ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Prosody Based Foreign Accent Identification

M. Sireesha
M.Tech,
Electronics and Communication,
VNRVJIET, Hyderabad.

G. Radha Krishna, M.Tech, Ph.D pursuing
Associate Professor,
Electronics and Communication,
VNRVJIET, Hyderabad.

Abstract:

Prosodic features of a language are not only critical in human perceptual process but also useful in Automatic Language Identification tasks like Accent Identification. Methods of accent classification which give high accuracy emulate human processes which attempt to capture phonetic variations. These are based on phonetic transcription and are impractical in real- world scenarios. Present approach is based on using acoustic data labeled only by accent, without phonetic labeling and to use machine learning techniques to model the phonological variations implicitly is attempted. Prosodic features were used to train Gaussian Mixture Models (GMMs) to distinguish languages. In this work speech corpus is collected from the native speakers of three south Indian states with mother tongue Telugu (TEL), Tamil(TAM) and Kannada(KAN). Results showed accuracies better than 82%.

Key words: Accent identification, GMM, SDC, MFCC, prosodic features.

I.INTRODUCTION :

Speaker characteristics vary either due to differences in physiological characteristics of speech production organs or learned habits. The acquired or learned habits are mostly influenced by the social environment and also by the characteristics of native language of the speaker in the critical period of learning. A speaker might use a blend of languages or vary their accent according to circumstances. Human listeners distinguish nativity by paying special attention to some particular phonetic patterns. Automatic speaker recognition approach of foreign accent identification can be utilized to study the uniqueness of the speech characteristics of a speaker. The approach to non-native speech recognition is to assume that non-native speakers will dominantly use their native phones, presumably by mapping the phones of the language they are speaking (L2) to their native language (LI).

Uniqueness of speaker may be attributed to several factors like shape and size of vocal tract; dynamics of the articulator; rate of vibration of vocal folds; Accent imposed by the speaker and speaking rate. Recognition systems in speech technology allow identifying the language, dialect, or speaker in a particular utterance based on a set of known previous recordings. Foreign accent deviates from native speech mainly in prosodic features namely Fluency, Word stress, Intonation. Challenging task is extraction and representing prosodic features for language or speaker recognition. Prosody is linked to syllables of a language which change in terms of Melody, Timing and Stress. These differences are represented using parameters derived from pitch contour, energy, and duration.

For parameterization of the segments, prosodic feature statistics and their local temporal slopes (tilt) within each segment are often used. Each voiced segment was parameterized by a piece-wise linear model whose parameters formed the features. In, the authors used N-gram counts of discretized feature values as features to a GMM classifier with promising results. In prosodic features were extracted using polynomial basis functions.

Prosody is known to play an important role in human speech perception process. Therefore, there is an increasing need to use prosodic features for the advancement of speech recognition technology. However, prosody is related to various levels of information, from linguistic, Para linguistic, to non-linguistic, and, therefore, its acoustic manifestation is rather complicated with large variations. This fact prevents prosody to be incorporated in speech recognition process. The organization of the thesis is as follows: Section 1 is introduction. In Section 2, Corpus collection is described. In section 3 Identifying Accents and Languages are described. The Modeling techniques employed in our experiments are given in Section 4. Results and discussion are contained in Section 5. Finally, Conclusion and scope for future work is given in Section 6.

II. Corpus Collection :

Speech samples from native speakers of three south Indian states namely A.P(Andhra Pradesh),Tamilnadu and Karnataka with mother tongue Telugu(TEL),Tamil(TAM) and Kannada(KAN) respectively constitute the training set. The training corpus details are given in table1. The speakers are separated into two groups, training and testing set. Different subsets of speakers in each language who are capable of speaking English Language are chosen as the testing set. The testing database used is the English utterance of the speakers with one of the above three languages as mother tongue.

Gender weightage was equally distributed in both training and testing sets. Entirely different sets of samples are used Training data for each language is obtained by concatenating speech data of different male and female speakers to make the model truly speaker-independent. For each language, speech data of duration of 1200 seconds is used for training. Ten test utterances in each language are used for evaluating the performance.Table1. Training conditions for three south Indian languages.

Language	Kannada (KAN)	Tamil (TAM)	Telugu (TEL)
Speakers	5	3	5
Training duration	30mt	30mt	30mt
Gender	3F, 2M	2F, 2M	2F, 3M

Table2. Testing conditions for three south Indian languages:

Language	Kannada (KAN)	Tamil (TAM)	Telugu (TEL)
Speakers	5	4	5
Testing duration	30mt	30mt	30mt
Gender	2F, 3M	3F, 2M	2F, 3M

III.IDENTIFYING LANGUAGES & ACCENTS:

Training the prosody of the foreign language is often neglected fully. This is also the case for today’s computer-assisted language learning systems: although some systems provide automatic pronunciation scoring for words, prosodic phenomena are mostly ignored. It is therefore desired to have an automatic method for evaluating the quality of the speaker’s prosody.

Methods of accent classification that emulate human processes attempt to capture some of the differences like using different phonemes and realizing phonemes differently across languages with the help of phoneme transaction gives high accuracy. This approach is rarely available in real- world scenarios. Present approach is to use acoustic data labeled only by accent, without phonetic labeling and to use machine learning techniques to model the above variations implicitly is attempted.

.

IV.FOREIGN ACCENT CLASSIFICATION SYSTEM:

A.System Building:

Incorporating prosodic information proceeds by first segmenting an utterance into syllables based on amplitude and pitch information. Various statistics based on the rhythmic and tonal characteristics are then computed. Information related to rhythm is encapsulated in terms of syllable timing and duration, and descriptors of amplitude patterns. Tone information is described in terms of phrase characteristics of the pitch (such as the range of pitch levels and the variation of pitch over a whole utterance) and syllable characteristics thereof (including descriptors of the shape of the pitch trajectory in each syllable, and measures of how strongly the pitch falls off at the end of each syllable).

Finally, the correlation between pitch and amplitude is also described by several measures Accent identification is related to speaker-independent speech recognition and speaker identification. It is practically easy to train phoneme models than training models of entire language. Though they are found to outperform those based on stochastic models, the phonemic approach has the following drawback. It needs phonemically labeled data in each of the target languages for use during the training.

The difference among languages, apart from their prosody lies in their short-term acoustic characteristics. Indian languages share many phones among themselves. Since there are many variants of the same phoneme, we need to consider the acoustic similarities of these phones. Combination of phonetic and acoustic similarities can decide a particular mother tongue. For text-independent language recognition, it is generally not feasible to construct word models in each of the target languages. So, models based on the sequential statistics of fundamental units in each of the languages are employed. Text independent recognizers use Gaussian mixture models (GMMs) to model the language dependent information. The modeling technique deciding the acoustic vectors should be multimodal, to represent the pronunciation variations of the similar phonemes in various languages. The language model used in this particular study is a GMM model of Mel Frequency Cepstral Coefficients MFCCs.

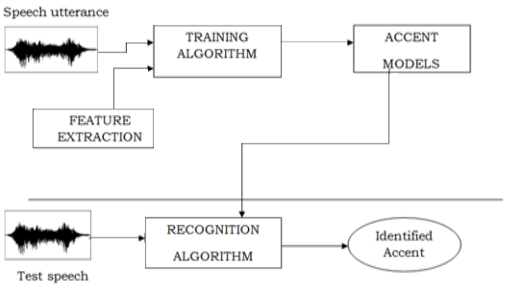


Fig 1 Accent Identification system

B.Spectral features for accent identification:

Speaker recognition systems rely on low-level acoustic information []. The classical front end analysis in speech recognition is a spectral analysis which parameterizes the speech signal into feature vectors. Mel-Frequency Cepstral Coefficients are most popular acoustic features used in speech recognition. Studies indicate that a cohesive representation of the acoustic signal is possible by using a set of Mel-frequency cepstral coefficients (MFCCs) which emulates the functioning of human perception. The speech signal is converted into a set of cepstral coefficients represented by a 13 dimensional MFCC feature vector. After collecting the multilingual speech data set, acoustic model parameters were estimated from the training data in each language. The extraction and selection of the parametric representation of acoustic signals is critical in developing any speaker recognition system.

Cepstral features capture the underlying acoustic characteristics of the signal. They characterize not only the vocal tract of a Speaker but also the prevailing characteristics of the vocal tract system of a phoneme. MFCCs provide information about the phonetic content of the language. Hence, we used MFCC coefficients as feature vectors to model the phonetic information.

C. Stochastic Models:

Gaussian Mixture Models (GMMs) belongs to the stochastic modeling and based on the modeling of statistical variations of the features. Therefore, it provides a statistical representation of how speaker produces sounds. The main purpose of using the Gaussian mixture models (GMM) in pattern recognition stage is because of its computational efficiency. It is a dominant approach for text-independent applications. GMM technique lies a hybrid between a parametric and non-parametric density model. Similar to parametric model it has structure and parameters that control the behavior of the density in known ways, but without constraints that the data must be of a specific distribution type. Like a nonparametric model, GMM has the freedom to allow arbitrary density modeling. In the present investigation, the Gaussian components can be considered to be modeling the broad phonetic sounds that characterize a person's voice.

V.RESULTS AND DISCUSSION:

Automatic speaker recognition systems can be implemented using widely used low level information such as sound of a person's voice via short term features related to speech spectrum. Beyond low level acoustics conveying speaker information, high level information related to learned habits and style such as particular word usage (idiolect) can be used successful. Both high level and low level features are studied and results are given below.

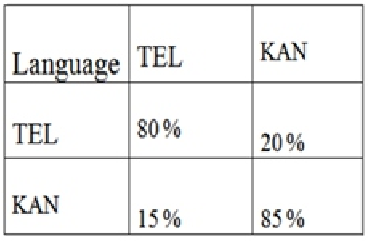
(a)Acoustic System :

A base line system consisted of standard approaches. The acoustic system is a standard cepstral GMM system using short term cepstral based features with 8 Gaussian mixtures built using a home grown corpus. This system produces accuracy up to 85%.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

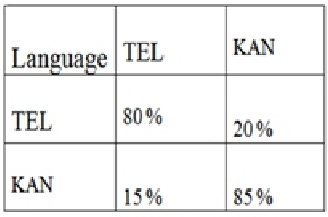
Table 3: Confusion matrix of pair-wise Accent identification task:

(a): Between Telugu and Kannada:



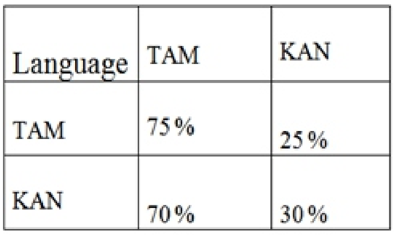
Language	TEL	KAN
TEL	80%	20%
KAN	15%	85%

(b): Between Telugu and Tamil



Language	TEL	KAN
TEL	80%	20%
KAN	15%	85%

(c): Between Tamil and Kannada



Language	TAM	KAN
TAM	75%	25%
KAN	70%	30%

(b). Prosodic Features:

For improvement in basic recognition accuracy by adding complimentary knowledge and also for robustness to acoustic degradation, to which low level features are highly susceptible high level information sources are explained. To begin with various measurements from data base like duration and pitch related features were extracted. The results are explained below.

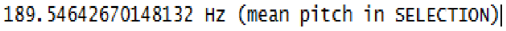
Pitch Frequencies for IdenticalTwins :

One way of focusing on voiced sound is by estimating the pitch (fundamental frequency) for each speech frame. The pitch is estimated with auto correlation method using the standard 'PRAAT' toolkit. Dynamics of fo contour corresponding to a sound unit is influenced by sveral factors such as identity of the sound unit, its position with respect to the phrase/word, its context, speaking style of the speaker, intonation rules of the language, type of the sentence etc.

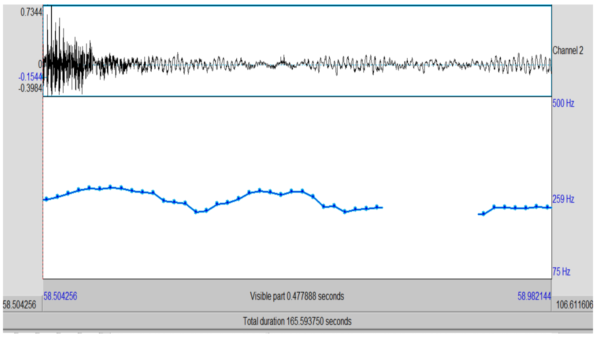
The dynamics of fo contour can be different among speakers due to different speaking style and accent.

Typical Pitch Frequencies for /ing/ sound of 3 speakers with their mother tongue

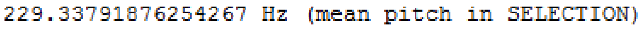
Telugu



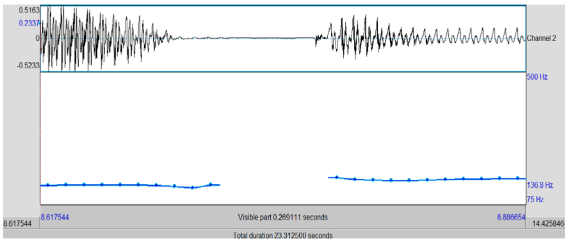
```
189.54642670148132 Hz (mean pitch in SELECTION)
```



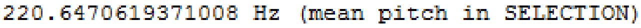
Kannada



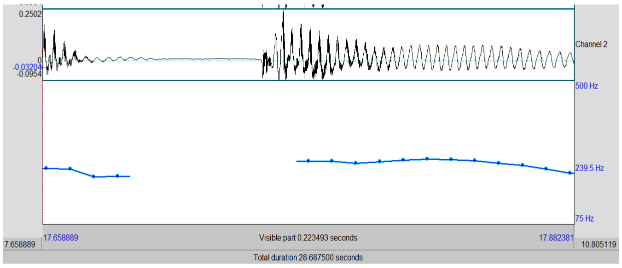
```
229.33791876254267 Hz (mean pitch in SELECTION)
```



Tamil

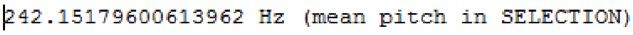


```
220.6470619371008 Hz (mean pitch in SELECTION)
```

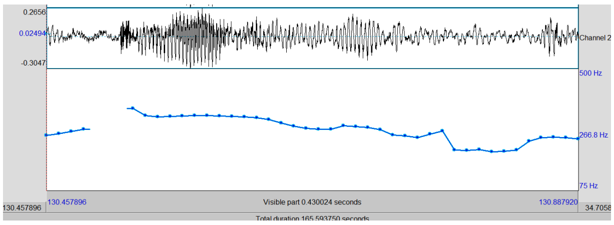


Typical Pitch Frequencies for /The/ sound of 3 speakers with their mother tongue

Telugu

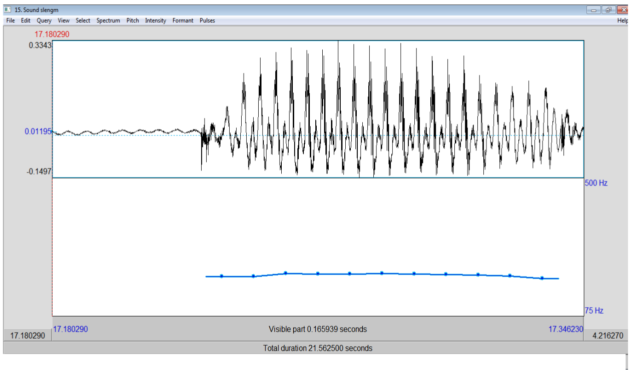


```
242.1517960613962 Hz (mean pitch in SELECTION)
```



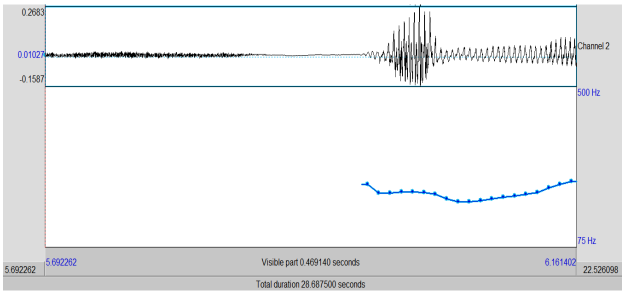
Kannada

```
240.88381615327523 Hz (mean pitch in SELECTION)
```


15. Sound slengm
File Edit Query View Select Spectrum Pitch Intensity Formant Pulses
Help
17.180290
0.3343
0.01195
-0.1497
500 Hz
75 Hz
17.180290
17.180290
Visible part 0.165939 seconds
17.346230
4.216270
Total duration 21.562500 seconds

Tamil

```
233.6968346955853 Hz (mean pitch in SELECTION)
```



VI. CONCLUSION AND FUTURE SCOPE :

Implement an Automatic Speaker Recognition system for identification of Accent and thus the native state of the speaker. Confusion is observed between Kannada and Tamil speakers. This confusion is found to be less when Acoustic prosodic features were introduced. We have proposed an effective approach to identify Accent in multilingual scenario by following the techniques available in Language and Speaker Identification.

A general purpose solution is proposed with a multilingual acoustic model. Further improvements can be made by including prosodic features and also covering techniques such as inclusion of SDC features and also the i-vector paradigm. Most important advances in future systems will be in the study of acoustic-phonetics, speech perception, linguistics, and psychoacoustics.

Next generation systems need to have a way of representing, storing, and retrieving various knowledge resources required for natural conversation particularly for countries like India. With the same training and testing procedures, apart from English and other regional languages, national language Hindi can be modeled and influence of any particular language on it can also be studied.

VII ACKNOWLEDGEMENT:

I am very thankful to The authors would like to acknowledge the cooperation of the management and staff of VNRVJIET.

VIII REFERENCES:

1.G. Doddington, P. Dalsgaard, B. Lindberg, H. Benner, and Z. Tan, "Speaker recognition based on idiolectal differences between speakers", in Proc. EUROSPEECH, pp. 2521–2524, Aalborg, Denmark, Sep. 2001.

2. "Multi Level Implicit features for Language and Speaker Recognition", Ph.D. Thesis, Leena Mary, Department of Computer Science, Indian Institute of Technology Madras, India ,June 2006.

3.A. Maier et.al. "Combined Acoustic and Pronunciation Modeling for Non-Native Speech Recognition" Interspeech 2007, pp1449-1452.

4. D. A. Reynolds," Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models", IEEE Transactions on Speech and Audio Processing Vol.3, No.1, 1995, 72-83.

5.A. Maier et.al "A Language Independent Feature Set for the Automatic Evaluation of Prosody" Interspeech 2009.

6.J. Cheng, N. Bojja, X. Chen "Automatic Accent Quantification of Indian Speakers of English" Interspeech 2011, pp2574-2578.

7.D.A. Reynolds, T.F. Quatieri and R.B.Dunn, "Speaker Verification using adapted Gaussian mixture models", Digital Signal Processing vol 10, pp19-41, 2000.

8. N. Scheffer, L. Ferrer, Martin Graciarena, S. Kajarekar, E. S. Stolcke, "The SRI NIST 2010 Speaker Recognition Evaluation System".