Identification of Different Sound Types from Speech using Prosodic Features

Srijita Jana

Asst. Prof., Dept. of Computer Science & Engineering

Dream Institute of Technology,Kolkata, India

Abstract— Identification of three different sound types is an important issue for speech processing. This paper presents a method of sound type identification based on the prosodic features of Indian languages. It includes the study of three types of sounds (i.e., Voice, Noise & Silence), their nature of periodicity, number of sound zones in a sentence and their duration etc. The primary goal of this study is to analyze the speech data and to develop a method to identify the different sound types, sound zones of that speech irrespective of language and their distribution pattern on different language. Sentences spoken by its different native speakers (both male and female) are analyzed and based on the detailed analysis, a sound type model has been developed. The model is evaluated using the raw test data set/sentences which were not included in the data set used for analysis i.e., using open test condition.

Keywords—Correlation coefficient; Power; Extrema Rate; Euclidean distance.

I. INTRODUCTION

Any sound identification system has three major components, that are database preparation, feature extraction and classification. As with sound type or speech recognition, humans are the most accurate identification systems in the world today. People have an inbuilt ability to recognize the sound type irrespective of language within seconds of hearing the speech. Each Languages have characteristic sound patterns like singsong, rhythmic, guttural, nasal etc. The frequency of occurrence of the phonological units that are used to produce words and the order of their occurrence in the words, is different from language to language. In several recent studies, articulatory parameters, spectral information, prosody, phonotactic and lexical knowledge etc are explored as various levels of speech features. Prosody of spoken language plays an important role in identification process of speech [8, 9]. It has been found that fluently spoken speech is not produced in a smooth, unvarying stream. Rather, it has some breaks or silences and both the voiced and unvoiced zones.

Now-a-days, speech synthesis is considered to be of primary concern for the need to empower (both the functionally illiterates & disabled) the population, using the man machine communication. Several researches covering this arena took place to develop various models to identify a language based upon different features . Direct speech mode of interaction with the machine may enable them to achieve mediator free direct access environment. Application of the knowledge of intonation and rhythm patterns of a language is extremely important for the speech synthesis with the aid of machine. We hypothesize that an LID system which exploits certain linguistic category in turn will have the necessary discriminative power to provide good performance on short utterances. The algorithms for LID (automatic Language Identification) can be roughly divided into two groups namely phonotactic modeling and acoustic modeling.

II. PREVIOUS WORK

Several approaches were taken and major studies were done to speech processing through years. Texas Instrument (1974-1980) effort was based on frequency occurrence of certain reference sounds in different languages [3]. The best published results were 80% on five languages. These studies embody notion of phonetic-base distinctiveness of languages. House and Neuberg's(1977) work was based on manually phonetic transcribed data[4]. They didn't use acoustic features, rather they exploit phonotactic information to achieve perfect discrimination between eight languages. Li and Edwars (1980) applied techniques suggested by House and Neuberg to real speech data [10]. They used broad phonetic classes to compute two statistical models: one based on segments and other based on syllables, with five languages. Foil (1986) examined two types of language identification system [11]. First approach extracted seven prosodic features (based on rhythm and intonation) from pitch and energy contour and second used formant frequencies, in terms of values and locations to represent the characteristic sounds patterns of language. Goodman(1989) extended Foil's work by modifying and adding parameters to improve the classification distance metric [8]. Muthusamy (1993) [6] discussed that broad phonetic, prosodic and acoustic information is required for segmental approach of LID. He experimented on four languages which explores segment ratios, duration and frequency of occurrence. He concluded with the opinion that phonetic level information may distinguish between languages with better accuracy than broad phonetic information. Through the study of the roles of phonotactic, prosodic and acoustic information Yan

(1965) provided a partial unification [12]. The correct recognition rate was 91%. Large vocabulary continuous speech recognition system (LVCSR) was used by Schultz et al (1996) where comparison of language identification was performed based on word level and phone level both with and without language model (LM). As per the dissertation, amount of knowledge incorporation in the word-based language identification system has a proportional relationship with the performance of that system [13]. Phone recognizer followed by language models (PRLM) was used by Berkling (1999) for evaluation and the studied features were phone duration, phoneme frequency of occurrence etc [5]. According to Hombert (1999) and Maddieson, the segments that are rare in different languages and easily could be identified have very importance in an LID system [1]. MIT group (2003) evaluated specific approach, phone recognition in phone-based LID system where new phoneme sets were used [7].

III. THE SPEECH MATERIAL

The present study aims at developing identification model for different sound types using different Indian languages spoken by its native speakers. The training set, covering different topics such as science, literature, general news, novels etc, are taken from different source. The speech data is recorded in a speech studio environment with 16 bit mono 22050 Hz (sampling frequency) digitization format.

IV. FACTORS AFFECTING THE SOUND TYPES

There are three types of sound i.e., voice, noise and silence present in every language. So, their presence is language independent though the distribution pattern is different in different languages. There are different factors that affect the continuous speech vigorously.

A. Speaking rate

Detection of spoken language also depends on the speaking rate[5,7] as it determines the duration of different sound zone specially the silence zone. In this study, only normal speaking rate is considered.

B. Emotion

Recognition of emotion in speech has many potentialities. It acts as an aid in speech understanding. Traditionally emotion was treated as ‘noise’ which detracts from understanding of an utterance. In different cultures and languages, emotions are often portrayed differently. The identification of different sound types (voice, noise and silence) also vary with the change of emotional condition of the speaker.

C. Stress/ Power

Stress can be considered as a factor to identify different languages as primary stress falls on different syllables of a word for different languages.

V. EXPERIMENTAL DETAILS

The aim and scope of current study is limited to the identification of different sound types and their distribution pattern based on some Indian languages those are being spoken by its native users [9] i.e., the natural speaker’s speech for that particular language will only be considered for the processing [2]. This study presents an approach to recognize the sound types from speech. The goal is to analyze the speech data and to develop a model to identify the different sound types and their distribution pattern in different languages.

In case of voiced speech, when the input excitation is nearly periodic impulse sequence, the corresponding speech looks visually almost periodic. In case of unvoiced speech, as the excitation is random noise-like, so the resulting speech is also without any periodic nature. Silence region can be described as the separator in the speech production process that generates unvoiced and voiced speech in succession.

D. Detection Methodologies

The language independent presence nature of three sound types (voice, noise and silence) is playing the main and crucial role in this detection methodology. Three different parameters were considered to identify different sound types in a sentence that are correlation coefficient, power and extrema rate. Due to the periodic nature, voice signals posses higher correlation value than noise signals while non-periodic nature of noise signal results in higher extrema rate value. Power can make clear distinction among voice, noise and silence.

In this study, we use the following formulae to get the correlation value-

$$r_{xy} = 1/(N-1)\sum_{i=0}^{N}((\mathrm{x_i}-\bar{\mathrm{x}})/\mathrm{S_x})((\mathrm{Y_i}-\bar{\mathrm{Y}})/\mathrm{S_Y}) \tag{1}$$

$$\mathrm{X_P} = \frac{\sqrt{\sum_{i=0}^{N}|\mathrm{x_i}|^2}}{N} \tag{2}$$

In equation (1), $\bar{\mathrm{X}}$ & $\bar{\mathrm{Y}}$ are mean of the data set x & y and S_X & S_Y are standard deviation of that data set and X_P is representing the Power of the sentence in (2).

The value of the extrema rate of a particular window is calculated as the average of the total value of local maxima and local minima of that window. During training phase, two types of windowing technique were considered for experiment. In case of overlapping window concept, each window has last 50% of the previous window and first 50% of the next window while, no part of a sound file falls in more than one window in case of non-overlapping window. In this study, Euclidean distance method is used for the classification purpose. Here, table I is representing window-wise correct recognition rate of different sound types for non-overlapping Window while, table II is used to represent the recognition rate for overlapping window. Table III is used to represent the scenario in case of file wise detection.

Table 1

Input sound type	**Output sound type**			***Overall % of correct recognition***
	Voice(%)	***Noise(%)***	***Silence(%)***	
Voice	96.73	3.27	0	90.5
Noise	8.74	91.26	0	
Silence	0	16.56	83.44	

Table 2

Input sound type	**Output sound type**			***Overall % of correct recognition***
	Voice(%)	***Noise(%)***	***Silence(%)***	
Voice	97.07	2.92	0	90.43
Noise	9.72	90.23	0.0004	
Silence	0	16.16	83.84	

Table 3

Input sound type	**Output sound type**			***Overall % of correct recognition***
	Voice(%)	***Noise(%)***	***Silence(%)***	
Voice	97.82	2.18	0	95.51
Noise	4.04	95.96	0	
Silence	0	7.25	92.75	

At the later stage of experiment, we imply some refinement strategy independently to improve the correct detection rate of different sounds (voice, noise, silence) for non-overlapping window. From the above tables, it is clear that some originally voice sound identified as noise, some silence identified as noise and some noise identified as voice. For much accurate detection, some combinations of parameters were implied on the wrongly identified sound. The combinations are i. Normal calculation without any refinement ii. Refinement using power iii. Refinement using power & correlation coefficient iv. Refinement using power & extrema rate v. Refinement using extrema rate & correlation coefficient vi. Refinement using power, extrema rate & correlation coefficient vii. Refinement using correlation coefficient viii. Refinement using extrema rate. Though all the different refinement strategies could not result equally for three sound types. If the input sound type is voice, the next level refinement combination, using correlation coefficient works very good where power improves correct recognition rate for noise type input. At the same time, when a certain strategy improves the detection rate of a specific sound type, the same one may decrease yields poor performance for the other two types of sound.

At the next stage of the study, i.e., detection of the different distribution pattern of the sound types in different languages, 3 parameters were chosen based on the number of different sound zones present in a file, total duration of a specific zone, total duration of the sentence etc.

The parameters are as follows:

A) Total no of voice zone/total duration of that particular file

B) Total no of noise zone/total duration of that particular file

C) Total no of silence zone/total duration of that particular file

We give alias of the above 3 parameters as P1, P2 & P3 Next we get values of these 3 parameters from the training set for three languages separately.

During testing phase, Equation (3) measures the value of Euclidean distance for a language.

$$\boldsymbol{ED_B} = \sqrt{\frac{\left(\boldsymbol{M^B_{P1}}-\boldsymbol{X_{P1}}\right)^2}{\boldsymbol{SDEV^B_{P1}}^2}+\frac{\left(\boldsymbol{M^B_{P2}}-\boldsymbol{X_{P2}}\right)^2}{\boldsymbol{SDEV^B_{P2}}^2}+\frac{\left(\boldsymbol{M^B_{P3}}-\boldsymbol{X_{P3}}\right)^2}{\boldsymbol{SDEV^B_{P3}}^2}} \qquad (3)$$

In figure 1, three different color lines in the graph represent total number of specific zone / total duration of that segment (i.e., parameter P1, P2 & P3). The X axis represents the number of input speech and the Y axis represents the value range for language A. Figure 2 represents the same for language B.

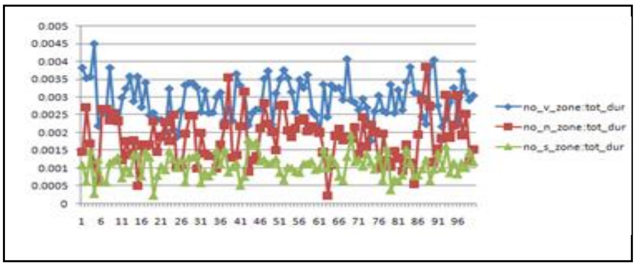

Fig. 1: plotting of parameter P1, P2 & P3 for language A

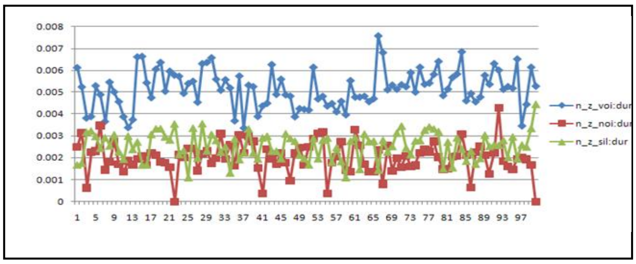

Fig. 2: plotting of parameter P1, P2 & P3 for language B

B. Result of the Experiment

The experiment explores that the differentiation between voice and silence sound types produces better result but differentiation between silence and noise sound types often yields poor performance. The distribution pattern of different sound types in the speech also varies from language to language. Here, language A and language B depicts totally different distribution pattern. In case of language B, the silence and noise sound types are overlapping while in case of language A , the distribution pattern is not overlapping. The overall correct recognition rate following this model is more than 90%.

It is revealed during the experimental process that the testing data speech below 3 sec duration yields poor and non-acceptable results.

VI. FUTURE SCOPE

As the distribution pattern of different sound types differs from language to language, so this nature could be used to develop a model to recognize the language being spoken, in next phase of experiment. A new aspect to automatic language identification from speech could be revealed through this model.

REFERENCES

[1] Hombert, J.M., Maddieson, I. The Use of ’Rare’ Segments for Language Identification. Proc. Eurospeech’99, vol. 1, pp. 379-382, September 1999.

[2] Jean-Luc Rouas, "Automatic Prosodic Variations Modeling for Language and Dialect Discrimination," Audio, Speech, and Language Processing, IEEE Transactions on, vol. 15, pp. 1904-1911, Aug. 2007.

[3] Leonard, R.G., Doddington, G.R. Automatic Language Identification. Technical Report RADC-TR-74-200, Air Force Rome Air Development Center, August 1974 .

[4] House, A.S., Neuberg, E.P. Toward Automatic Identification of the Languages of an Utterance: Priliminary Methodological Considerations. Journal of the Acoustical Society of America, 62(3), pp. 708-713, 1977.

[5] Berkling, K., Reynolds, D., Zissman, M.A. Evaluation of Confidence Measures for Language Identification. Proc. Eurospeech’99, vol. 1, pp. 363-366, September 1999

[6] Muthusamy, Y.K. A Segmental Approach to Automatic Language Identification. PhD thesis, Oregon Graduate Institue of Science and Technology, October 1993.

[7] Singer, E., Torres-Carrasquillo, P.A., Gleason, T.P., Campbell, W.M., Reynolds, D.A. Acoustic,Phonetic,and Discriminative Approaches to Automatic Language Identification. Proc. Eurospeech’03, pp. 1345-1348, September 2003

[8] Goodman, F.J., Martin, A.F., Wohlford, R.E. Improved Automatic Language Identifi- cation in Noisy Speech. Proc. ICASSP’89, pp. 528-531, May 1989

[9] Yuan-Fu Liao, Shuan-Chen Yeh, Ming-Feng Tsai, Wei-Hsiung Ting, and Sen-Chia Chang,” Latent Prosody Model-Assisted Mandarin Accent Identification”

[10] Li, K.P., Edwards, T.J. Statistical Models for Automatic Language Identification. Proc. ICASSP’80, pp 884-887, April 1980.

[11] Foil, J.T. Language Identification Using Noisy Speech, Proc. ICASSP’86, pp. 861-864, April 1986.

[12] Yan, Y. Development of an Approach to Language Identification Based on Languagedependent Phone Recognition. PhD thesis, Oregon Graduate Institue of Science and Technology, October 1995.

[13] Schultz, T., Rogina, I., Waibel, A. LVCSR-Based Language Identification. Proc. ICASSP’96, pp. 781-784, May 1996.