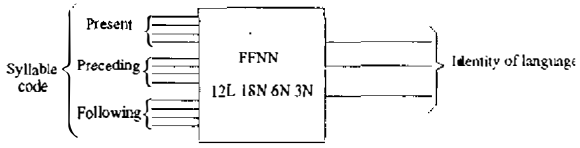

Figure 1. Training of neural network classifier for language identification using phonotactic features.

Consider a test sequence $S = \{t_1, t_2, ...t_n\}$ containing n trisyllabic units, where t_j $1 \leq j \leq n$ represents each trisyllabic unit. Then

$$P(L_i|S) = \prod_{j=}^{n} P(L_i|t_j) \quad (4)$$

This is equivalent to accumulating the log-likelihood probability

$$log P(L_i|S) = \sum_{j=}^{n} log P(L_i|t_j)) \quad (5)$$

The identity of the language can be decided based on the highest accumulated log probability at the output. This is equivalent to finding the language that maximizes the log probability.

$$L^* = \arg\max_i log P(L_i|S) \quad (6)$$

4. PROSODIC FEATURES FOR LANGUAGE IDENTIFICATION

Prosodic features (suprasegmental) are those aspects of speech which go beyond phonemes/syllables, and deal with auditory features of sound. In spoken communication we can use these features without really thinking about them. Prosodic features are shown to be more robust to acoustic and environmental mismatches than cepstral features [7].

4.1 Rhythm

Among the prosodic features, rhythm is known to carry a substantial information about language identity. Rhythm is produced by the regularity of patterns of some language specific unit. The rhythm and CV patterns are closely related to the articulation during speech production [8].

The syllable structure is important in rhythm modeling. The number of constituents in a syllable N_t, number of constituents before and after vowel.N_{c1}, N_{c2}, respectively, can be used for representing the syllable structure. Since it is difficult to compute the durations of syllable constituents (consonants and vowels) separately, the duration of the syllable is used instead, which may help to represent the rhythmic characteristics of the syllable. A syllable in isolation can not be associated with representative rhythm. The rhythm is formed by a sequence of syllables, which are closely related to certain linguistic properties. Hence the rhythm can be approximated at a trisyllabic level by combining the features deri ed from three nearby syllables.

4.2 Intonation

Pitch or fundamental frequency (F_0) variation over a larger span (larger than a syllable or as large as a phrase/sentence) is normally referred as intonation. Pitch analysis generates pitch curves with finer variations. But intonation is often perceived as the tendencies and inflections of the contour, and not by the fine variations.

The language-specific pitch characteristics may be represented by the following quantitative measures.

(a) The range of pitch within the syllable ΔF_0
(b) The location of maximum F_0 with respect to the onset of syllable, which is important in determining the accent
(c) The average pitch F_{syl} as the representative pitch of the syllable
(d) The positional details of the syllable with respect to the word and phrase

In order to account for the local variations of pitch, the above measures of the preceding and following syllable may be used to form a trisyllable level representation.

4.3 Neural network classifier using phonotactic and prosodic features

Human beings acquire the prosodic knowledge over a period of time. The process by which this happens can not be explained or formulated in terms of algorithms. So it is difficult to represent them in a machine learnable form. Hence rule-based approaches are impractical for prosody. Neural networks can be used for capturing the implicit language-specific prosodic features for language identification. The role of prosody and phonotactic features in syllable sequence for identification of language can be studied with the help of segmented and transcribed database using a FFNN as shown in Figure 2.

Let the test speech is represented by sequence of

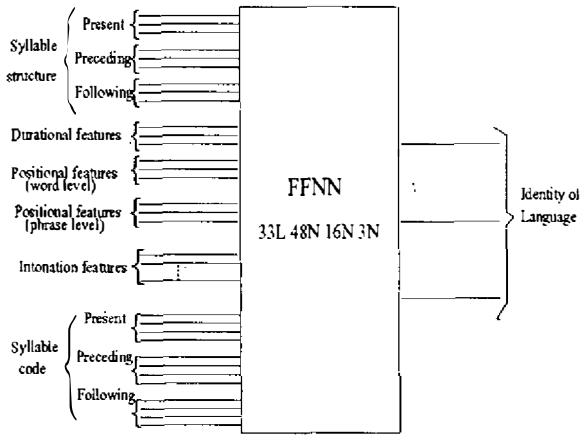

Figure 2. Neural network classifier based on phonotactic and prosodic features for language identification.

feature vectors $X = \{x_1, x_2, ...x_n\}$ where $x_j, 1 \leq j \leq n$ represents the phonotactic and prosodic features of the j^{th} trisyllabic unit. Then

$$logP(L_i|X) = \sum_{j=1}^{n} logP(L_i|x_j) \qquad (7)$$

The identity of the language can be decided based on the highest accumulated log-likelihood probability at the output terminal of the FFNN classifier.

5. RESULTS FROM EXPERIMENTAL STUDY

The database consists of speech segments excised from continuous speech in broadcast television news bulletins for Indian languages. It contains data of three different languages namely, Hindi, Tamil, and Telugu. The database contains several segmented and transcribed short segments

Table 1. Performance of phonotactic feature based neural network classifier LID system using explicit syllable description to capture different phonotactic features. The different columns 2 to 5 represent the percentage of languages identified correctly.

Method	20 syllables		50 syllables	
Language	Rank -based	Accumulation	Rank -based	Accumulation
Tamil	99	98.5	100	100
Telugu	72.4	85.8	83.8	95.6
Hindi	96.2	98.5	100	100

Table 2. Performance of phonotactic feature based neural network classifier LID system where broad labeling is used to capture different phonotactic features. The different columns 2 to 5 represent the percentage of languages identified correctly.

Method	20 syllables		50 syllables	
Language	Rank -based	Accumulation	Rank -based	Accumulation
Tamil	99.8	99.25	100	100
Telugu	47.65	82.7	49.86	92.6
Hindi	29.6	88.3	81.71	96.34

of speech, each approximately of 3 sec duration. This is further parsed into syllables by human experts using Indian Language Transliteration (ITRANS) code. The choice of this database was mainly to study the phonotactics of these languages without any errors at the segmentation and transcription level. Further, since the word boundaries are available, frequently occurring words containing less than four syllables can be modeled using the trisyllabic structure. For each language, approximately 40,000 syllables were used for training the classifier. While testing, an average of 600 test cases, each having length of 20 syllables and 250 test cases, each having length of 50 syllables were used.

5.1 Using explicit syllable codes

The input features were obtained from the transcribed database by explicitly coding the syllables assuming that each syllable has four constituents. The absence of any constituent was represented by a unique code. A trisyllabic model was used for modeling the phonotactic regularities. The performance of the neural network based classifier using phonotactic features is given in Table 1. In rank-based method, languages were ranked based on the classifier output value for each syllable in the test speech. The number of first ranks obtained are counted for length of the test syllable sequence. The language having maximum first ranks is decided as the winner. In the second method, accumulation of evidence is done as per (5). It is seen that the LID system gives better performance when evidences are accumulated.

5.2 Using broad category syllable codes

The input features were obtained from the transcribed database by broadly segmenting the syllables and then they are coded. Syllable constituents were labeled in terms of broad phonetic categories namely vowels, nasals, semivowels, fricatives, unvoiced unaspirated stop, unvoiced aspirated stop, voiced unaspirated stop and voiced aspirated stop. The input features were obtained by coding the syllables in terms of this broad classification. The stop consonants were classified based on the manner of articulation since it is known that aspirated sounds do not exist in Tamil. A trisyllabic model was used for modeling the phonotactic regularities. The performance of the NN based classifier for LID using the broad phonotactic features is given in Table 2. The evidence obtained from the broad phonotactic features suggests that labeling of syllables in terms of broad categories is sufficient to represent the phonotacic constraints of languages. This will eliminate the need for proper transcribed speech for training as well as for testing of the neural network. The broad phoneme categories used for labeling the syllable constituents should be optimized for languages in the identification task.

5.3 Using phonotactic and prosodic features

Rhythmic features of languages are represented by the structure of syllable and its duration. Intonation parameters are represented by the average syllabic pitch, range of pitch and location of maximum pitch. Since the features corresponding to a single syllable alone is not sufficient for representing the prosodic pattern, a trisyllabic structure is taken as the basic unit. The trisyllabic structure is obtained by including features from the preceding and following syllables along with the present syllable features.

For each language, features derived from approximately 25,000 syllables were used for training the classifier. Prosodic features along with phonotactic features in terms of explicit syllable codes is used to train the FFNN based classifier with language identity as output. The results shown in Table 3 reveals that by including the prosodic features along with the phonotactic features, the classifier shows an improvement in performance even when the training was done with fewer examples.

Table 3. Performance of phonotactic and prosody based neural network classifier LID system. The entries from columns 2 to 5 represent the percentage of languages identified correctly.

Method	20 syllables		50 syllables	
Language	Rank -based	Accumulation	Rank -based	Accumulation
Tamil	95.15	96.46	99.53	100
Telugu	67	82	90	100
Hindi	99.40	100	100	100

6. SUMMARY AND CONCLUSIONS

In this paper we have shown that the neural network based classifiers are capable of performing language identification using phonotactic and prosodic features. The FFNN classifiers are able to distinguish between languages from the phonotactic regularities/constraints of languages, and language-specific prosody represented in terms of quantitative measures. Also the phonotactic features in terms of broad phonetic categories were tried in order to capture the phonotactic regularities of languages. This study can be extended for more languages. At the next level, our goal is to design a system which can identify the language using features which can be directly derived from the speech signal, without the use of syllable transcription.

REFERENCES

[1] M. A. Zissman and K. M. Berkling, "Automatic language identification," Speech communication, vol. 35, pp. 115–124, 2001.

[2] Alex Waibel, "Multilinguality in speech and spoken systems," Proc. IEEE, vol. 88, no. 8, pp. 1297–1313, Aug. 2000.

[3] Leena Mary, and B. Yegnanarayana, "Autoassociative Neural Network Models for Language Identification," in Proc. Int. Conf. Intelligent Sensing and Information Processing, Jan. 2004, pp. 317–320.

[4] Leena Mary, K. Srirama Murty, S. R. Mahadeva Prasanna, and B. Yegnanarayana, "Features for Speaker and Language Identification," in Proc. of Odyssey-2004, pp. 156–159., Jun. 2004

[5] M. A. Zissman, "Comparison of four approaches to automatic langage identification of telephone speech," inIEEE Trans. Acoust., Speech, and Signal Processing, vol. 4,no. 1, pp. 31-44, Jan. 1996.

[6] Yeshwant K. Muthusamy, Etienne Barnard, and Ronald A. Cole, "Reviewing automatic language identification," inIEEE signal Processing Magazine, vol. 11 no. 4 pp. 33-41 Oct. 1994

[7] Ann E. Thyme-Gobbel, and Sandra E. Hutchins, "On using Prosodic Cues in Automatic Language Identification," in Proc. Int. Conf. Spoken Language Processing, vol. 3, pp. 1768–1772, Apr. 1996.

[8] Jean Luc Rouas, Jerome Farinas, Francois Pellegrina, and Regine Andre obrech, "Modeling Prosody for Language Identification on Read and Spontaneous Speech," in Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing, vol. 1, pp. 40–43, May. 2003.