International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-2, December 2019

An Experimental Analysis of Speech Features for Tone Speech Recognition

Utpal Bhattacharjee, Jyoti Mannala

Abstract*: Recently Automatic Speech Recognition (ASR) has been successfully integrated in many commercial applications. These applications are performing significantly well in relatively controlled acoustical environments. However, the performance of an Automatic Speech Recognition system developed for non-tonal languages degrades considerably when tested for tonal languages. One of the main reason for this performance degradation is the non-consideration of tone related information in the feature set of the ASR systems developed for non-tonal languages. In this paper we have investigated the performance of commonly used feature for tonal speech recognition. A model has been proposed for extracting features for tonal speech recognition. A statistical analysis has been done to evaluate the performance of proposed feature set with reference to the Apatani language of Arunachal Pradesh of North-East India, which is a tonal language of Tibeto-Burman group of languages.

Keywords: Feature Selection, LPCC, MFCC, Tonal Language, Prosodic Features, Speech Recognition

I. INTRODUCTION

Automatic speech recognition (ASR) research has made remarkable progress since its inception in the mid of 20th century making it a viable option for human-machine interaction. However, there are few issues which are still hindering its wide spread use in commercial applications. One such issue is the language dependency of the speech recognition systems. Based on the use of tone for discriminating phones, the languages may be divided into two broad categories - Tone language and Non-tone language. A language is regarded as `Tone Language' if the change in the tone of the word results in changing the meaning of the word [1]. The basis of tone is the pitch of the sound. Pitch is the perceived fundamental frequency or the rate of vibration of the vocal folds during the production of the sound. The most general definition of tone language was proposed by D.M. Beach in the year 1924 [2]. Beach defined tone language as a language that uses pitch constructively in any manner of its articulation. According to this definition all the languages are tone language since intonation in terms of pitch modulation is inherent to the articulation of any language.

Revised Manuscript Received on December 30, 2019.
***** Correspondence Author
Utpal Bhattacharjee*, Department of Computer Science and Engineering, Rajiv Gandhi University, Rono Hills, Doimukh, Arunachal Pradesh, India. Email: utpal.bhattacharjee@rgu.ac.in
Jyoti Mannala, Department of Computer Science and Engineering, Rajiv Gandhi University, Rono Hills, Doimukh, Arunachal Pradesh, India. Email: mannalajoy@gmail.com

However, this definition fails to distinguish the languages where tone is used to distinguish words of different meaning otherwise phonetically alike. Tone or intonation is the musical modulation of the voice in speech and as such integral part of the speech production in any language [3]. According to C.M.Doke [4] tones may be classified into two broad categories - characteristics tone and significance tone. Characteristic tone is the method of grouping of musical pitch which characterize a particular language, language group or language family. Significant tone on the other hand plays an active part in the grammatical significance of the language, may be a means of distinguishing words of different meaning otherwise phonetically alike. A generally accepted definition of tone language was proposed by K.Pink [5]. According to this definition, a tone language must have lexical constructive tone. In generative phonology, it means tone of a tonal phonemes are no way predictable, must have to specify in the lexicon of each morpheme [3]. For any tone language, the basic building block is tonal syllable. A tonal syllable consist of two components - a syllabic sound unit and an associated lexical tone. If the tone is ignored, it is called base syllable. Each syllable consist of vowel and consonant sounds. Tone is realized in voiced segment, therefore, tonal base units (TBU) in most of the time are voiced vowels [6]. Since tone associated with the vowels are sufficient to express the tone associated with the syllable, in the present study, only tonal vowels will be analysed to determine the tonal phoneme discrimination capability of the feature sets. Tone may be broadly classified into two categories -- Level tone and Contour tone. Level tones are the tones which remain constant throughout the TBU. Level tones are classified as High, Low and Middle. In construct, contour tones shows a clear shifting from one level to another within the syllabic boundary. Contour tones may be classified into rising and falling. Woo [3] argued that contour tones can be considered as collection of multiple level tones. Her argument was supported by other scholar like Leben [7], Goldsmith [8] and Yip [1] with suitable evidence to justify the fact. However, many other scholars did not support that contour tone should be decomposed into level tones [6].

A major section of world population spreading across south-east Asia, East Asia and Sub-Sahara Africa are speakers of tonal language [9]. In the present study, an attempt has been made to analyse the tonal phoneme discrimination capability of popular feature extraction techniques namely Mel frequency cepstral coefficient (MFCC), Linear predictor cepstral coefficient (LPCC) and prosodic features.

Selection of suitable feature set is one of the most crucial design decision for the development of a speech based system.

Speech signal not only conveys the linguistic information, but lots of other information like information about the speaker, gender, social and regional identity, health and emotional status etc. Different speech features represent different aspects of the speech signal. Moreover, the information present in different speech features are redundant and overlapping.

Therefore, it is difficult to identify and separate which aspect of the speech signal is represented by which feature. In speech research, very often features are selected on experimental basis, and sometimes using the mathematical approach like Principal component analysis (PCA).

The Apatani language of Arunachal Pradesh of North East India is belongs to the Tani group of language. Tani languages constitute a distinct subgroup within Tibeto-Burman group of languages [10]. The other languages of the group are Adi, Bangni, Bokar, Bori, Damu, Gaol, Hill Miri, Milang, Na, Nyishi, Tagin, Tangam and yano. The Tani languages are found basically in the continuous areas from the Kamng river to the Siang river of Arunachal Pradesh. A small number of Tani speakers are found in the contiguous area of Tibet and only the speakers of Missing language are found in the Brahmaputra valley of Assam [11]. The Apatani language has 06(six) vowels and 17 (seventeen) consonants [12].

The Table. 1 presents the Apatani vowels and Table. 2 presents the Apatani consonants with their manner and position of articulation.

Table1: Apatani vowels.

Tongue Height	Tongue position		
	Front	central	Back
High	ɪ		ʊ
Mid	ɛ	ə	ɔ
Low	ɑ:		

Table 2: Apatani consonants with their manner and place of articulation

Manner of Articulation	Place of Articulation				
	Labial	Alveolar	Palatal	Velar	Glottal
Stop	p, b	t, d	ʧ, ʤ	k, g	
Nasals	m	n		ŋ	
Fricative		s		k^h	h
Flap			r		
Approximate		l	ɹ		

II. THE SPEECH FEATURES

Speech is the output of a vocal tract system excited by an excitation source signal. Characteristics of both the vocal tract response and excitation source signal vary with time to produce different sounds. At the time of speech production, human beings impose duration and intonational pattern on top of the vocal tract response to convey the intended message [13]. Speech signal not only conveys the linguistic information but lots of other information like information about the speaker, gender, social and regional identity, health and emotional status etc. The first step of automatic speech recognition system is to form a compact representation of the speech signal emphasizing phonetic information of the signal over other information. Choosing suitable features for developing a speech based system is one of the most crucial design decision for speech based system development. The speech features can be categorize into three categories -- Excitation source features, vocal tract features and prosodic features.

Speech features extracted from excitation source signal is called source features. Excitation source signal is obtained by discarding the vocal tract information from the speech signal. This is achieved by first predicting the vocal tract information using linear predictor filter coefficients extracted from the speech signal and then separating it by using inverse transformation. The resulting signal is called linear predictor residual signal [14]. The features extracted from LP residual signal is called excitation source features or source features. A sound unit is characterized by a sequence of shapes assumed by the vocal tract during production of the sound. The vocal tract system can be considered as a cascade of cavities of varying cross sectional areas. During speech production, the vocal tract act as a resonator and emphasizes certain frequency components depending on the shape of the oral cavity. The information about the sequence of shapes of vocal tract that produce the sound unit is captured by vocal tract features also called system or spectral features. The vocal tract characteristics can be approximately modelled by spectral features like linear predictor coefficients (LPC) and ceptral coefficients (CC) [13]. Prosody plays a key role in the perception of human speech. The information contained in prosodic features is partly different from the information contained in source and spectral features. Therefore, more and more researchers from the speech recognition area are showing interests in prosodic features. Generally, prosody means "the structure that organizes sound". Pitch (tone), Energy (loudness) and normalized duration (rhythm) are the main components of prosody for a speaker. Prosody can vary from speaker to speaker and relies on long-term information of speech.

Very often, prosodic features are extracted with larger frame size than acoustical features as prosodic features exist over a long speech segment such as syllables. The pitch and energy contours change slowly compared to the spectrum, which implies that the variation can be captured over a long speech segment [15].

The source, system and prosodic features are distinct from each other in speech production, feature extraction and perception point of view. They are mostly non-overlapping in nature and represent different aspects of the speech production system. The basic objective of ASR system is to recognize the phonetic content of the speech signal discarding other irrelevant information.

Most of the state-of-the-art ASR systems are developed using only system or spectral features as these features are concern with the shape of the vocal tract during production of different sound units, which in turn reveals the information about the sound unit produced. However, in case of tonal speech recognition, speech unit having the same phonetic structure but of different tones convey different meaning.

Therefore, the system feature itself is not sufficient for the recognition of the tonal speech. To enhance the performance of tonal speech recognition system, the prosodic information, which represents the tonal characteristics of the speech must have to be incorporated in the feature set. The major challenge in incorporating prosodic features with the spectral features comes from the extraction process itself.

The spectral features are short-term features. The change pattern of the spectral features can be recorded with high resolution if the observation window size is 15~25 microseconds.

However, due to the slow-varying nature of the prosodic features, in this observation window the changes in the prosodic features of the speech signal cannot be captured. To overcome the problem, fusion of the features extracted from this two domains has been carried out. In speech processing, two commonly used methods of fusion are - feature-level fusion and score-level fusion. In feature-level fusion, prosodic features like pitch and temporal energy were computed frame by frame and they are appended to the spectral features. To capture the dynamic property of the features, their first-order and second-order derivatives are also added. However, vital information which can be observed only in long-duration observation window are missed out in this approach. In the second approach, the spectral and prosodic features are extracted from the tonal base unit (TBU) using separate observation window. The spectral features are then feed to a classifier that computes a class label for the base acoustical unit of the TBU and the prosodic features are feed to a classifier that computes a class label for the tone associate with the TBU. One of the major problem with this approach is that correlation between the spectral and prosodic features are completely ignored at classifier level.

In this paper we have proposed a hybrid method where the features are extracted with different observation windows and then combined together to take a decision on class boundary of the TBU.

III. PROPOSED METHOD

The block diagram of the proposed model is given in Fig. 1. The pre-emphasized speech signal is first blocked into frame of 100 ms duration with 50% overlapping. From each block, two types of features have been extracted -- spectral features and prosodic features. The spectral features considered in the present study are Mel Frequency Cepstral Coefficients (MFCC) and Linear Predictor Cepstral Coefficients (LPCC). To extract the spectral features, each speech frame of 100 ms has been re-framed into frame of size 20 ms with 50% overlapping. The spectral features namely MFCC and LPCC have been extracted from each 20 ms frame separately. In the present study we have proposed a modified k-mean clustering algorithm which preserve the temporal information of the speech feature. We are calling it temporal k-mean (TKM) algorithm. The algorithm is given below:

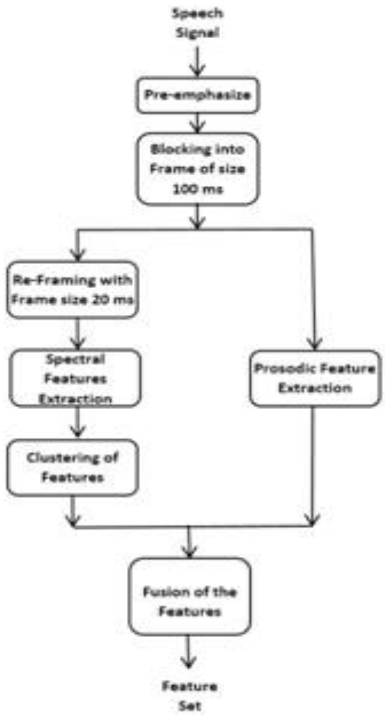

Fig. 1. Block diagram of the hybrid feature extraction system

Temporal K-Mean (TKM) Algorithm

1. Compute the initial value for the ith cluster centroid as follows:

$$c_{ij} = \frac{1}{M} \sum_{1+(i-1)*M}^{i*M} c_j \qquad \dots (1)$$

where $M = \frac{N}{k}$, N and k are the total number of frames and number of clusters respectively, c_j is the value of the jth coefficient of the feature and c_{ij} is the initial value of the ith cluster for jth coefficient

2. Use a data structure for the centroid as (centroid_values, proximity_index), the proximity_index referred to the central location of each cluster derived in the time scale.
3. For each frame j repeat step 4 to 6
4. Select the two nearby clusters m and k for jth frame based on proximity index. The cluster with two consecutive proximity index m and k are nearby clusters to j if

$$M * m \leq j \leq k * M \qquad \dots (2)$$

5. Compute the distance of the jth frame from this two cluster centroids.
6. Assign the frame to the nearby cluster and update its cluster centroid.

The algorithm has been applied separately to both MFCC and LPCC features and reduced feature sets have been extracted which represents the spectral characteristic of the speech signal for the entire 100 ms duration. These features are combined with prosodic features extracted from the 100 ms frame considering it as a single unit.

The prosodic features extracted are maximum, minimum and average values of F0 and Energy computed over the entire 100 ms period. These prosodic features are combined with MFCC and LPCC features separately and two different sets of features have been computed. Each feature set is evaluated for their relative performance in tonal speech recognition.

IV. EXPERIMENTAL SETUP

In the present study, each tonal instance of a vowel has been considered as different tonal vowel. For example, the vowel [ɑ:] has three associated tones -- rising, falling and level. Thus vowel [ɑ:] gives raise to the tonal vowels [ɑ́:] ([ɑ:] rising), [ɑ̀:] ([ɑ:] falling) and [ɑ̄:] ([ɑ:] level). We referred to these vowels as tonal vowels. Considering the tonal instances as a separate vowel, we get sixteen tonal vowels in Apatani language. The vowels are given in Table. 3. Since the vowel [ə] has only one tone, it is not taken into consideration while evaluating the performance of the feature vectors.

A speech database of Apatani tonal words has been prepared to carry out the experiments. The database consist of 12 isolated tonal words spoken by 20 different speakers (13 males and 7 females). The recording has been done in a controlled acoustical environment at 16 KHz sampling frequency and 16 bit mono format. A headphone microphone has been used for recoding the database. The words are selected in such a way that each tonal instance of the vowel has at least 5 instances among the words. Thus, for each tonal vowel, we have minimum 100 instance recorded from 20 speakers.

Table. 3. Apatani Tonal vowels.

[ɑ̄:]	Vowel ɑ: with level tone
[ɑ́:]	Vowel ɑ: with rising tone
[ɑ̀:]	Vowel ɑ: with falling tone
[ɪ̄]	Vowel ɪ with level tone
[ɪ́]	Vowel ɪ with rising tone
[ɪ̀]	Vowel ɪ with falling tone
[ɔ̄]	Vowel ɔ with level tone
[ɔ́]	Vowel ɔ with rising tone
[ɔ̀]	Vowel ɔ with falling tone
[ɛ̄]	Vowel ɛ with level tone
[ɛ́]	Vowel ɛ with rising tone
[ɛ̀]	Vowel ɛ with falling tone
[ʊ̄]	Vowel ʊ with level tone
[ʊ́]	Vowel ʊ with rising tone
[ʊ̀]	Vowel ʊ with falling tone
[ə̄]	Vowel ə with level tone

A feature would be effective in discriminating between different tonal vowels if the distribution of different tonal vowels are concentrated at widely different location in the parameter space although they are different from each other only in associated tone[16]. A good measure of effectiveness would be the ratio of inter-vowel to intra-vowel (within the class) variance for the tonal vowels, referred to as F-ratio, which is defined as

$$\mathrm{F} = \frac{\text{Variance within the class}}{\text{Average variance across all classes}} \quad \ldots (3)$$

To compute the overall F-ratio values across all class. The equation is:

$$F = \frac{\frac{1}{N}\sum_{i=1}^{N}(\mu_i - \bar{\mu})}{\frac{1}{N}\sum_{i=1}^{N} S_i} \quad \ldots (4)$$

Where N is the number of tonal vowels, μ_i is the mean of a particular coefficient of the feature vector for ith tonal vowel, $\bar{\mu}$ is the overall mean value for that coefficient of the feature vector for all the tonal vowels. S_i, within a tonal vowel variance is given by

$$S_i = \frac{1}{M_i}\sum_{j=1}^{M_i}(x_{ij} - \mu_i) \quad \ldots (5)$$

where x_{ij} is the value of the coefficient for jth observation of the ith tonal vowel and M_i is the number of observations for ith tonal vowel. Higher F-ratio value for a coefficient indicates that it can be used for good classification

Another metric used for measuring the performance of features in discriminating among the tonal instances of a vowel is the Kullback-Leibler distance (KLD). The KLD provides a natural distance between a probability distribution and a target probability distribution. KL distances have been measure among features extracted from the tonal vowel and their average has been taken. If the distance is higher, the feature has better tonal phoneme discrimination capability.

V. RESULTS AND DISCUSSIONS

All the experiments were carried out using the database described in Section - IV. The vowels are segmented from the isolated words for all its tonal instances. The segmentation has been done using PRAAT software which is followed by subjective verification. The speech signal is first segmented into frame of 100 ms with 50% overlapping. We will refer to this as 1st level frame. Each 1st level frame is now passed through two parallel system. The 1st system extracts the spectral features –MFCC and LPCC separately. To extract the spectral features, whose characteristics are correctly visible only in short duration frame, we have re-framed the 1st level frame into frame of size 20 ms with 50% overlapping. We refer to this as 2nd level frame. The MFCC and LPCC features are extracted from each 2nd level frame. The MFCC feature has been computed using a 21-channel filter bank resulting in a 13-dimensional cepstral features consisting of c_0 to c_{12} coefficients. The LPCC has been computed using a 10th dimensional predictor signal aggregated to a 13-dimenaional cepstral coefficients. Now, the MFCC and LPCC features are clustered into 3 clusters using temporal k-mean algorithm. The cluster centroids are clubbed together and we get a 39-dimentional MFCC and 39-dimensional LPCC feature vector for the 1st level frame of the speech signal. These two set of features are then combined with the prosodic features separately.

The prosodic features – maximum, minimum and average F0 and Energy are computed from each 1st level frame directly. Thus, we get two sets of 45-dimensional feature vectors (39 spectral features and 6 prosodic features) for each 1st level frame. We will refer to this features as High-level MFCC and High-level LPCC features respectively.

To perform a comparative study of the proposed feature set, we have extracted baseline MFCC and LPCC features from the speech signal with 20 ms frame size and 50% overlapping considering the same experimental setup as described above. To capture the dynamic property of the speech signal, the 1st order and second order derivatives of the coefficients are also added. Thus we get a 39-dimensional MFCC feature vector and 39-dimensional LPCC feature vector. The result of the experiment carried out is given in the Table. 4.

Table. 4. Average F-ratio and KL Distance for the features.

Feature vector	F-ratio	KL Distance
Baseline MFCC + Δ + ΔΔ	2.0136	0.4404
Baseline LPCC + Δ + ΔΔ	2.5569	0.6956
High-Level MFCC	5.3350	0.8727
High-Level LPCC	4.3350	0.8754

From the above experiments it have been observed that as a result of adding prosodic features along with the MFCC and LPCC features, the overall tonal phoneme discrimination capability increases considerably compared to baseline MFCC and LPCC features.

In the second set of experiments, we have computed the intra-tone phoneme discrimination capability of the proposed feature set. We have computed the F-ratio value considering all the phonemes of a particular tone (level, rising or falling) intra-class. Similarly, KL-distance has been measures only with other vowels of the same tone. The result is summarized in Table. 5.

Table. 5. Average F-ratio and KL Distance for the features for intra-tone phoneme discrimination capability

Feature vector	F-ratio	KL Distance
Baseline MFCC + Δ + ΔΔ	3.0731	0.4721
Baseline LPCC + Δ + ΔΔ	3.7763	0.3846
High-Level MFCC	4.2870	0.4258
High-Level LPCC	4.4580	0.3516

From the above results it has been observed that the proposed features have better intra-tone phone discrimination capability. This observation justify the fact that these features can be used for both tonal and non-tonal speech recognizer.

In the third set of experiments, we have evaluated the performance of features for their inter-tone discrimination capability. In this experiment, we have computed F-ratio value considering all the instances of a tonal vowel as intra-class and other tonal instances of the same vowel as inter-class. Further, KL-distances have been measures among the tonal instances of the same base vowel only. The results of the experiments are given in Table. 6.

Feature vector	F-ratio	KL Distance
Baseline MFCC + Δ + ΔΔ	0.7365	0.0538
Baseline LPCC + Δ + ΔΔ	0.8383	0.293
High-Level MFCC	4.7813	0.5754
High-Level LPCC	3.9852	0.2958

From the above results it has been observed that the proposed features are performing significantly well in inter-tone discrimination of the phoneme when the base phoneme is the same and different tonal instances are distinct from each other only due to change in tone. In this scenario the baseline MFCC and LPCC features are completely failed to discrimination among the phonemes.

VI. CONCLUSION

This paper presents a feature set for tonal speech recognition. The spectral and prosodic features are combined together using a late fusion technique to produce a feature set for the classifier. The proposed feature extraction technique has been evaluated for tonal phoneme discrimination task. It has been observed that the proposed feature set is performing significantly well in tonal as well as tone-independent evaluation scenario. Therefore, the proposed feature set can be used as a universal feature vector for both tonal and non-tonal speech recognition systems which is a long standing need for global acceptability of automatic speech recognition system.

ACKNOWLEDGMENT

This work is supported by UGC major project grant MRP-MAJOR-COM-2013-40580.

REFERENCES

1. M. Yip, The Tonal Phonology of Chinese, New York: Garland Publishing, 1991.
2. D. M. Beach, “The Science of Tonetics and Its Application to Bantu Languages”, in Bantu Studies, 2nd Series, Vol. 2, PP. 75-106, 1924.
3. N. H. Woo, Prosody and Phonology, Doctoral dissertation, MIT, 1969.
4. C. M. Doke, A Comparative Study in Shona Phonetics, Johannesburg, University of Witwatersrand Press, 1931.
5. K. Pink, “Tone Languages”, Ann Arbor, University of Michigan Press, 1964.
6. P. Sarmah, “Tone Systems of Dimasa and Rabha: A Phonetic and Phonological Study”, Doctoral dissertation, University of Florida, 2009.
7. W. Leben, “Suprasegmental Phonology”. Ph.D. dissertation, MIT, 1973.
8. J. Goldsmith, “An overview of autosegmental phonology”, Linguistic Analysis, 2(1): 23-68, 1976.
9. U. Bhattacharjee, “Recognition of the Tonal Words of Bodo Language”, In International Journal of Recent Technology and Engineering, Volume-1, Issue-6, 2013.
10. M.W. Post and T. Kanno, “Apatani Phonology and Lexicon, with a Special Focus on Tone”, Himalayan Linguistics, Vol. 12(1):17-75, 2013.
11. J. T. Sun, “Tani languages”, In The Sino-Tibetan Languages, edited by G. Thurgood and R. LaPolla, pp. 456-466, London and New York: Routledge, 2003.
12. P. T. Abraham, Apatani-English-Hindi Dictionary, Central Institute of Indian Language, Mysore, India, 1987.
13. K. S. Rao, “Application of prosody models for developing speech systems in Indian languages”, International Journal of Speech Technology, 14(1), 19-33, 2011.
14. J. Makhoul, “Linear prediction: A tutorial review”, Proceedings of the IEEE, 63(4), 561-580, 1975.
15. E. E. Shriberg, “Higher Level Features in Speaker Recognition”, In C. Muller (Ed.) Speaker Classification I. Volume 4343 of Lecture Notes in Computer Science / Artificial Intelligence, Springer: Heidelberg / Berlin / New York, pp. 241-259, 2007.

16. G. S. Raja and S. Dandapat, “Sinusoidal model based speaker identification”, Proc. NCC-2004, vol. 1, pp. 523–527, 2004.

AUTHORS PROFILE

Utpal Bhattacharjee received his Master Degree from Dibrugarh University, India and Ph.D. from Gauhati University, India in the year 1999 and 2008 respectively. Currently he is working as a Professor in the department of Computer Science and Engineering of Rajiv Gandhi University, Arunachal Pradesh, India. His research interest is in the field of Speech and Natural language Processing and Machine Learning

Jyoti Mannala received her Master Degree from Rajiv Gandhi University, Arunachal Pradesh, India in the year 2012. Presently she is working as a research scholar in the department of Computer Science and Engineering of the University. Her research area is natural language processing.