**PROJECT COMPLETION REPORT ON**

# *Design and Development of a Speech Recognizer in the context of tonal languages of Arunachal Pradesh*

**Submitted to**
**UNIVERSITY GRANTS COMMISSION**
**BAHADUR SHAH ZAFAR MARG**
**NEW DELHI – 110 002**

**Submitted by**
**UTPAL BHATTACHARJEE**

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**
**RAJIV GANDHI UNIVERSITY, RONO HILLS,**
**DOIMUKH-791 112, ARUNACHAL PRADESH, INDIA**

**1.** TITLE OF THE PROJECT

***Design and Development of a Speech Recognizer in the context of Tonal Languages of Arunachal Pradesh. (MRP-MAJOR-COMP-2013-40580)***

**2. NAME AND ADDRESS OF THE PRINICIPAL INVESTIGATOR**


a) Name : Dr. Utpal Bhattacharjee

b) Designation : Professor

c) Department : Computer Science and Engineering

d) Address : Department of Computer Science and Engineering, Rajiv Gandhi University, Rono Hills, Doimukh-791112, Arunachal Pradesh, India.
Phone. +91 9435086480/+91 9774900310 (M)
Email: *utpal.bhattacharjee@rgu.ac.in*


**3. NAME AND ADDRESS OF THE INSTITUTE**

Rajiv Gandhi University, Rono Hills, Doimukh-791112, Arunachal Pradesh, India.

***4.*** **UGC APPROVAL LETTER NO AND DATE** : ***F.No. – 43 -281/2014 (SR) dated 30th October, 2015***

**5. DATE OF IMPLEMENTATION** : ***1st May, 2015***

**6. TENURE OF THE PROJECT** : ***01 May 2015 to 30 June 2018***

**7. TOTAL GRANT ALLOCATED** : ***Rs. 14,20,000/-***

***8.*** **TOTAL GRANT REALLOCATED** : ***Rs. 11,91,000/-***

***9.*** **TOTAL GRANT RECEIVED** : ***Rs. 9,95,000/-***

**10. TOTAL EXPENDITRUE** : ***Rs***.

**11. TITLE OF THE PROJECT**

Design and Development of a Speech Recognizer in the context of Tonal Languages of Arunachal Pradesh.

**12. OBEJECTIVES OF THE PROJECT**

a. Develop a speech recognition database for the tonal languages of Arunachal Pradesh.
b. Characterize the acoustic-phonetic parameters of speech signal to identify their intra-phoneme and inter-phoneme discriminating capability with reference to the tonal languages.
c. Identification of features that can be used as feature vector for a universal speech recognizer that can recognize both tonal as well non-tonal speech efficiently.

d. Developing a prototype for universal speech recognition system using those feature vectors.

## 13. WHETHER OBJECTIVES WERE ACHIEVED: *Yes*

- A speech database for the tonal language of Arunachal Pradesh has been developed
- Characterization of the acoustic-phonetic features has been done and based on the analysis, a new feature set has been proposed.
- A prototype for a speech recognition system has been developed for the tonal languages of Arunachal Pradesh that can efficiently recognize tonal and non-tonal words.

## 14. ACHIEVEMENT OF THE PROJECT

The languages of Arunachal Pradesh in North East India are low-resource tonal languages, making them unique from other languages such as English, Hindi, and Assamese. A tonal language is one in which the lexical tone has a significant impact on the meaning of the words. The research conducted as part of the project is the first of its kind. During the project's work, a speech recognition database for tonal languages was created, and commonly used speech parameterization techniques were tested for tonal recognition performance both statistically and using a speech recognizer. In addition, a new feature set has been proposed that may be used for tonal and non-tonal languages. The suggested feature set was tested, and a prototype for a speech recognition system was created using the proposed feature set as the parameterization technique and the Hidden Markov Model as the classifier.

## 15. SUMMARY OF THE FINDINGS

A speech recognition system for the tonal languages of Arunachal Pradesh has been developed during the project work. The languages of Arunachal Pradesh are low resource languages and primarily spoken languages. There is no scientific analysis of the language done prior to this study. In this work, we have developed an automatic speech recognition system for the tonal languages of Arunachal Pradesh. The languages can be broadly categorised into two categories – tonal and non-tonal. Tone plays a vital role in distinguishing among the syllables of a tonal language, whereas in non-tonal language, tone cannot change the lexical meaning of a syllable. English, Hindi, Assamese etc., are examples of non-tonal language and Chinese, Japanese, Apatani, Nyishi etc., are examples of tonal language. Due to the active participation of the tone-related information in determining the meaning of a syllable, the tonal

speech recognition systems are different from non-tonal speech recognition. In this work, we have presented a detailed analysis of the performance of the most commonly used speech parameters for tonal speech recognition. Analysis of the features has been done using statistical evaluation metrics and Hidden Markov Model-based recognizer. The tonal phoneme recognition consists of two subtasks – base phoneme recognition and associate tone recognition. Considering the fact that some of the speech features are inherently good in discriminating among the base-phonemes and some other parameters are good in discriminating among the tones, different combinations of the speech features are evaluated for their tonal phoneme discrimination capability. A multi-window feature concatenation algorithm has been proposed and its performance is evaluated in the context of tonal speech recognition. A speech recognition database for the languages of Arunachal Pradesh has been developed for the study of the languages of Arunachal Pradesh and we named the database as Arunachali Tonal Speech Recognition Database Version -1 (ATSRD-V.1). A prototype for a speech recognition system has been developed for the tonal languages of Arunachal Pradesh that can efficiently recognize tonal and non-tonal words. Some of the major findings of the project are:

- All the features, except the prosodic feature exhibit change in entropy due to the change in tone and base-phoneme together. However, the major contributor to the change in entropy is the change in base-phoneme. Therefore, the change in tone without the change in base-phoneme remains undetected.
- The prosodic features can capture the change in tone of the Tonal Base Unit (TBU). However, it fails to identify the change in base-syllable itself.
- Combining the features from multiple sources can improve the performance of the tonal speech recognition system.
- The features are broadly classified as segmental and supra-segmental features. The segmental features can be extracted with high resolution only from short observation windows like MFCC, LPCC etc. whereas supra-segmental features like prosodic features can be captured efficiently from long observation window. Therefore, in order to combine the features when a common window size is considered, their combined feature set lose significant information.
- The time-varying property of the speech signal contributes significantly in detection of the sound unit represented by the speech signal. When features from multiple windows size combined together, the temporal information of the smaller observation windows

have to be preserved.

- The Hidden Markov Model (HMM) based automatic speech recognition system models the speaker specific information in addition to the phonetic information. Therefore, when normalization techniques are used to minimize the intra-speaker and inter-speaker variability, there speech recognition performance improves.

## 16. CONTRIBUTION TO THE SOCIETY

It is the first scientific researches on dialects/languages of the tribals of Arunachal Pradesh. This work will help in conduct comparative studies with other languages of mongoloid stock of North East India and their counterpart in China, Korea, Japan etc. Further, the system developed during the project work recognize seamlessly both tonal and non-tonal words, which is still a major technological bottleneck for presently successful commercial speech recognition system. The system can be used as a Phoneme Recognizer, Speech to Text Converter etc. This work will serve as the foundation for further research into the languages/dialects of other ethnic groups in order to develop an automated speech-based system.

**17.** WHETHER ANY Ph.D. ENROLLED/PRODUCED OUT OF THE PROJECT

| | | |
|---|---|---|
| Ph.D. Enrolled | : | Yes |
| Name | : | Jyoti Mannala |
| Title of the Theis | : | SPEECH RECOGNITION IN UNCONTROLLED AMBIENT CONDITION IN THE CONTEXT OF TONAL LANGUAGES OF ARUNACHAL PRADESH |
| Status | : | Awarded |

**18.** NO OF PUBLICATIONS OUT OF THE PROJECT

1. Bhattachajee, U. and Mannala, J.: An Experimental Analysis of Speech Features for Tone Speech Recognition, International Journal of Innovative Technology and Exploring Engineering, vol. 9(2), pp. 4355-4360 (2019). (Scopus indexed)
2. Bhattachajee, U. and Mannala, J.: Feature Level Solution to Noise Robust Speech Recognition in the context of Tonal Languages, International Journal of Engineering and Advanced Technology, Vol. 9(2), pp 3864-3870 (2019). (Scopus indexed)
3. Bhattachajee, U. and Mannala, J. and G. Yubbey, Statistical Evaluation of Spectral Features for Tonal Phoneme Discrimination Capability, IEEE International Conference on Electrical, Communication, Electronics, Instrumentation and Computing (ICECEIC), 2019.
4. Jyoti Mannala, Bomken Kamdak and Utpal Bhattacharjee, An Analysis of Phase-based Speech Features for Tonal Speech Recognition, Advances in Electrical and Computer Technologies 2020, April 2020

# An Analysis of Phase-Based Speech Features for Tonal Speech Recognition


**Jyoti Mannala, Bomken Kamdak, and Utpal Bhattacharjee**

**Abstract** Automatic speech recognition (ASR) technologies and systems have made remarkable progress in the last decade. Now-a-days ASR based systems have been successfully integrated in many commercial applications and they are giving highly satisfactory results. However, speech recognition technologies as well as the systems are still highly dependent on the language family for which it is developed and optimized. The language dependency is a major hurdle in the development of universal speech recognition system that can operate at any language conditions. The language dependencies basically come from the parameterization of the speech signal itself. Tonal languages are different category of language where the pitch information distinguishes one morpheme from the others. However, most of the feature extraction techniques for ASR are optimized for English language where tone related information is completely suppressed. In this paper we have investigated short-time phase-based Modified Group Delay (MGD) features for parameterization of the speech signal for recognition of the tonal vowels. The tonal vowels comprises of two categories of vowels—vowels without any lexical tone and vowels with lexical tone. Therefore, a feature vector which can recognize the tonal vowels can be considered as a speech parameterization technique for both tonal as well as non-tonal language recognizer.

**Keywords** Feature analysis · MGD feature · Phase-based features · Speech recognition · Tonal language

J. Mannala · B. Kamdak · U. Bhattacharjee (✉)
Rajiv Gandhi University, Arunachal Pradesh, Rono Hills, Doimukh 791112, India
e-mail: utpal.bhattacharjee@rgu.ac.in

J. Mannala
e-mail: mannalajoy@gmail.com

B. Kamdak
e-mail: bomken.kamdak@rgu.ac.in



T. Sengodan et al. (eds.), *Advances in Electrical and Computer Technologies*, Lecture Notes in Electrical Engineering 711,
https://doi.org/10.1007/978-981-15-9019-1_54

## 1 Introduction

Natural languages are broadly classified into two categories—tonal and non-tonal based on their dependency on lexical tone. In tonal language, the lexical tone plays an important role in distinguishing the syllables otherwise similar whereas in non-tonal language the lexical tone has no significant role in distinguishing the syllables. English, Hindi, Assamese are the example of non-tonal language whereas Chinese, Japanese, language of South East Asia, Sweden, Norway and Sub-Sahara Africa are tonal languages [1]. Modern speech recognition research has a half century long legacy. The technology and the systems developed speech recognition have already registered significant progress and many systems are already commercialized. However, those systems are optimized with non-tonal languages, particularly for English language. As a result, when these systems are used for tonal speech recognition their performance degrades considerably. Since the large sections of the world population are speaker of tonal language, for the global acceptability of the speech recognition technology and system, it must be efficient in recognizing in tonal as well as non-tonal language.

One of the major reasons for the system developed for non-tonal language fail to deliver consistent performance in tonal language is due to the non-consideration of the lexical tone related information. Lexical tones are produced as a result of excursion of the fundamental frequency and these informations are discarded in non-tonal speech recognition system as a measure of performance optimization and due to robustness issues as it contains very little useful information for non-tonal speech recognition system.

In the recent years many attempts have been made for developing tonal speech recognition system [2–4]. Such systems are developed considering the fact that a tonal syllable has two components—phonetic and tone. The phonetic component gives information about the base phonetic unit which is similar with non-tonal speech and a tonal unit which gives information about the tone associated with that phonetic unit. As a result, the tonal speech recognition system relies on two sets of features—Spectral features like MFCC for base phonetic unit recognition and prosodic features for associated lexical tone recognition. The scores obtained from both are combined together to arrive at a decision on underlying syllabic unit. However, the prosodic features are highly sensitive to ambient conditions. As a result, the tonal speech recognition systems based on prosodic features are highly susceptible to ambient conditions.

The speech recognition system relies on short-term spectral property of the speech signal in order to exploit the short-term stationary property of the speech signal. To extract the short-term property, Short Term Fourier Transform (STFT) is used. STFT returns the short-term magnitude and phase spectral of the speech signal. However, in most of the cases magnitude spectra is retain to extract spectral features like Mel Frequency Cepstral Coefficient (MFCC) and phase spectral is completely discarded due to the practical difficulty in phase wrapping [5, 6]. However, the recent research has established the importance of phase spectra in speech processing

applications like speech recognition, speaker recognition, emotion recognition and speech enhancement [7].

In this paper we have analyzed the tonal phoneme discrimination capability of phase-based features. The performances of phase-based features have been evaluated for tonal phoneme discrimination.

## 2 Feature Vector for the Representation of Tonal Phonemes

The Fourier transform of a discrete time speech signal $x(n)$ is given by.

$$X(\omega) = |X(\omega)|e^{j\phi(\omega)} \tag{1}$$

where $|X(\omega)|$ is the magnitude spectra and $\phi(\omega)$ is the phase spectra of the speech signal. There are number of speech processing difficulties in using the phase spectra directly in Automatic Speech Recognition (ASR). Two most critical problems are—firstly a phase spectrum is highly sensitive to the exact positioning of the short-time analysis window. It has been observed that for a small shift in analysis window, the phase spectrum changes dramatically [8]. Secondly, the phase spectrum values are only computable within the range $\pm\pi$, called principal phase spectrum. The value changes abruptly due to the wrapping effect beyond this range. However, for better representation of the phase spectra for automatic speech recognition, the spectra must be unwrapped. The major problem with this unwrapping is that any multiple of $2\pi$ is added to the phase spectra without changing the value of $X(\omega)$. Recent studies have shown that phase spectrum can be used for speech applications and gives promising results [9, 10]. Among the phase based features extraction techniques, Group Delay Function (GDF) and All-pole Group Delay Function (APGD) are widely used. In the present study we have used a modified version of GDF called Modified Group Delay (MGD) function for extracting the phase based features due to their promising performance in speech recognition [11].

The Group Delay Function is derived by taking the negative derivation of the Fourier phase spectrum $\phi(\omega)$, written as [12, 13]:

$$\begin{aligned}\tau(\omega) &= -\frac{d(\phi(\omega))}{d(\omega)} \\ &= \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{|X(\omega)|^2}\end{aligned} \tag{2}$$

the angular frequency $\omega$ is limited to $(0,2\pi)$, $Y(\omega)$ is the magnitude of the Fourier transform of the time-weighted version of the speech signal given by $y(n) = nx(n)$. The subscript R and I denotes the real and imaginary parts of the signals. The features derived from GDF often leads to an erroneous representation near the point of discontinuity. It is due to the denominator $|X(\omega)|^2$ which tends to 0 near the point of

discontinuities. Therefore, the group delay function is modified, which is given as [14]

$$\tau(\omega) = \frac{\tau_p(\omega)}{\left|\tau_p(\omega)\right|}\left|\tau_p(\omega)\right|^{\alpha} \tag{3}$$

where

$$\tau_p(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{|S(\omega)|^{2\gamma}} \tag{4}$$

where $S(\omega)$ is the cepstrally smoothed form of $|X(\omega)|$. $\alpha$ and $\gamma$ controls the range dynamics of the modified group delay function. Here,

$$P(\omega) = X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega) \tag{5}$$

is called the product spectra of the speech signal which includes both magnitude and phase information [15].

## 3 Speech Database

In the present study, we have created a speech database of Apatani Language of Arunachal Pradesh of North East India to analyze the performance of phase-based features for tonal speech recognition in mismatched environmental conditions. The Apatani language belongs to the Tani group of language. Tani languages constitute a distinct subgroup within Tibeto-Burman group of languages [16]. The Tani languages are found basically in the contiguous areas 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 Assam [17]. The Apatani language has 06(six) vowels and 17 (seventeen) consonants [18]. To record the database, 24 phonetically rich isolated tonal words have been selected. The words are 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. Since the tone associated with the vowel is sufficient to identify the tone associated with the entire syllable [3, 19], therefore, in the present study we have evaluated the phone discrimination capability and robustness issue of the phase-based features with reference to their tonal vowel discrimination capability. Each tonal instance of a vowel has been considered as different tonal vowel. For example, the vowel [a:] have three associated tones—rising, falling and level. Thus vowel [a:] gives raise to the tonal vowels [ á:] ([a:] rising), [ à:] ([a:] falling) and [ā:] (([a :] level). Considering the tonal instances as a separate vowel, we get sixteen

**Table 1** Apatani vowels and their tonal instances

| Vowel | Tonal instances | | |
|---|---|---|---|
| | Rising | Level | Falling |
| ɪ | [ ɪ́] | [$\overline{\text{ɪ}}$] | [ ɪ̀] |
| ʊ | [ ʊ́] | [$\overline{\text{ʊ}}$] | [ ʊ̀] |
| ɑː | [ ɑ́ː] | [$\overline{\text{ɑ}}$ː] | [ ɑ̀ː] |
| ɛ | [ ɛ́] | [$\overline{\text{ɛ}}$] | [ ɛ̀] |
| ɔ | [ ɔ́] | [$\overline{\text{ɔ}}$] | [ ɔ̀] |
| ə | - | [$\overline{\text{ə}}$] | - |

tonal vowels in Apatani language. The vowels and their tonal instances are given in Table 1. Since the vowel [ə] has only one tone, it is not taken into consideration while evaluating the performance of the feature vectors.

All the experiments are carried out using this database. 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.

## 4 Experiment and Results

To evaluate the performance of the features for tonal phoneme discrimination capability, both statistical methods and Hidden Markov Model based recognizer have been used.

Euclidean distances between the feature values extracted from each pair of tonal phoneme have been computed. The Euclidean distance gives an indication of the linear separation among the tonal vowels with reference to phase-based features. Higher the value of Euclidean distance indicates better discrimination capability for the feature vector.

Fisher's Discrimination ration ($F$-ratio) [20] has been used as a quantitative measure for the tonal phoneme discrimination capability of the phonemes. $F$-ratio is defined as:

$$F = \frac{\text{Variance of the tonal phoneme mean}}{\text{Average intra} - \text{phoneme variance for all phonemes}}$$

The above ratio can be computed as:

$$F = \frac{\frac{1}{P}\sum_{i\in P}\sqrt{(\mu_i - \overline{\mu})^2}}{\frac{1}{P}\sum_{i\in P}\left(\frac{1}{T}\sum_{\beta\in T}\sqrt{\left(\left|x_\beta^{(i)} - \mu_{\beta,i}\right|^2\right)}\right)} \tag{7}$$

where $\overline{\mu}$ is the average mean for all the tonal phonemes, $\mu_i$ is the average mean for the base phoneme $i$, $\mu_{\beta,i}$ is the average mean for phoneme $i$ for tone $\beta$, $x_\beta^{(i)}$ indicates an instance of the phoneme $i$ for tone $\beta$. Higher the value of F-ratio indicates that the feature is capable of discriminating among the tonal phonemes.

To evaluate the performance of the phase-based feature set in recognizing the tonal phonemes, a Left-to-Right Hidden Markov Model (LRHMM) has been used. The LRHMM is suitable for speech recognition due to its capability to model the time varying property of the speech signal. The number of HMM states is determined experimentally. In the present model, 6 (six) states have been used. Each state is represented by a single Gaussian distribution function given by [21].

$$P\left(x|\mu,\sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) \tag{8}$$

where x is the observation vector, $\mu$ is the Gaussian mean vector and $\sigma^2$ is the variance. The forward–backward algorithm has been used for training the HMM model. Clean speech signals have been used for training the models.

To extract the short-time MGD features the speech signal is first pre-emphasized with emphasizing factor 0.97 and then framed by a Hamming windows of 30 ms duration and 10 ms frame rate. The phase-based MGD features are extracted from the windowed speech signal using the method described in the Sect. 2.

In the first set of experiments we have evaluated the phoneme discrimination capability of the MGD features in the context of tonal vowel recognition. The feature values are computed from each instance of the tonal vowels. For each tonal vowel, the average value for each dimension of the feature vector has been computed ignoring the outliers. The Euclidean distances have been computed between each tonal vowel with all the other tonal vowels and their average has been taken. Table 2 gives the average Euclidean distances of each tonal vowel from all the other tonal vowels. Table 3 presents the average Euclidean distances among different categories of tonal vowels.

From the above experiments it has been observed that phase-based MGD features are suitable in discriminating the tonal vowels. They possess discrimination ability even when the base phoneme of the tonal vowels is same and distinction among them is due to underlying tone only or vice versa.

To assess the suitability of the MGD features for tonal vowel recognition, we have computed the F-ratio values for the features. Higher the value of F-ratio among different groups indicates better discrimination ability of the feature with respect to that grouping factor. In the present study we have evaluated the computed F-ratio

**Table 2** Average euclidean distances of each tonal vowel from all the other vowels

| Tonal Vowel | Average euclidean distance from the other tonal vowels | Tonal vowel | Average euclidean distance from the other tonal vowels | Tonal Vowel | Average euclidean distance from the other tonal vowels |
|---|---|---|---|---|---|
| [ ɪ́] | 0.7513 | [ ʊ̀] | 0.9267 | [ɛ̄] | 1.2091 |
| [ī] | 0.7292 | [ ɑ́ː] | 1.4317 | [ ɛ̀] | 1.1002 |
| [ ɪ̀] | 0.5993 | [ɑ̄ː] | 1.9577 | [ ɔ́] | 1.6260 |
| [ ʊ́] | 0.9437 | [ ɑ̀ː] | 2.7468 | [ɔ̄] | 1.3167 |
| [ʊ̄] | 1.0653 | [ ɛ́] | 1.1449 | [ ɔ̀] | 2.0015 |

**Table 3** Average Euclidean distance among different categories of tonal vowels

| | |
|---|---|
| Average Euclidean distance among the vowels with same base phoneme but different tone | 1.1496 |
| Average Euclidean distance among the vowels with different base phoneme but same tone | 0.9698 |
| Average distance from the vowels with different base phoneme and tone | 1.3589 |

value with grouping factors—same base-phoneme, same tone and different base phoneme and tone. The F-ratio values are listed in Table 4.

From the above experiments, it has been established that short-time phase based feature MGD has the capability to identify the tonal vowels even when they are distinct from each other only by tone or only by base-phoneme. This observation assets the fact that short-time phase based MGD feature is a better alternative than the combination of MFCC and Prosodic based features for tonal vowel recognition which have been evaluated in our earlier works [22].

In the next set of experiments, we have evaluated the performance of MGD feature for their tonal vowel recognition in terms of recognition accuracy of the HMM based recognizer. The model has been trained using clean speech database. 60% of the tonal

**Table 4** F-ratio values under different grouping factors

| | |
|---|---|
| Average Euclidean distance among the vowels with same base phoneme but different tone | 3.5463 |
| Average Euclidean distance among the vowels with different base phoneme but same tone | 3.8222 |
| Average distance from the vowels with different base phoneme and tone | 4.6514 |

**Table 5** Evaluation metric for the HMM based recognizer

| | |
|---|---|
| Correctly recognized the tonal vowel | 89.23% |
| Incorrectly recognized as a tonal vowel with same base phoneme but different tone | 6.46% |
| Incorrectly recognized as a tonal vowel with same tone but different base phoneme | 2.91% |
| Incorrectly recognized as a tonal vowel with different tone and different base phoneme | 1.40% |

instances of each vowel have been used for training and the remaining 40% for testing the system. The performance of the MGD features have been evaluated in terms of recognition accuracy, which is the percentage of times the recognizer has been able to recognize the tonal vowel correctly. The error cases have been further in-depth investigated to get an insight into the confusion created at modeling level. Table 5 presents an analysis of the performance of HMM based tonal vowel recognition.

From the experiments it has been observed that the short-term phase based MGD feature vector is efficient in representing both tone variation as well as base-phoneme variation in case of tonal vowels. Only in the case of 6.46% cases the recognizer has been unable to recognize the tone variation of the same base-phone whereas in 2.91% cases tone takes more dominants over base-phone for tonal vowel recognition. This facts reassures the suitability of MGD feature for tonal vowel recognition in particular and language recognition in general.

## 5 Conclusion

It this paper we have investigated the performance of MGD features for their tonal vowels discrimination capability. It has been observed that phase-based MGD feature extracted from different tonal vowels is statistically separate from each other in the feature space even when they are different from each other only by tone or base-phone. This fact has been established by statistical measures Euclidean distance and F-ratio test. The performances of the features have been evaluated with a HMM based recognizer in terms of recognition accuracy. In 89.23% cases, the tonal vowels are recognized correctly by the HMM based recognizer trained and tested with MGD features. In the present investigation, it has been observed that MGD features are equally efficient in representing vowels with lexical tone (rising and falling) and vowels without any lexical tone (level tone). This observation appeals more in-depth investigation of the MGD feature for using it as a parameterization technique for language independent ASR system.

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

## References


1. U. Bhattacharjee, Recognition of the tonal words of bodo language. Int. J. Recent Technol. Eng. 1, (2013)
2. H.M. Wang, J.L. Shen, Y.J. Yang, C.Y. Tseng, S.L. Lee, Complete Chinese dictation system research and development. in *Proceedings ICASSP-94*, vol. 1. (1994), pp. 59–61
3. C.J. Chen, H. Li, L. Shen, G.K. Fu, Recognize tone languages using pitch information on the main vowel of each syllable, acoustics, speech, and signal processing. in *Proceedings (ICASSP'01), 2001 IEEE International Conference on*, vol. 1. (IEEE, 2001)
4. C.J. Chen, R.A. Gopinath, M.D. Monkowski, M.A. Picheny, K. Shen, in *New Methods in Continuous Mandarin Speech Recognition, 5th European Conference on Speech Communication and Technology*, vol. 3. (1997), pp. 1543–1546
5. P. Mowlaee, R. Saeidi, Y. Stylianou, Phase importance in speech processing applications. in *Fifteenth Annual Conference of the International Speech Communication Association* (2014)
6. B. Yegnanarayana, J. Sreekanth, A. Rangarajan, Waveform estimation using group delay processing. IEEE Trans. Acoust. Speech Signal Process. **33**(4), 832–836 (1985)
7. J. Deng, X. Xu, Z. Zhang, S. Frühholz, B. Schuller, Exploitation of phase-based features for whispered speech emotion recognition. IEEE Access **4**, 4299–4309 (2016)
8. L.D. Alsteris, K.K. Paliwal, Short-time phase spectrum in speech processing: a review and some experimental results. Digital Signal Process 17.3, 578–616 (2007)
9. R.M. Hegde, H.A. Murthy, V.R.R. Gadde, Signi_cance of the modi_ed group delay feature in speech recognition. IEEE Trans. Audio Speech Lang. Process. **15**(1), 190–202 (2007)
10. I. Hernáez, I. Saratxaga, J. Sanchez, E. Navas, I. Luengo, Use of the harmonic phase in speakerrecognition. in *Twelfth Annual Conference of the International Speech Communication Association* (2011)
11. B. Bozkurt, L. Couvreur, On the use of phase information for speech recognition. in *2005 13th European Signal Processing Conference* 2005 Sep 4. (IEEE, 2005), pp. 1–4
12. H. Banno, J. Lu, S. Nakamura, K. Shikano, H. Kawahara, Efficient representation of short-time phase based on group delay. in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98* (1998)
13. H.A. Murthy, B. Yegnanarayana, Speech processing using group delay functions. Signal Process. **22**(3), 259–267 (1991)
14. H. Murthy, V. Gadde, The modi_ed group delay function and its application to phoneme recognition, in *Proceedings ICASSP* (Hong Kong, 2003), pp. 68–71
15. D. Zhu, K.K. Paliwal, Product of power spectrum and group delay function for speech recognition. in *Proceedings ICASSP 04* (2004), pp. 125–128
16. M.W. Post, T. Kanno, Apatani phonology and lexicon, with a special focus on tone. Himalayan Linguist. **12**(1), 17–75 (2013)
17. J.T. Sun, Tani languages, in *The Sino-Tibetan Languages.* ed. by G. Thurgood, R. LaPolla (Routledge, London and New York, 2003), pp. 456–466
18. P.T. Abraham, *Apatani-English-Hindi Dictionary* (Central Institute of Indian Language, Mysore, India, 1987).
19. U. Bhattachajee, J. Mannala, An experimental analysis of speech features for tone speech recognition. Int. J. Innov. Technol. Exploring Eng. **9**(2), 4355–4360 (2019)
20. H. Patro, G.S. Raja, S. Dandapat, Statistical feature evaluation for classification of stressed speech. Int. J. Speech Technol. **10**(2–3), 143–152 (2007)
21. L. Rabiner et al, HMM clustering for connected word recognition. in *International Conference on Acoustics, Speech, and Signal Processing* (IEEE, 1989)
22. U. Bhattachajee, J. Mannala, Feature level solution to noise robust speech recognition in the context of tonal languages. Int. J. Eng. Adv. Technol. **9**(2), 3864–3870 (2019)

# Feature Level Solution to Noise Robust Speech Recognition in the context of Tonal Languages

**Utpal Bhattacharjee, Jyoti Mannala**

***Abstract:****Performance of a speech recognition system is highly dependent on the operational environments. The mismatched ambient conditions have adverse impact on the performance of an Automatic Speech Recognition (ASR) system. The speech parameterization techniques for tonal speech recognition are different from those used for non-tonal speech recognition. It is due to the fact that tonal speech has two components – basic linguistic unit and tone. The basic linguistic unit with different tones convey different meanings. Therefore, the feature set used for tonal speech recognition must have the capability to representing both of them. Tone is determined by the fundamental frequency of the speech signal which is highly sensitive to noise. Since at the time of parameterization of the non-tonal speech recognition systems, these highly noise-sensitive tone related information are discarded, the traditional noise elimination methods used for non-tonal speech recognition fail to deliver robust performance in tonal speech recognition. In the present study, we have analyze the performance of different commonly used feature sets for noisy tonal speech recognition. Hidden Markov Model (HMM) based speech recognizer has been used for performance evaluation. Noise elimination techniques sub-band spectral subtraction and Wiener filter have been used for noise reduction and their relative performance have been evaluated.*

***Keywords :****HMM, Noise elimination, Sub-band spectral subtraction, Tonal speech recognition, Wiener Filter*

## I. INTRODUCTION

Feature extraction is the front-end of any speech recognition system. The feature extraction for a speech recognition system is the process of reliable, compact and robust parameterization of the speech signal. The efficiency of the entire speech recognition system is highly dependent on proper parameterization of the speech signal. The feature vector extracted from the speech signal must have the capability to discriminating among different phonemes and must be robust to the environment and intra-speaker variability. The significance of cepstral features for speech recognition have been established by many researchers [1][2][3]. However, there are practical limitation in the use of cepstral features due to its sensitivity towards the background and channel noises [4].Mel frequency cepstral coefficients (MFCC) and linear predictor cepstral coefficients (LPCC) are two extensively used feature vector in speech science. MFCC feature is based on magnitude spectrum. A perceptually motivated frequency wrapping filter-bank is applied to the magnitude spectrum. The filters are evenly spaced on a perceptually motivated frequency wrapping scale call Mel-scale, first suggested by Stevens and Volkman [5]. The log-energy of each filter output is computed and accumulated. Finally, Discrete Cosine Transformation (DCT) is applied to produce the Mel frequency cepstral coefficients [6]. In the present study, a filter bank of 24 triangular filters spread across the whole frequency range from 0 to Nyquist frequency has been used. The first 12-cepstral coefficients and log energy have been considered as the MFCC feature vector. Linear predictor cepstral coefficient (LPCC) is a feature vector based on Linear predictor coefficient (LPC). The LPC are obtained using a $p^{th}$ -order All-pole approximation in the windowed waveform [7]. The autocorrelation method has been used to evaluate the linear predictor coefficients. The LPCC have been computed directly from LPC as [8]:

$$c_n = \begin{cases} a_n + \frac{1}{n}\sum_{m=1}^{n-1} mc_m a_{n-m}, & 1 \le n \le p \\ \frac{1}{n}\sum_{m=n-p}^{n-1} mc_m a_{n-m}, & n \ge p \end{cases} \quad \ldots(1)$$

where$p$ is the order of the predictor coefficients and $n$ is the number of cepstal coefficients. In the present study, 10$^{th}$ order LP analysis has been performed and 13 LPCC coefficients have been computed. To capture the dynamic property of the speech signal, along with baseline MFCC and LPCC features their first and second order derivatives are also added. Thus we get a 39-dimensinal MFCC feature set and a 39-dimensional LPCC feature set

Prosody plays an important role in understanding the meaning of a conversation in human to human communications. Prosodic features of speech characterize the paralinguistic information of a conversation like speaker habits, discourse structure, speaker intension, emotion etc. In general, prosody means the organization of a sound. Normally, it is represented by fundamental frequency ($F_0$), energy and normalized duration of syllable. The prosodic features are very important to identify the tone associated with a syllable. In the present study, in order to use only frame-based features, fundamental frequency and energy have been considered for the representation of prosodic information. Fundamental frequency and frame energy are static features, calculated frame by frame. In order to include temporal information, their first (Δ)- and second (ΔΔ)-order derivatives have been calculated and added to the feature set. Thus, we get a 6-dimensional prosodic feature vector for each frame.

**Revised Manuscript Received on December 15, 2019.**
***** Correspondence Author
**UtpalBhattacharjee*,**Department of Computer Science and Engineering, Rajiv Gandhi University, Rono Hills, Doimukh, Arunachal Pradesh, India, Pin - 791 112 Email: utpal.bhattacharjee@rgu.ac.in
**JyotiMannala,**Department of Computer Science and Engineering, Rajiv Gandhi University, Rono Hills, Doimukh, Arunachal Pradesh, India, Pin - 791 112 Email: mannalajoy@gmail.com

Left-to-Right Hidden Markov Model (LRHMM) has been used as baseline speech recognition system to recognize the tonal vowels of Apatani language of Arunachal Pradesh of North East India. The main reason for using LRHMM is that it can model the time varying property of the speech signal. A number of HMM states is determined empirically. In the present model, 6 (six) states have been used. Each state is represented by a single Gaussian distribution function given by [9]

$$P(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) \quad \dots (2)$$

Where $x$ is the observation vector, $\mu$ is the Gaussian mean vector and $\sigma^2$ is the variance. To initialize the model, speech signal from known vowels have been divided into 6 equal parts and each part from left to right has been used to initialize a state. The forward-backward algorithm has been used for training the HMM model. Clean speech signals have been used for the training purpose.

## II. NOISE ELIMINATION METHODS

The most commonly used de-noising techniques are based on either spectral subtraction or Weiner's filter. These techniques are based on the assumption that the speech signal $s(n)$and the additive noise $d(n)$are uncorrelated with each other. Therefore, the equation for noisy speech signal $x(n)$ can be represented as [10]

$$x(n) = s(n) + d(n) \quad \dots (3)$$

The original signal can be estimated from the noisy speech signal by using Wiener filter as:

$$\hat{S}(\omega) = H(\omega).X(\omega) \quad \dots (4)$$

Where $H(\omega)$, $X(\omega)$, $\hat{S}(\omega)$ are the Wiener filter response function, noisy signal and the estimated clean speech signal in frequency domain respectively. Wiener filter is an optimal filter that minimize the mean square error. The mean square error is represented by the function

$$\begin{aligned} E(\omega) &= S(\omega) - \hat{S}(\omega) \\ &= S(\omega) - H(\omega).X(\omega) \end{aligned} \quad \dots (5)$$

The $H(\omega)$ value is determined by minimizing the expectation of mean square error, which is obtained by taking first order derivative of the error function with respect to response function of the Weiner filter $H(\omega)$ and equating it to 0. The expectation of mean square error is given by

$$E[|E(\omega)|^2] = E[|S(\omega) - H(\omega).X(\omega)|^2] \quad \dots (6)$$

where $E[.]$ stands for expectation operation. Taking the derivatives of eq(6) and equating it to 0, we get

$$\begin{aligned} \frac{\delta E[|E(\omega)|^2]}{\delta H(\omega)} &= 2H(\omega)E[|X(\omega)|^2] - 2E[|X(\omega)S(\omega)^*|] \\ &= 2H(\omega)P_X(\omega) - 2P_{XS}(\omega) = 0 \end{aligned} \quad \dots (7)$$

where$P_X(\omega)$ and $P_{XS}(\omega)$ are power spectra of noisy speech and cross power spectra between noisy speech signal and clean speech respectively.In case of no correlation between the speech signal $s(n)$ and the additive noise $d(n)$, we get

$$\begin{aligned} P_X(\omega) &= E[|X(\omega)|^2] = E[|S(\omega) + D(\omega)|^2] \\ &= E[|S(\omega)|^2] + E[|D(\omega)|^2] + E[|S(\omega)D(\omega)|] \\ &= P_S(\omega) + P_D(\omega) \end{aligned} \quad \dots (8)$$

Similarly

$$\begin{aligned} P_{XS}(\omega) &= E[|X(\omega)S(\omega)^*|] \\ &= E\left[\left|\big(S(\omega) + D(\omega)\big)S(\omega)^*\right|\right] \\ &= E[|S(\omega)|^2] = P_S(\omega) \end{aligned} \quad \dots (9)$$

Therefore, the Wiener filter can be represented by:

$$H(\omega) = \frac{P_S(\omega)}{P_S(\omega) + P_D(\omega)} \quad \dots (10)$$

The signal-to-noise ratio is defined by

$$SNR = \frac{P_S(\omega)}{P_D(\omega)} \quad \dots (11)$$

Therefore, the impulse response of the Wiener filter can be represented in term of SNR as:

$$H(\omega) = \left[1 + \frac{1}{SNR}\right]^{-1} \quad \dots (12)$$

In the present work, we have implemented the adaptive Wiener filter based on the model proposed by El-Fattah et al[11] for speech enhancement. The mean $m_x$ and standard deviation ${\sigma_x}^2$of the speech signal have been estimated. It is assumed that the additive noise is of zero mean and variance${\sigma_d}^2$. The variance ${\sigma_d}^2$ has been estimated exploiting the silent period of the speech signal. Thus the power spectrum of noise has been estimated as

$$P_D(\omega) = {\sigma_d}^2 \quad \dots (13)$$

Considering a small segment of the speech signal, in which speech $x(n)$ is assumed to be stationary, the signal can be modelled as:

$$x(n) = m_x + {\sigma_x}^2 w(n) \quad \dots (14)$$

where$m_x$ and ${\sigma_x}^2$ mean and standard deviation of the speech signal for a small segment of the speech signal and $w(n)$ is unit variance noise. Therefore, for a small segment of the speech signal, the Wiener filter transfer function can be represented by:

$$H(\omega) = \frac{{\sigma_s}^2}{{\sigma_s}^2 + {\sigma_d}^2} \quad \dots (15)$$

Since $H(\omega)$ is constant over this small segment of speech, the impulse response of the Wiener filter can be obtained by

$$h(n) = \frac{{\sigma_s}^2}{{\sigma_s}^2 + {\sigma_d}^2}\delta(n) \quad \ldots (16)$$

The enhanced speech signal $\hat{s}(n)$ for the local segment can be expressed as:

$$\hat{s}(n) = m_x + (x(n) - m_x) * \frac{{\sigma_s}^2}{{\sigma_s}^2 + {\sigma_d}^2}\delta(n)$$
$$= m_x + \frac{{\sigma_s}^2}{{\sigma_s}^2 + {\sigma_d}^2}(x(n) - m_x) \quad \ldots (17)$$

If $m_x$ and ${\sigma_s}^2$ are updated for each segment, we can say

$$\hat{s}(n) = m_x(n) + \frac{{\sigma_s}^2}{{\sigma_s}^2(n) + {\sigma_d}^2}(x(n) - m_x(n)) \quad \ldots (18)$$

when${\sigma_s}^2$is much larger than ${\sigma_d}^2$, there will be no attenuation and the estimated speech signal $\hat{s}(n)$ will be basically due to $x(n)$. However, if ${\sigma_s}^2$ is smaller than${\sigma_d}^2$, there will be attenuation and the filtering will be done.The value of $m_x(n)$ can be estimated from the $x(n)$ as:

$$\hat{m}(n) = \frac{1}{2M+1}\sum_{k=n-M}^{n+M} x(k) \quad \ldots (19)$$

where$(2M + 1)$ is the number of sample in the short segment used for estimation. Since ${\sigma_x}^2 = {\sigma_s}^2 + {\sigma_d}^2$, ${\hat{\sigma}_s}^2(n)$ may be estimated from $x(n)$ as:

$${\hat{\sigma}_s}^2(n) = \begin{cases} {\hat{\sigma}_x}^2(n) - {\hat{\sigma}_d}^2 \\ 0, otherwise \end{cases}, if {\hat{\sigma}_x}^2(n) > {\hat{\sigma}_d}^2 \quad \ldots (20)$$

Where

$${\hat{\sigma}_x}^2(n) = \frac{1}{2M+1}\sum_{k=n-M}^{n+M}\left(x(k) - \hat{m}(n)\right)^2 \quad \ldots (21)$$

Another method for de-noising uncorrelated additive noise is spectral subtraction. The power spectra of the corrupted speech signal can be approximated from eq. (3) as

$$|X(k)|^2 = |S(k)|^2 + |D(k)|^2 \quad \ldots (22)$$

where$|S(k)|^2$ and $|X(k)|^2$ are the magnitude spectra of clean and the noise respectively. Since the noise spectra cannot be obtained directly, an estimate $\hat{D}(k)$is obtained from the silent period [12]. The estimation of clean speech spectrum is obtained by

$$\left|\hat{S}(k)\right|^2 = |X(k)|^2 - \alpha\left|\hat{D}(k)\right|^2 \quad \ldots (23)$$

where$\alpha$ is the over subtraction factor, which is a function of SNR. This model is based on the assumption that the noise affects the speech signal uniformly. However in case of real world operational conditions, this assumption is not true. It has been observed that impact of noise is different for different frequency range. Kamath and Loizou [13] proposed a multiband model for spectral subtraction. The entire frequency range of the speech signal is divided into *N* non-overlapping sub-bands and band-specific subtraction factor is computed for each frequency band. The estimation for clean speech of the $i^{th}$ band is obtained by

$$\left|\hat{S}(k)\right|^2 = |X_i(k)|^2 - \alpha_i\delta_i\left|\hat{D}(k)\right|^2, b_i \le k \le e_i \quad \ldots (24)$$

where$b_i$ and $b_i$ are beginning and ending frequency bins of the$i^{th}$ frequency band and $\delta_i$ is the tweaking factor for the $i^{th}$ band. The band specific SNR is computed using the magnitude of the noisy spectra and estimated noise spectra as follows:

$$SNR_i = 10\ \log_{10}\left(\frac{\sum_{b_i}^{e_i}|X_i(k)|^2}{\sum_{b_i}^{e_i}\left|\hat{D}_i(k)\right|^2}\right) \quad \ldots (25)$$

Using the SNR value $\alpha_i$ is computed as:

$$\alpha_i = \begin{cases} 5 & SNR_i < -5 \\ 4 - \frac{3}{20}SNR_i & -5 \le SNR_i \le 20 \\ 1 & SNR_i > 20 \end{cases} \quad \ldots (26)$$

The negative value of the enhanced spectra is floored to the noisy spectra.

## III. SPEECH DATABASE

A speech database of Apatani tonal words has been prepared to carry out the experiments. The Apatani language of Arunachal Pradesh of North Eastern India is a 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 [14].Apatani has two lexical tones raising (ˊ) and falling (ˋ) [15]. In addition to these two tones, Apatani has words without any associated tone, which are referred to as normal tone. Except the vowel [ə] all the other vowels have 3 tonal instances namely raising, falling and level. In case of vowel [ə] only level tone has been observed. In the evaluation of the speech recognition system for tonal speech recognition task, the vowel [ə] has not been taken into consideration. The database for the present research consist of 24 isolated tonal words spoken by 20 different speakers (13 males and 7 females).The words chosen for recording are:

**Table-1: Tonal words considered for recording**

| Sl no. | Apatani Tonal Words | Meaning in English |
|---|---|---|
| 1 | tá | Bite |
| 2 | tɑ | Listen |
| 3 | tà | Drink |
| 4 | khɛ̀ | Cry |
| 5 | khɛ | To get angry |
| 6 | khɛ́ | Remove |
| 7 | cɪ | Cut with scissor |
| 8 | cì | Bring together two things |
| 9 | jì | Be black |

| | | |
|---|---|---|
| 10 | jí | Roll |
| 11 | jɪ | Bind |
| 12 | àlɔ̀ | Salt |
| 13 | àlɔ | Dry |
| 14 | kɔrɔ | Day before yesterday |
| 15 | kɔ́rɔ | Fence |
| 16 | ɑpó | Blossom |
| 17 | ápʊ | Wrap Up |
| 18 | kʊ | Beg |
| 19 | kʊ̀ | Spray |
| 20 | kʊ́ | Wave like movement |
| 21 | ɑnó | Young Brother |
| 22 | ɑnʊ | Uncle |
| 23 | mədɔ́ | Rain |
| 24 | mədɔ | Doing |

For any tone language, the basic building blocks are tonal syllables. 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 [16]. The tone associated with the vowels are sufficient to express the tone associated with the syllable. Therefore, in the present study we will evaluate robustness of a tonal speech recognition system in terms of its capability to recognize tonal vowel at different noise conditions. The words are recorded 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. Each speaker uttered the same words 5 times. From the recorded isolated words, a vowel database has been created by segmenting the vowels from the isolated words. The segmentation has been done by using PRAAT software which is followed by subjective verification. Thus we get at least 100 instances for each tonal vowel. The database has been divided into two parts – training set and testing set. The training set consist of 50 instances of each tonal vowel and the testing set consist of remaining 50 instances of each tonal vowel.

From the clean database noisy versions of the database has been created by adding noise from the AURORA database [17]. The noises added to the database are babble, car, exhibition, restaurant, street, subway and train noises. The noises are added at -15dB, -10dB, -5dB, 0dB, 5dB, 10dB and 15dB signal-to-noise ratio (SNR).

## IV. RESULTS AND DISCUSSION

The speech has been analyzed using a Hamming windows of length 25 ms, frame rate 100 Hz and pre-emphasis factor of 0.97. MFCC, LPCC and prosodic features have been extracted from each frame. Now from the extracted features two tonal feature sets have been created by appending the prosodic features with MFCC and LPCC features separately. We call them MFCC tonal feature and LPCC tonal feature respectively.

To study the suitability of the feature sets for tonal speech recognition their probability density function (PDF) characteristics have been analyzed. If the same vowel with different tone have different PDF characteristics for a particular feature set, then the feature set will be efficient in recognizing the tonal instances of the vowels. PDF characteristics of the MFCC and LPCC tonal feature sets are given in Fig-1 and Fig-2 respectively.

From the Figures it has been observed that both MFCC and LPCC feature sets the peak of the distribution are at different positions. For the vowel [ɔ] the MFCC tonal feature has more tonal phoneme discrimination capability while for vowel [ɑ] the LPCC tonal feature exhibits more tone discrimination capability. In case of vowels [ɛ] and [ʊ], both MFCC and LPCC tonal features display tone discrimination capability. This observation justify the fact that tone discrimination capability of a feature set depends on the underlying vowels.

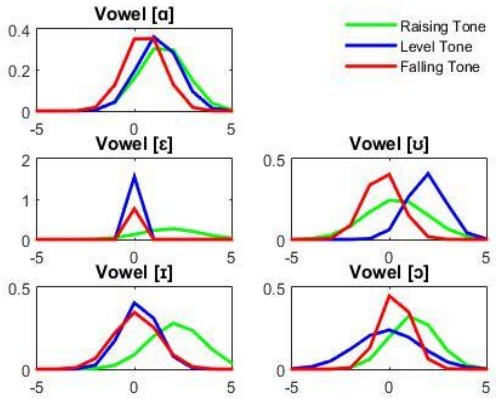


**Fig. 1 PDF characteristics of tonal vowels for MFCC tonal feature set**

To evaluate the efficiency of the feature set in recognizing the tonal vowels, a Hidden Markov Model based recognizer has been trained using the clean training set. The testing has been done using the testing set and the confusion matrices for recognition of the tonal vowels have been prepared. The confusion matrices for the MFCC and LPCC tonal feature sets based HMM recognizer for recognizing the tonal vowels have been given in Table – 2 and Table -3 respectively.

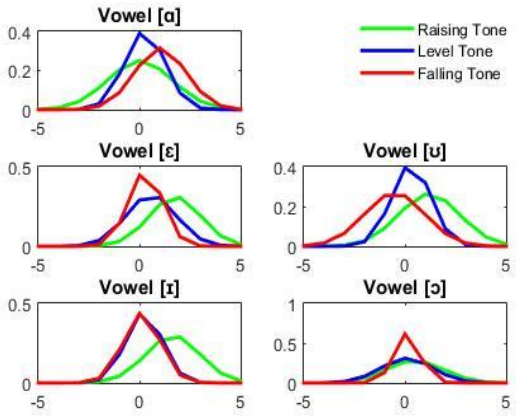


**Fig. 2 PDF characteristics of tonal vowels for LPCC tonal feature set**

**Table – 2: Confusion matrix for tonal phoneme recognition with tonal MFCC and HMM based recognizer (50 test for each tonal vowel)**

| | [ā:] | [ á:] | [ à:] |
|---|---|---|---|
| [ā:] | 41 | 6 | 3 |
| [ á:] | 4 | 43 | 3 |
| [ à:] | 1 | 2 | 47 |
| | [ī] | [ í] | [ ì] |
| [ī] | 43 | 7 | 0 |
| [ í] | 6 | 44 | 0 |
| [ ì] | 0 | 0 | 50 |
| | [ɔ̄] | [ ɔ́] | [ ɔ̀] |
| [ɔ̄] | 49 | 1 | 0 |
| [ ɔ́] | 2 | 48 | 0 |
| [ ɔ̀] | 1 | 0 | 49 |
| | [ɛ̄] | [ ɛ́] | [ ɛ̀] |
| [ɛ̄] | 48 | 1 | 1 |
| [ ɛ́] | 0 | 48 | 2 |
| [ ɛ̀] | 1 | 1 | 48 |
| | [ū] | [ ú] | [ ù] |
| [ū] | 47 | 1 | 2 |
| [ ú] | 1 | 48 | 1 |
| [ ù] | 0 | 1 | 49 |

**Table 3: Confusion matrix for tonal phoneme recognition with tonal LPCC and HMM based recognizer (50 test for each tonal vowel)**

| | [ā:] | [ á:] | [ à:] |
|---|---|---|---|
| [ā:] | 48 | 1 | 1 |
| [ á:] | 0 | 49 | 1 |
| [ à:] | 0 | 0 | 50 |
| | [ī] | [ í] | [ ì] |
| [ī] | 39 | 8 | 3 |
| [ í] | 5 | 40 | 5 |
| [ ì] | 2 | 2 | 46 |
| | [ɔ̄] | [ ɔ́] | [ ɔ̀] |
| [ɔ̄] | 41 | 6 | 3 |
| [ ɔ́] | 9 | 37 | 4 |
| [ ɔ̀] | 2 | 10 | 38 |
| | [ɛ̄] | [ ɛ́] | [ ɛ̀] |
| [ɛ̄] | 47 | 0 | 3 |
| [ ɛ́] | 1 | 49 | 0 |
| [ ɛ̀] | 0 | 2 | 48 |
| | [ū] | [ ú] | [ ù] |
| [ū] | 49 | 0 | 1 |
| [ ú] | 2 | 48 | 0 |
| [ ù] | 1 | 1 | 48 |

From the above confusion matrices it has been observed that the tonal phoneme recognition accuracy depends on the feature set and the underlying vowel. The recognition accuracy of the HMM based recognizer in tonal phoneme discrimination using different feature sets have been summarized in table-4.

**Table -4: Recognition accuracy of the HMM based recognizer for tonal phoneme recognition for different feature sets**

| Tonal Vowel | MFCC tonal Feature set (in %) | LPCC tonal feature set (in %) |
|---|---|---|
| [ā:] | 82 | 96 |
| [ á:] | 86 | 98 |
| [ à:] | 94 | 100 |
| [ī] | 86 | 78 |
| [ í] | 88 | 80 |
| [ ì] | 100 | 92 |
| [ɔ̄] | 98 | 82 |
| [ ɔ́] | 96 | 74 |
| [ ɔ̀] | 98 | 76 |
| [ɛ̄] | 96 | 94 |
| [ ɛ́] | 96 | 98 |
| [ ɛ̀] | 96 | 96 |
| [ū] | 94 | 98 |
| [ ú] | 96 | 96 |
| [ ù] | 98 | 96 |
| Average | 93.6 | 90.27 |

In the next set of experiments, we have considered the noisy versions of the database and their performances have been evaluated using the same HMM model which is trained with clean speech. The recognition accuracy under different noise types and noise levels is given in table-5 and table-6.

**Table – 5: The recognition accuracy of HMM and MFCC tonal feature based speech recognizer for recognizing noisy tonal vowels**

| Noise Type | -15 dB | -10 dB | -5 dB | 0 dB | 5 dB | 10 dB | 15 dB |
|---|---|---|---|---|---|---|---|
| **Babble** | 23.4 | 25.4 | 26.7 | 33.8 | 37.4 | 54.2 | 67.3 |
| **Car** | 24.0 | 26.5 | 27.6 | 32.8 | 38.6 | 49.8 | 69.6 |
| **Exhibition** | 22.6 | 28.6 | 29.1 | 33.1 | 40.7 | 52.8 | 73.3 |
| **Restaurant** | 22.8 | 25.3 | 24.6 | 31.7 | 34.4 | 58.0 | 62.0 |
| **Street** | 28.2 | 25.9 | 25.2 | 35.4 | 35.3 | 48.9 | 63.5 |
| **Subway** | 24.6 | 29.5 | 28.8 | 38.6 | 40.3 | 52.3 | 72.6 |
| **Train** | 25.8 | 28.2 | 26.1 | 37.2 | 37.8 | 52.6 | 68.0 |

**Table – 6: The recognition accuracy of HMM and LPCC tonal feature based speech recognizer for recognizing noisy tonal vowels**

| Noise Type | -15 dB | -10 dB | -5 dB | 0 dB | 5 dB | 10 dB | 15 dB |
|---|---|---|---|---|---|---|---|
| **Babble** | 21.1 | 22.4 | 23.8 | 29.9 | 33.2 | 48.0 | 59.7 |
| **Car** | 23.3 | 20.7 | 24.2 | 27.1 | 32.9 | 41.8 | 58.8 |
| **Exhibition** | 22.4 | 25.7 | 27.5 | 30.5 | 38.0 | 49.0 | 68.2 |
| **Restaurant** | 19.8 | 21.3 | 21.0 | 26.9 | 29.3 | 49.3 | 52.7 |
| **Street** | 24.8 | 25.4 | 23.4 | 33.8 | 33.3 | 46.4 | 60.0 |
| **Subway** | 21.2 | 23.6 | 23.9 | 31.5 | 33.1 | 42.8 | 59.6 |
| **Train** | 23.0 | 22.3 | 21.9 | 30.3 | 31.3 | 43.2 | 56.1 |

From the above results, it has been observed that the recognition accuracy of the HMM based recognizer degrades considerably when noise presents in the speech signal. The performance deterioration is different for different noise types.

Further, it has been observed that MFCC tonal feature outperforms LPCC based tonal feature under all operational conditions. Therefore, MFCC tonal feature may be considered as better parameterization technique for tonal speech recognition under all operational conditions. Therefore, the performance of the de-noising techniques have been evaluated with MFCC tonal feature only.

To de-noise the corrupted speech signal, we apply Wiener Filter and sub-band spectral subtraction methods separately and the performance haves been evaluated. The result of the experiments are given in table-7 and table-8.

**Table – 7: The recognition accuracy of HMM and MFCC tonal feature based speech recognizer for recognizing tonal vowels at different noise conditions with Wiener Filter de-noising technique**

| Noise Type | -15 dB | -10 dB | -5 dB | 0 dB | 5 dB | 10 dB | 15 dB |
|---|---|---|---|---|---|---|---|
| **Babble** | 33.7 | 35.8 | 38.0 | 47.9 | 53.1 | 66.8 | 83.0 |
| **Car** | 41.9 | 37.2 | 43.5 | 48.9 | 59.1 | 61.9 | 87.7 |
| **Exhibition** | 35.8 | 41.2 | 44.0 | 48.9 | 60.8 | 71.3 | 89.6 |
| **Restaurant** | 31.7 | 34.0 | 33.7 | 43.0 | 46.9 | 65.6 | 70.3 |
| **Street** | 39.7 | 40.6 | 37.5 | 54.1 | 53.2 | 69.0 | 89.0 |
| **Subway** | 37.2 | 41.5 | 42.1 | 55.4 | 58.3 | 60.5 | 84.3 |
| **Train** | 41.3 | 40.1 | 39.5 | 54.6 | 56.3 | 62.6 | 81.5 |

**Table – 8: The recognition accuracy of HMM and MFCC tonal feature based speech recognizer for recognizing tonal vowels at different noise conditions with sub-band spectral subtraction de-noising technique**

| Noise Type | -15 dB | -10 dB | -5 dB | 0 dB | 5 dB | 10 dB | 15 dB |
|---|---|---|---|---|---|---|---|
| **Babble** | 29.7 | 30.8 | 33.2 | 41.5 | 46.2 | 76.9 | 95.5 |
| **Car** | 39.8 | 28.4 | 37.3 | 39.6 | 49.3 | 75.2 | 95.9 |
| **Exhibition** | 34.7 | 36.3 | 40.7 | 44.2 | 55.6 | 78.4 | 94.2 |
| **Restaurant** | 27.1 | 28.0 | 28.2 | 35.7 | 39.1 | 78.8 | 84.4 |
| **Street** | 34.2 | 39.0 | 34.2 | 50.6 | 49.2 | 74.2 | 96.1 |
| **Subway** | 31.4 | 32.6 | 34.2 | 44.2 | 47.0 | 75.4 | 97.9 |
| **Train** | 36.0 | 31.0 | 32.5 | 43.6 | 45.7 | 77.8 | 94.9 |

From the above experiments it has been observed that the Wiener filter gives better performance in high noise condition whereas the sub-band spectral subtraction gives better performance at low noise condition. At 10dB and 15dB noise level, the sub-band spectral subtraction method outperforms the Wiener filter in noise compensation.

In the next experiment, we have combined the sub-band spectral subtraction and Wiener filter method. The speech spectra first goes through a sub-band spectral subtraction method and then Wiener filter is applied. The result of the experiment is summarized in table-9.

**Table-9: The recognition accuracy of HMM and MFCC tonal feature based speech recognizer for recognizing tonal vowels at different noise conditions with sub-band spectral subtraction and Wiener filter de-noising techniques**

| Noise Type | -15 dB | -10 dB | -5 dB | 0 dB | 5 dB | 10 dB | 15 dB |
|---|---|---|---|---|---|---|---|
| **Babble** | 38.0 | 40.0 | 42.7 | 53.6 | 59.6 | 86.2 | 89.3 |
| **Car** | 49.0 | 39.4 | 48.4 | 53.1 | 65.1 | 82.3 | 96.8 |
| **Exhibition** | 42.3 | 46.5 | 50.8 | 55.8 | 69.9 | 89.8 | 94.0 |
| **Restaurant** | 35.3 | 37.2 | 37.1 | 47.2 | 51.6 | 86.7 | 77.3 |
| **Street** | 44.4 | 47.8 | 43.0 | 62.8 | 61.4 | 85.9 | 92.6 |
| **Subway** | 41.2 | 44.5 | 45.8 | 59.8 | 63.2 | 81.5 | 94.6 |
| **Train** | 46.4 | 42.7 | 43.2 | 58.9 | 61.2 | 84.2 | 91.2 |

From the above results it has been observed that under certain noise conditions one de-noising technique gives better performance over the other technique. However, when both the methods are combined together, it gives a consistently optimal performance under all operational conditions.

## V. CONCLUSION

In this paper, the robustness issue of MFCC and LPCC features combined with prosodic features has been evaluated for tonal speech recognition. In case of tonal speech recognition only the spectral features like MFCC and LPCC are not sufficient as they does not conation tone related information. Therefore, prosodic features must have to be combined with them. Prosodic feature, which is determined by fundamental frequency and energy is highly sensitive to noise. Therefore, at noisy environmental conditions the performance of the speech recognition system degrades considerably. In the present study it has been observed that under controlled environmental conditions, both MFCC + Prosodic features and LPCC + prosodic features perform well in recognizing the tonal speech. However, with increasing level of noise, the performance degrades considerably. The degradation is more in case of LPCC + prosodic features compared to MFCC + prosodic features. Considering all operational conditions it has been observed that MFCC + prosodic feature is a better option for recognizing tonal speech. Two most commonly used de-noising techniques sub-band spectral subtraction and Wiener filter have been used for noise elimination in

the present work. It has been observed that Wiener filter perform significantly well in high noise conditions whereas sub-band spectral subtraction gives better performance in low noise condition. Combining both the methods, we have observed that the performance has consistently improves in all noise conditions. However, for some noise conditions, this performance is lower than the performance of individual techniques. Considering an optimal operational scenario, we have suggested that sub-spectral subtraction combined with Wiener filter is a viable noise reduction technique for tonal speech recognition.

## ACKNOWLEDGMENT

This work is supported by UGC major project grant MRP-MAJOR-COM-2013-40580, Ministry of Human Resource Development, Government of India.

## REFERENCES

1. M. Baloul, E. Cherrier, and C. Rosenberger. "Challenge-based speaker recognition for mobile authentication." Biometrics Special Interest Group (BIOSIG), 2012 BIOSIG-Proceedings of the International Conference of the. IEEE, 2012.
2. N. Desai, K. Dhameliya, and V. Desai. "Feature extraction and classification techniques for speech recognition: A review." International Journal of Emerging Technology and Advanced Engineering 13.12: 367-371, 2013.
3. W. Han, et al. "An efficient MFCC extraction method in speech recognition."2006 IEEE international symposium on circuits and systems. IEEE, 2006.
4. X. Zhao, and D. Wang. "Analyzing noise robustness of MFCC and GFCC features in speaker identification." 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2013.
5. S. S. Stevens and J. Volkman, The Relation of Pitch to Frequency, A Revised Scale. In: American Journal of Psychology, 53, 1940.
6. B.J. Shannon and K. K. Paliwal. "A comparative study of filter bank spacing for speech recognition." Microelectronic engineering research conference. Vol. 41. 2003.
7. L. Rabiner and M. Sambur. "Application of an LPC distance measure to the voiced-unvoiced-silence detection problem." IEEE Transactions on Acoustics, Speech, and Signal Processing 25, no. 4: 338-343, 1977.
8. B.S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification." the Journal of the Acoustical Society of America 55, no. 6 : 1304-1312. 1974.
9. L. Rabiner, et al. "HMM clustering for connected word recognition." International Conference on Acoustics, Speech, and Signal Processing,. IEEE, 1989.
10. J. Meyer and K.U. Simmer. "Multi-channel speech enhancement in a car environment using Wiener filtering and spectral subtraction." In 1997 IEEE international conference on acoustics, speech, and signal processing, vol. 2, pp. 1167-1170. IEEE, 1997.
11. M.A. Abd El-Fattah, et al. "Speech enhancement using an adaptive wiener filtering approach." Progress in Electromagnetics Research 4: 167-184, 2008.
12. M. Berouti, R. Schwartz and J. Makhoul, “Enhancement of speech corrupted by acoustic noise,” Proc.IEEE Int. Conf. Acoust., Speech, Signal Process., pp.208-211, Apr. 1979.
13. S. Kamath and P.Loizou. "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise." In ICASSP, vol. 4, pp. 44164-44164. 2002.
14. M. Yip, The Tonal Phonology of Chinese, New York: Garland Publishing, 1991.
15. 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.
16. P. Sarmah, “Tone Systems of Dimasa and Rabha: A Phonetic and Phonological Study”, Doctoral dissertation, University of Florida, 2009.
17. http://ecs.utdallas.edu/loizou/speech/noizeus/ accessed on 23rd October, 2019.

## 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

**JyotiMannala**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.

# 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. 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.

**Revised Manuscript Received on December 05, 2019.**

**Utpal Bhattacharjee**, Department of Computer Science and Engineering, Rajiv Gandhi University, Rono Hills, Doimukh, Arunachal Pradesh, India, Pin - 791 112
Email: utpal.bhattacharjee@rgu.ac.in
**Jyoti Mannala**, Department of Computer Science and Engineering, Rajiv Gandhi University, Rono Hills, Doimukh, Arunachal Pradesh, India, Pin - 791 112 Email: mannalajoy@gmail.com

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:

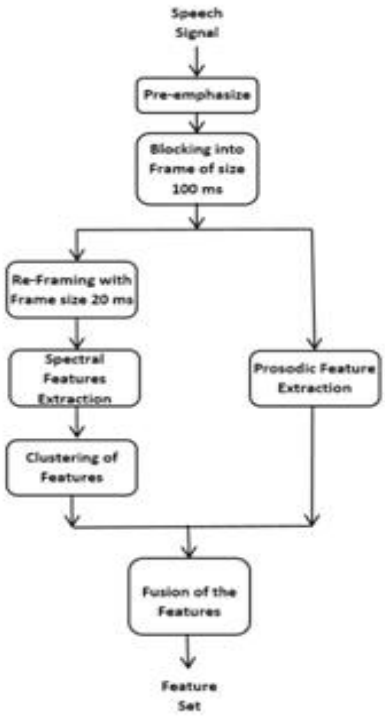


**Fig. 1. Block diagram of the hybrid feature extraction system**

**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, $\mu_i$ 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.

# Statistical Evaluation of Spectral Features for Tonal Phoneme Discrimination Capability

Utpal Bhattacharjee[1], Jyoti Mannala[2] and Gyati Yubbey[3]
*Department of Computer Science and Engineering*
*Rajiv Gandhi University, Rono Hills*
Doimukh, Arunachal Pradesh, India - 791112
[1]utpal.bhattacharjee@rgu.ac.in, [2]mannalajoy@yahoo.co.in and [3]gyatiappa@gmail.com

***Abstract*—Speech recognition is a well know problem in the area of speech science and machine learning. Lots of progress have already been made in this direction. However, there are some areas which need in-depth study. One such issue associated with the global acceptability of speech recognition system. Speech recognition system and technology developed for a particular linguistic group fails to deliver when tested with another completely distinct linguistic group. In the present study, attempt has been made to characterize the features used in speech recognition system to identify their intra-phoneme and inter-phoneme discriminating capability with reference to the tonal languages. The objective of the study is to identify features that can be used for both tonal and non-tonal speech recognition.**

***Index Terms*—Feature Analysis, Tonal Language, Speech Recognition, Statistical Evaluation**

## I. Introduction

In the recent years, significant progress has been made in speech recognition technology, making it a strong modality for human-machine interaction. The current use of speech recognition technology in some commercial appliances is just tip of the iceberg and its full power and potential is yet to explore. However, to realize such a potential, speech recognition technology must be able to deliver nearly human-like recognition performance in all operational conditions. Many leading researcher in the field understand the fragile nature of the current speech recognition systems [1]. The above observation triggers the need for in-depth study of all aspects of speech recognition technology to identify areas that need improvement.

Language portability is a major aspect for global acceptability of speech recognition system. A system developed for a particular language should be susceptible to any another language with minimal set of training data. However, it has been observed that the technology and systems which have been developed for non-tonal languages such as Indo-Aryan languages perform very poorly for the recognition of tonal languages. A language is said to be tonal if words with same phonetic contents but different lexical tone patterns convey different meaning. Tone information is generated by excursion of the fundamental frequency. As the lexical tones do not contain any meaningful information for non-tonal languages, the feature extraction process of speech recognition systems developed for non-tonal languages discard those information. As a result, the systems which are very efficient in recognizing non-tonal languages fail to perform satisfactorily for tonal languages. A major section of world population spreading across south-east Asia, East Asia and Sub-Sahara Africa are speaker of tonal language [2]. In the present study, statistical analysis of five most commonly used features namely - Mel Frequency Cepstral Coefficients (MFCC), Linear Frequency Cepstral Coefficients (LFCC), Linear Predictor Cepstral Coefficients (LPCC), Reflection Coefficients (RC) and Log Area Ratio (LAR) has been done for their relative effectiveness in tonal speech recognition.

In the last three decades, many attempts have been made for the development of speech recognition system for tonal languages. A popular method for recognizing tonal language is the two step method [3], first to recognize the based syllable by its phonetic contents. In the second step, recognize the tone of the syllable by classifying the pitch contour of that syllable using discriminating rules. Recognition of tonal syllable is a combination of the recognition of base syllable and the associated tone. The above method works well in isolated-syllable speech recognition but difficult to handle continuous speech. To overcome the problems of two step method, one step method has been developed consideration disyllable approach [4]. In this approach, each syllable is decomposed into two demi-syllables. The first demi-syllable contain tone-independent phone information. The second demi-syllable called toneme, carries the tone information of the whole syllable. In this approach, the demi-syllable with different toneme are considered as different phonemes. This approach work fine with tonal language with small set of tones associate with the phonemes. However, for languages with large number of tones associated with each phoneme, the number of phonemes increases exponentially. As a result, the search space for the recognizer increases and the entire recognition task is slow down. In another approach, which is based on the observation that pitch information of the main vowel is sufficient to determine the tone of the whole syllable [5]. Using this approach, number of phonemes can be drastically reduced. In addition, phonemes from this method are close to the Indo-Aryan languages.

To analyse the relative performance of different speech features in recognition of tonal and non-tonal phonemes,

This work has been supported by UGC Major Project Grant MRP-MAJOR-COMP-2013-40580, Ministry of HRD, Government of India.

three statistical evaluation methods have been used. They are, probability density function (PDF) characteristics, Analysis of variance (F-ratio) and Kullback-Leibler divergence. Data distribution of a class close to the normal distribution leads to better classification [6]. Probability density functions (PDF) of a feature vector have been estimated for different phonemes as well as different tonal instances of each phoneme to evaluate their relative effectiveness in discriminating among the phonemes and discriminating among the tonal instances of the same phoneme. A feature will perform well if the the peaks of the distributions are different for various phonemes even when they are different from each other only because of of tone. Both mean and variance of a feature is important from the point of view of its discriminating power in a particular application. A feature set with higher F-ratio value will produce better recognition accuracy [7]. The F-ratio value for each coefficient of a feature vector has been evaluated separately and the average has been taken as F-ratio for that feature. To measure the quantity of deviation in the distribution of feature vector under different tones, Kullback-Leibler divergence (KLD) is used. KLD is defined as the relative entropy of two density functions. The relative entropy between two distributions is null if the distributions are identical. Thus, the divergence between two distributions indicate how distinct they are. In the present study, a coefficient with higher KLD value among different tones will have better tonal-syllable discriminating capability.

## II. Spectral Features

Feature is the compact representation of the acoustic properties of a speech signal. The commonly used features for speech recognition are Mel Frequency Cepstral Features (MFCC), Linear Frequency Cepstral Features (LFCC), Linear Prediction Cepstral Features (LPCC) and a set of reflection coefficients (RC). Davis and Mermelstein [8] classify the features into two categories - frequency based features and linear predictor spectrum based features. The frequency based features are extracted directly from the frequency domain representation of the speech signal. MFCC and LFCC are in the first category. The second group includes linear predictor cepstral coefficient (LPCC) derived from linear predictor coefficients (LPC), reflection coefficients (RC) and log-area ratio (LAR).

MFCC are obtained from a spectrum filtered by mel scale [9]. MFCC are the most widely used features for speech and speaker recognition. It is based on human perception of critical bandwidths. It is based on the observation that high frequencies are captured by humans ear with less precision in comparison to low frequencies. Therefore, it gives linear frequency resolution up to 1000 Hz and logarithmic resolution at higher frequencies. MFCC are computed from the filter-bank output as [10]

$$MFCC_i = \sum_{k=1}^{N} X_k \left[ i(k - \left(\frac{1}{2}\right) \frac{\pi}{N} \right] i = 1, 2, 3...M \quad (1)$$

where $M$ is the number of cepstal coefficients, $X_k, k = 1, 2, ., N$ represents the log energy output of the $k^{th}$ filter bank. $N$ is the number of triangular filters in the filter bank.

LFCCs are computed from the log-magnitude spectra of the speech signal as

$$LFCC_i = \sum_{k=0}^{N-1} Y_k \left( \frac{\pi ik}{K} \right), i = 1, 2, ....M \quad (2)$$

where $K$ is the number of log-magnitude DFT coefficients $Y_k$.

Linear predictor based features has wide application in the area of speech science [11]. In linear predictor method, the current sample is estimated from the past p samples using a linear predictor

$$\hat{x}[n] = \sum_{k=1}^{p} a_k x[n-k] \quad (3)$$

where $\hat{x}[n]$ is the $n^{th}$ predicted sample of the speech signal, $a_k$ represents the $k^{th}$ predictor coefficient and $p$ is the order of the linear predictor. The LPCC were obtained from the $p^{th}$ order LP coefficients directly as [12]

$$\begin{aligned} LPCC_i &= \lg(G), i = 0 \\ &= a_i + \sum_{k=0}^{i-1} \left( \frac{k}{i} \right) LPCC_i.a_k, i = 1, 2, ...p \\ &= \sum_{k=i-p}^{i-1} \left( \frac{k}{i} \right) LPCC_k.a_{i-k}, i > p \end{aligned} \quad (4)$$

where $LPCC_k$ is the $k^{th}$ linear predictor cepstral coefficient.

The reflection coefficient (RC) were obtained by a transformation of the LP coefficients. It is equivalent to matching the inverse of the LP spectrum with a transfer function spectrum that corresponds to an acoustic tube consisting of p sections of various cross section areas [13]. The change in cross-sectional area of the tube boundaries can be represented by p+1 reflection coefficients. If the volume velocity of air flow at the glottis $v_g$ and at the lips is $v_l$, then the transfer function is [13]

$$\frac{v_l}{v_g} = \frac{0.5\,(1 + r_g) \prod_{k=1}^{p} (1 + r_k)}{1 - \sum_{k=1}^{p} a_k z^{-k}} \quad (5)$$

where $a_k$ are predictor coefficients, $r_g$ reflection coefficient at glottis and $r_k$ is the $k^{th}$ reflection coefficient. If the value of $r_g$ is known or assumed then the rest of the reflection coefficients can be calculated from $a_k$. It is assumed that $r_g = 1$.

Log-area ratio coefficients are the natural logarithm of the ratio of the areas of adjacent sections of a lossless tube equivalent to the vocal tract. It is possible to estimate the ratio of adjacent sections, though the absolute values of those areas could not be computed. The Log-area ratios can be found from the reflection coefficients as

$$g_k = ln \left( \frac{1 - r_k}{1 + r_k} \right) \quad (6)$$

where $g_k$ is the LAR and $r_k$ is the corresponding reflection coefficient.

Furui [14] observed that combination of instantaneous and dynamic features of the speech spectrum increases the recognition accuracy of the speech recognition system. To capture the dynamic features of the speech spectrum, the first order and second order derivatives for each feature vector has been calculated and combined with the respective feature.

## III. Speech Database

A speech database for the Apatani language of Arunachal Pradesh of India has been collected for tonal speech recognition. Apatani belongs to Sino-Tibetan group of languages. Apatani has 6 vowels and 17 consonants [15]. The vowels and the consonants are listed in the table given below:

TABLE I: Vowels and Consonants of Apatani Language

| **Category** | ***List of Phonemes*** |
|---|---|
| Vowel | *a, i, u, ü, e, o* |
| Consonants | *k, $k^h$, g, ŋ, c, j, t, d, l, n, p, b, m, y, l, s, h* |

The Apatani language has two lexical tones: raising and falling. The tones are associated with all the six vowels. In addition, Apatani language has words without any change in lexical tone, which is considered as level tone. A database of 58 isolated Apatani words has been created. Each word is uttered by 20 speakers (12 male and 8 female). The words are selected in such a way that they include each vowel for its all three tonal instances. Recording has been done at 16 KHz sampling frequency at mono channel format with 16 bit resolution in a controlled environment.

## IV. Experiment and Results

All the experiments were carried out using the database described in Section III. The vowels are segmented from the isolated words for all its tonal instances. For each speaker there are 4 occurrences of each vowel for each tonal instance. Thus there are 12 occurrences of a vowel for each speaker containing all the tones associated with it. The segmentation has been done using PRAAT software which is followed by subjective verification. In the present study, each tonal instance of a vowel has been considered as a separate vowel. That is, with vowel /*a*/, there are three instances /$\bar{a}$/ with level tone, /á/ with raising tone and /à/ with falling tone. Therefore, we consider them as three different phonemes. Each feature has been evaluated for their inter phoneme variability to intra-phoneme variability considering those tonal instance as separate phoneme. Higher inter-phoneme variability indicate better discrimination capability for the feature vector.

Statistical modelling techniques such as HMM used Gaussian probability density function to represent the area in the feature space occupied by a particular phoneme class. The Gaussian PDFs are characterized by their mean vector and a covariance matrix estimated during training. For different phoneme, the mean and variance are different. Speech features will perform well if these two properties have different values for different phonemes. For tonal phoneme representation, two phonemes only distinct by tone should also have the different mean and variance. A feature set with a sharper probability density characteristic produces better recognition accuracy [16].

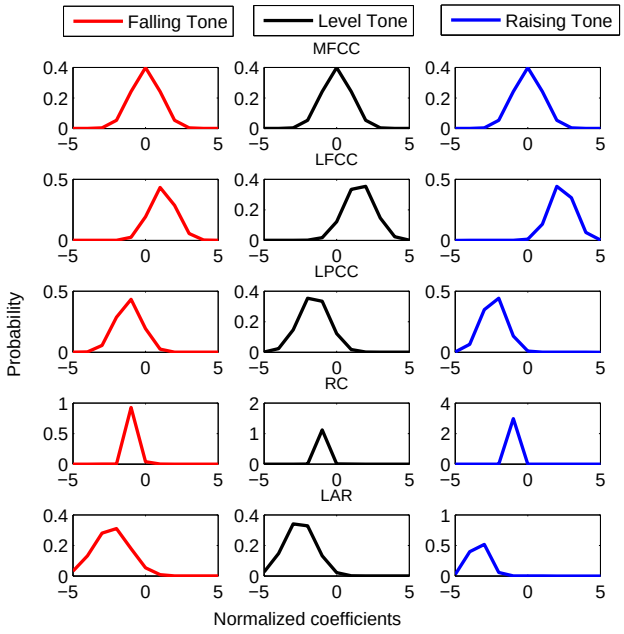


Fig. 1: Probability density function for the 1st coefficient of the features

Fig. 1 represents the probability density characteristics of the first coefficient of all the five feature vectors. It has been observed that the PDF characteristics resemble Gaussian distribution for all the features. Further, for different tonal instances of the vowels, the the peaks are at different positions for all the features except for MFCC coefficient. Therefore, the features may be utilized for the representation of tonal phonemes with good discriminating capability among different tonal versions of the same phoneme.

Fishers Discrimination ration (F-ratio) [5] has been used as a quantitative method for evaluation the discriminating capability among different tonal versions of the phonemes. F-ratio has been defined as

$$F = \frac{\text{Variance of the tonal phoneme mean}}{\text{Average intra-phoneme variance for all tones}} \tag{7}$$

The above ratio can be represented as

$$F = \frac{\frac{1}{P}\sum_{i\in P}\sqrt{|\mu_i - \bar{\mu}|^2}}{\frac{1}{P}\sum_{i\in P}\left(\frac{1}{T}\sum_{\beta\in T}\sqrt{\left|x_\beta^{(i)} - \mu_{\beta,i}\right|^2}\right)} \tag{8}$$

where $\bar{\mu}$ is the average mean for all the phonemes across all tones. $\mu_i$ is the average mean of phoneme $i$ across all tones.

$\mu_{\beta,i}$ is the mean value of phoneme $i$ for tone $\beta$. $x_{\beta}^{(i)}$ indicate the instance of a phoneme $i$ with $\beta$ tone. $P$ is the number of phonemes and $T$ is the total number of tones associated with the phoneme $i$.

The F-ratio has been calculated for all the five speech features to evaluation their inter-phoneme to intra-phoneme variation with change in associated tone of the phoneme. Table. II shows the average F-ratio value for different features under different tonal conditions.

TABLE II: Average F-ratio value for the features

| Feature Type | F-ratio |
|---|---|
| MFCC | 0.1129 |
| LFCC | 0.0812 |
| LPCC | 0.0895 |
| RC | 0.0590 |
| LAR | 0.0538 |

KL-Distance has been measured among different tonal instances of the same phoneme. Vowel /a/, which has three tonal instances level tone instance $/\bar{a}/$, raising tone instance /á/ and falling tone instance /à/. KLD of a feature vector extracted from all three instances have been measure and their average has been taken. The average of these distances has been presented in the Table given below.

TABLE III: Average KL Distance among the different tonal instances of the same phoneme

| Feature Type | Phonemes | | | | | | Average |
|---|---|---|---|---|---|---|---|
| | */a/* | */i/* | */ii/* | */u/* | */e/* | */o/* | |
| MFCC | 1.3255 | 0.5363 | 0.8796 | 2.0798 | 0.7303 | 1.4909 | **1.1737** |
| LFCC | 0.6294 | 0.3294 | 0.4507 | 1.4293 | 0.4364 | 1.3750 | **0.7750** |
| LPCC | 0.9432 | 0.5652 | 1.1641 | 1.4092 | 1.1046 | 1.9635 | **1.1916** |
| RC | 0.5422 | 0.3359 | 0.5881 | 1.0390 | 0.7417 | 1.4349 | **0.7803** |
| LAR | 0.1743 | 0.0898 | 0.1651 | 0.4582 | 0.4502 | 0.8455 | **0.3639** |

The Table III shows that all the features exhibit change in entropy with change in tone. Thus they are capable of capturing the changes due to change in tone. However, MFCC and LPCC can discrimination among the different tonal instances of the same phoneme more prominently compare to other features. Further, some of the phonemes have inherently better tone discrimination capability compare to others. It has been observed that vowel */u/* and */o/* have better tone discriminating capability compared to other vowels.

## V. Conclusion

In the present study, two different yardsticks have been used to measure the effectiveness of a feature vector in discriminating the tonal phonemes. F-ratio value is an indicator of tonal phoneme discrimination capability of a feature while considering all other variabilities. KLD have been computed for each tonal instances of a vowel. It gives an indicator of change in entropy of a feature extracted from a vowel due to change in associated tone. Since both the measures are different and complementary to each other, a feature selection algorithm, which will use the information from both the sources will provide better feature set for tonal as well as non-tonal phoneme recognition. Further, the probability density characteristic shows that all the features exhibit Gaussian distribution for all tonal instances and their peaks are distinguishable except MFCC. This observation suggests that if Gaussian based speech recognizer with MFCC feature perform poorly in recognizing tonal phonemes. However, MFCC shows significant entry change with change in associated tone of a phoneme. Further, F-ratio value of MFCC is also high. This observation suggest that MFCC captures the change in tone effectively. However, to utilize that property, non-Gaussian classifier must have to be used. This observation suggests that the perforation of a speech recognition system for tonal phoneme recognition depends on choice of feature as well as the classifier.

## References


[1] L. Deng and H. Xuedong "Challenges in adopting speech recognition." Communications of the ACM 47.1 (2004): 69-75.

[2] U. Bhattacharjee, "Recognition of the Tonal Words of Bodo Language." International Journal of Recent Technology and Engineering. Volume-1(2013).

[3] H. M. Wang, J. L. Shen, Y. J. Yang, C. Y. Tseng, and S. L. Lee, Complete Chinese dictation system research and development, Proceedings ICASSP-94, Vol. 1, pp. 59-61.

[4] C.J. Chen, H. Li, L. Shen and G.K. Fu "Recognize tone languages using pitch information on the main vowel of each syllable." Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on. Vol. 1. IEEE, 2001.

[5] C. J. Chen, R. A. Gopinath, M. D. Monkowski, M. A. Picheny, and K. Shen, "New Methods in Continuous Mandarin Speech Recognition", 5th European Conference on Speech Communication and Technology, Vol. 3, pp. 1543 - 1546, 1997.

[6] B. S. Atal, Automatic recognition of speakers from their voices, Proc. IEEE, vol. 64, pp. 460476, April 1976.

[7] G. S. Raja, and S. Dandapat, Performance of Selective Speech Feature for Speaker Identification, Institution of Engineers (India) Journal.

[8] S. B. Davis and P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, IEEE Trans. Acoust., Speech, and Signal Process., vol. 28, pp. 357366, Aug 1980.

[9] L. R. Rabiner, A. E. Rosenberg, and S. E. Levinson, Considerations in dynamic time warping algorithms for discrete word recognition, IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-26, pp. 575-586, Dec. 1978.

[10] Sumitra Shukla, S R Mahadeva Prasanna and S. Dandapat, Stressed Speech Processing: Human vs Automatic in Non-professional Speakers Scenario IEEE Proc. NCC 2011, Bangalore, India.

[11] H. Patro, G. S. Raja, and S. Dandapat, Statistical Feature Evaluation for Classification of Stressed Speech, Revised and submitted (August-2006) to International Journal of Speech Technology.

[12] V. Mitra, N. Hosung, C.Y. Espy-Wilson, E. Saltzman and L. Goldstein, "Articulatory Information for Noise Robust Speech Recognition," Audio, Speech, and Language Processing, IEEE Transactions on , vol.19, no.7, pp.1913-1924, Sept. 2011

[13] H. Wakita, Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms, IEEE Trans. Audio and Electroacoust., vol. AU-21, pp. 417-427, Dec. 1973.

[14] S. Furui, Speaker independent isolated word recognition using dynamic features of speech spectrum, IEEE Trans. Acoust., Speech, Signal Process., vol. 34, pp. 5259, 1986.

[15] P.T. Abraham, Apatani-English-Hindi Dictionary, Central Institute of Indian Language, 1987.

[16] G. S. Raja and S. Dandapat, Sinusoidal model based speaker identification, Proc. NCC-2004, vol. 1, pp. 523527, Jan.-Feb. 2004.

**A PROJECT REPORT ON**

# DESIGN AND DEVELOPMENT OF A SPEECH RECOGNIZER IN THE CONTEXT OF TONAL LANGUAGES OF ARUNACHAL PRADESH.

**Submitted to**
**UNIVERSITY GRANTS COMMISSION**
**BAHADUR SHAH ZAFAR MARG**
**NEW DELHI – 110 002**

**Submitted by**

**UTPAL BHATTACHARJEE**

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**
**RAJIV GANDHI UNIVERSITY, RONO HILLS,**
**DOIMUKH-791 112, ARUNACHAL PRADESH, INDIA**

# Table of Contents

# 1. Introduction

Automatic Speech Recognition is a computer programme that detects the linguistic information encoded in a speech signal. Speech recognition research needs inputs from diverse disciplines like Phonetics, Linguistics, Physics, Computer Science, Psychology, Pattern recognition, Communication and information, Signal processing, etc., [1]. Pioneering work in the area of automatic speech recognition was reported in the 1950s. Various researchers in 1950 attempted to explain the theory and fundamental ideas of acoustic phonetics. As a result, attempts were made for the development of the first generation of ASR systems. An ASR System at Bell Laboratories for speaker-dependent isolated digits identification was developed in 1952. [2]. At the RCA Laboratories, a system to recognize ten (10) distinct syllables in 10 monosyllabic words for a particular speaker was attempted in 1956 [3]. In 1959, Fry and Denes had tried to develop a system to recognize phonemes of four vowels and nine consonants at University College in England [4, 5]. At MIT Lincoln Laboratories in 1959, a speaker-independent system to identify 10 vowels was developed [6].

On different fundamental ideas in speech recognition, various research works had been performed in the 1960s. Nakata and Suzuki of the Radio Research Lab, Tokyo, described the development of a Japanese hardware vowel recognizer system during this period [7]. In 1962, Sakai, along with Doshita of Kyoto University developed a recognizer of hardware phonemes. In 1963, at NEC Laboratories, Nagata and co-workers developed hardware to recognize digits [8]. Three research projects that were also started in the same decade played a significant role in the development of speech recognition research. Martin and his colleagues of RCA laboratories began the first project [9]. Vintsyuk in the Soviet Union started the second project [10]. Reddy performed the third project on continuous speech recognition [11].

In the 1970s decade, ASR research has made significant progress, especially in isolated word recognition. Some of the notable works in this period were the works done by Zagoruyko & Velichko in Russia [12], Chiba &Sakoe in Japan [13] & in the United States by Itakura [14]. IBM started research work on Large Vocabulary speech recognition during that period [15, 16, 17]. With a sequence of ASR experiments, AT&T Bell Laboratories started research on speaker-independent speech recognition systems in the second half of the 1970s [18].

In the 1980s, research work on connected word recognition was started [19, 20, 21, 22]. In the research field of speech, statistical modeling methods like Hidden Markov Model (HMM) were introduced during that period [23, 24]. Another contemporary development during that

period was the sponsorship of the Defense Advanced Research Projects Agency (DARPA) for Continuous and Large Vocabulary speech recognition. Speech research laboratories at CMU [25], BBN [26], Lincoln Labs [27], SRI [28], MIT [29], and AT & T Bells Labs [30] were among the few major Research laboratories sponsored by DARPA during that period. Maximum Mutual Information (MMI) criterion was introduced in that period. The central concept of MMI training is using most of the mutual information between the sound observations and their corresponding words [31]. The neural network was introduced for speech recognition applications during the later 1980s [32, 33].

In the 1990s, several new pattern recognition approaches were introduced. During that period, the traditional pattern recognition problem was transformed into an optimization problem based on minimizing the empirical recognition error [34]. As a result, the distribution functions for the speech signal could not be accurately chosen or defined, rendering the Bayes decision theory used in the traditional pattern recognition problem inapplicable. Several learning methods have been introduced, such as discriminative training and kernel-based methods. The Minimum Classification Error (MCE) criterion was proposed for discriminative training to optimize recognizer parameters to minimise error rate [35]. The Generalized Probabilistic Descent (GPD) training algorithm was also discussed to approximate the error rate for optimization. Both the MMI and MCE outperform the maximum likelihood (ML)-based approach in speech recognition performance [36].

In the 2000s, Variational Bayesian (VB) estimation and clustering techniques were introduced, which is based on the posterior distribution of parameters [37]. Giuseppe Richardi [38] has proposed an active learning algorithm for ASR to minimize human supervision in training acoustic and language models and maximize ASR performance. In 2005, research works were carried out for improving the performance of the Large Vocabulary Continuous Speech Recognition (LVCSR) system [39]. In 2007, utilizing a database of large-scale spontaneous speech known as the "Corpus of Spontaneous Japanese (CSJ)," an investigation of the differences in acoustic features of spontaneous speech with the acoustic features of reading speech was presented [40]. Sadaoki Furui [41] had performed research works on speech recognition methods where adaptation of speech variation was implemented using many models that are trained based on clustering techniques. Rajesh M. Hegde et al. [42] proposed the group delay function (GDF) as an alternative method in 2007 to process the Fourier transform phase to extract speech features directly from the speech signals. De-Wachter et al. [43] used a straightforward template matching method to overcome the time dependencies problems in speech recognition. In the case of speech recognition, Xinwei Li et al. [44] had

presented a new optimization method termed semidefinite programming (SDP) to solve the large margin estimation (LME) problem of continuous density HMM (CDHMM). Jeih-Weih et al. [45] introduced three new temporal filtering approaches based on constrained versions of linear discriminant analysis (LDA), principal component analysis (PCA), and minimal class distance (MCD), in which the statistics of the modulation spectra of the speech features are used. In 2009, G.Zweig et al. [46] proposed a framework for LVCSR called a segmental conditional random field framework. In 2009, the Open Source Speech Recognition toolbox from RWTH Aachen University was made public [47]. Deep Belief Networks (DBNs) were proposed in 2010 for phone recognition, with only frame-level information being employed in the training phase [48]. Subspace Gaussian mixture models (GMMs) was proposed for speech recognition in 2010 [49]. The IBM Attila speech recognition toolkit was described by H.Soltau et al. [50]. In 2011, the Kaldi speech recognition toolkit, a free and open-source toolkit for speech recognition, was published by D.Povey et al. [51]. In 2011, J.F.Gemmeke et al. [52] proposed exemplar-based sparse representations to improve the noise robustness of speech recognition systems by modelling noisy speech signals with a sparse linear mixture of speech and noise exemplars. Context-Dependent Deep Neural Network (CDDNN) HMM was applied around 2011 for speech transcription [53]. Deep Neural Networks (DNNs) were considered in 2012 as an alternative to the GMM-HMM approach for acoustic modelling in speech recognition systems [54]. Around 2012, a context-dependent model was proposed for large vocabulary speech recognition (LVSR) purpose where a pre-trained DNN HMM was used to train the DNN. Convolutional Neural Network (CNN) was used for automatic speech recognition in a hybrid architecture of CNN and HMM.

Recently Automatic Speech Recognition (ASR) has been successfully integrated into 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 reasons 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 project we are trying to develop an ASR system that work efficiently for both tonal and non-tonal languages. Language portability is a major aspect for global acceptability of speech recognition system. A system developed for a particular language should be susceptible to any another language with minimal set of training data. However, it has been observed that the technology and systems which have been developed for non-tonal languages such as Indo-Aryan languages perform very poorly for the recognition of

tonal languages. A language is said to be tonal if words with same phonetic contents but different lexical tone patterns convey different meaning. Tone information is generated by excursion of the fundamental frequency. As the lexical tones do not contain any meaningful information for non-tonal languages, the feature extraction process of speech recognition systems developed for non-tonal languages discard those information. As a result, the systems which are very efficient in recognizing non-tonal languages fail to perform satisfactorily for tonal languages. A major section of world population spreading across South-East Asia, East Asia and Sub-Sahara Africa are speaker of tonal language [55]. In the last three decades, many attempts have been made for the development of speech recognition system for tonal languages. A popular method for recognizing tonal language is the two step method [56], first to recognize the based syllable by its phonetic contents. In the second step, recognize the tone of the syllable by classifying the pitch contour of that syllable using discriminating rules. Recognition of tonal syllable is a combination of the recognition of base syllable and the associated tone. The above method works well in isolated syllable speech recognition but difficult to handle continuous speech. To overcome the problems of two step method, one step method has been developed considering disyllable approach [57]. In this approach, each syllable is decomposed into two demi-syllables. The first demi-syllable contain tone independent phone information. The second demi-syllable called toneme, carries the tone information of the whole syllable. In this approach, the demi-syllable with different toneme are considered as different phonemes. This approach works fine with tonal language with small set of tones associate with the phonemes. However, for languages with large number of tones associated with each phoneme, the number of phonemes increases exponentially. As a result, the search space for the recognizer increases and the entire recognition task is slowed down. Another approach is based on the observation that pitch information of the main vowel is sufficient to determine the tone of the whole syllable [58]. The number of phonemes can be significantly reduced using this method. Furthermore, since the phonemes generated by this technique are similar to those found in Indo-Aryan languages, this approach might be used as a feature vector in a universal speech recognition system that recognizes both tonal and non-tonal languages.

## 2. Objective:

a. Develop a speech recognition database for the tonal languages of Arunachal Pradesh.
b. Characterize the acoustic-phonetic parameters of speech signal to identify their intra-phoneme and inter-phoneme discriminating capability with reference to the tonal languages.
c. Identification of features that can be used as feature vector for a universal speech recognizer that can recognize both tonal as well non-tonal speech efficiently.
d. Developing a prototype for universal speech recognition system using those feature vectors.

## 3. Material and Methods

### 3.1 Development of a Speech Database for Tonal Language of Arunachal Pradesh

A speech database for Automatic Speech Recognition research with reference to the tonal language has been developed using the Apatani language of Arunachal Pradesh. Arunachal Pradesh of North East India is one of the linguistically richest and most diverse regions in all of Asia, being home to at least thirty and possibly as many as fifty distinct languages in addition to innumerable dialects and subdialects thereof. The vast majority of languages indigenous to modern-day Arunachal Pradesh belong to the Tibeto-Burman language family. The majority of these in turn belong to a single branch of Tibeto-Burman, namely Tani. Almost all Tani languages are indigenous to central Arunachal Pradesh, while a handful of Tani languages are also spoken in Tibet. Tani languages are noticeably characterized by an overall relative uniformity, suggesting relatively recent origin and dispersal within their present-day area of concentration. Most Tani languages are mutually intelligible with at least one other Tani language, meaning that the area constitutes a dialect chain. In addition to these non-Indo-European languages, the Indo-European languages Assamese, Bengali, English, Nepali and especially Hindi are making strong inroads into Arunachal Pradesh primarily as a result of the primary education system in which classes are initially taught by immigrant teachers from Hindi-speaking parts of northern India. Because of the linguistic diversity of the region, English is the only official language recognized in the state [59, 60].

Automatic speech recognition 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 [61]. 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 [62]. 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. 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 [63]. According to C.M.Doke [64] 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 of 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 [65]. 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.

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 [64] argued that contour tones can be considered as collection of multiple level tones. Her argument was supported by other scholar like Leben [66], Goldsmith[67] and Yip [61] with suitable evidence to justify the fact.

The Apatani language of Arunachal Pradesh of North East India is belongs to the Tani group of language [55]. Tani languages constitute a distinct subgroup within Tibeto-Burman group of languages [59, 60]. The other languages of the group are Adi, Bangni, Bokar, Bori, Damu, Gaol, Hill Miri, Milang, Na, Nyishi, Tagin, Tangam and yano [69]. 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. The Apatani language has 06(six) vowels and 19 (nineteen) consonants. Post and Kanne [60] presents a list of Apatani phonemes which is given in Table -1and Table -2.

Table.1: 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 position 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 | | |

Apatani is a tone language. Apatani tones are represented in two levels – morpheme level and word level. Apatani morphemes may be specified for one of the two lexical tones – High and Low. Since the Apatani morphemes are bound and unpronounceable, these underlying tones are in principle inaudible. They are assigned high and low on the basis of their refluxes at word level. Morphologically simplex monosyllabic words with single high root are realized with high level tone and low root is realized with falling-to-low tone respectively. Most of the Apatani words are morphologically complex and mostly dimorphic and disyllabic. The complex disyllabic word has one of the following contour tones –high-level, high-to-low falling and low-

to-high rising [67]. In determining the meaning of a word, instead of precise pitch height, the overall structure of the pitch contour – level, rising and falling plays the decisive role. In the present study, we have classified the Apatani tones as – level, rising and falling. It has been observed that except the short vowel [ə], all the other vowels are associated with these three tones. Only level tone has been observed in case of vowel [ə].

The Apatani language has six different types of accents. To capture all the accents, we have visited six Apatani villages which are known for using those different types of Apatani accents. The villages visited during the study are:

I) Hari
II) Hong
III) Hija
IV) Bulla
V) Mudang
VI) Tage

A speech database is prepared for Apatani language. The database comprises 18 isolated tonal words spoken by 50 different speakers (33 male and 17 female). Apatani language has two lexical tones raising (˴) and falling (˪) [9]. The one which does not contain either of these two tones are referred to as level tone. As there is slight different in the accent of each speakers from different villages so equal number of speakers are chosen from each village. Following are the particular isolated words which are chosen for recording. The words are selected in consultation with Phonetic experts to capture all the tonal and non-tonal instances of each vowel is given in table 3.

Table. 3: Apatani Words selected for Recording

| ***Sl no.*** | ***Apatani Tonal Words*** | ***Meaning in English*** |
|---|---|---|
| 1 | */alá/* | Soup |
| 2 | */àlà/* | Hand |
| 3 | */ala/* | Coming |
| 4 | */tàpe/* | Pumpkin |
| 5 | */tape/* | Leech |
| 6 | */ámi/* | Cat |
| 7 | */àmì/* | Tail |
| 8 | */amì/* | Eye |
| 9 | */álò/* | Put to Dry |
| 10 | */àló/* | Bone |
| 11 | */àlò/* | Salt |
| 12 | */àlo/* | Day |
| 13 | */alò/* | Drop |
| 14 | */apú/* | Blossom |
| 15 | */ápu/* | Wrap Up |
| 16 | */àpu/* | Arrow |
| 17 | */müdó/* | Rain |
| 18 | */müdo/* | Doing |

To collect the recorded speech data, we prepared a room that is nearly soundproof and to further reduce the reverberation, we have layered the walls and windows with thick curtains. Each speaker is asked to utter the same word three (03) times. The recording specification for the database is given below:

Table.4: Recording specification for the Speech Recognition Database

| | |
|---|---|
| Number of Speakers | 50 (Male=33, Female=17) |
| Number of sessions | 01 |
| Number of words | 18 |
| Number of instances for same word for the same speaker | 03 |
| Data types | Speech |
| Sampling rate | 16 KHz |
| Sampling format | Mono-channel, 16 bits resolution |
| Microphones | [1]. Table Microphone (AHUJA ACM-66) connected to workstation (PC) Z440.<br>[2]. Digital Recorder (Zoom H1 Handy digital recorder). |
| Acoustic environment | Controlled environment |
| Languages | Apatani with six different accents |

We named the database as Arunachali Tonal Speech Recognition Database Version -1 (ATSRD-V.1).

### 3.2 Subjective Evaluation of the database

A human perceptual test was conducted to determine the validity of the database, precisely to determine if a specific word kept in the database to represent a particular tone conveys that tone or not. For each recorded speech file, a serial number has been assigned to hide the actual identity of the file from the listener. Each file has been played to 15 persons who belong to Apatani tribes of Arunachali Pradesh. Five (05) of them are phonetic experts and the remaining are naïve speakers of Apatani language. Only those files identified correctly by more than 60% listeners have been considered for further processing. It has been observed that some of the listeners have an inherent problem in recognizing some of the tones. Therefore,

if a listener recognizes more than 50% of a particular tone incorrectly, their response for that tone was not taken into account.

## 3.3 Analysis of Features

Feature is the compact representation of the acoustic properties manifested in the speech signal [67]. Choosing suitable features for developing any of the speech systems is a crucial design decision. The features are to be chosen to represent the required information for the functioning of the proposed system. Different speech features represents different information of the speech signal in a highly overlapping manner. Therefore, for the development of a speech based system, the features are selected experimentally in most of the cases. In some of the cases, the features are also selected using mathematical approach like principal component analysis (PCA) [68]. The speech features may be broadly classified into the following categories – (i) Excitation source features (ii) Spectral features and (iii) 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 [69]. The features extracted from LP residual signal is called excitation source features or source features. The state-of-the-art phone recognition systems are developed only with vocal tract information. However, a sound unit is produced as a result of active involvement of excitation source and vocal tract. Just the shape of the vocal tract is not sufficient enough for the characterization of a sound unit. The bilabial plosive consonants *b* and *p* are produced by the same manner and place of articulation. The different between these two sounds is coming as a result of difference in their excitation type. The consonant *b* is voiced and *p* is unvoiced. Similarly, for all the vowels, the excitation type is nearly similar. The difference between the vowel sounds are coming as a result of place and manner of articulation. Thus, we can conclude that each sound is produced as a result of unique combination of excitation source and vocal tract participation. Therefore, to characterize a sound unit, excitation source parameter as well as vocal tract parameter are necessary. The most commonly used source parameters are Reflection coefficient (RC), Log area ratio (LAR) and Arc-sin reflection coefficients (ARC)[69]. In the present study, source features have not been considered as it has more information about speaker identity than information about the spoken words.

A sound unit is characterized by a sequence of shapes assumed by the vocal tract during production of the sound [70]. 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. Formants are the resonances of the vocal tract at a given point of time characterized by bandwidth and amplitude [71]. These parameters are unique for a sound unit. 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 features are clearly visible in the frequency domain. Frequency domain analysis of the speech signal is performed by segmenting the speech signal into frame of 20-30 ms, with the frame shift of 10 ms. Most commonly used spectral features are linear predictor cepstral coefficients (LPCC), mel frequency cepstral coefficients (MFCC), perceptual linear predictor coefficients (PLPC) and their derivations [72]. In the present study, LPCC, MFCC and PLPCC features have been considered to analyze their phoneme discrimination capability with reference to tonal phonemes. Due to the state of the art performance of MFCC features, it has been considered as a de facto feature for speech recognition, specially for non-tonal languages.

Prosody represents the suprasegmental aspects of speech production. Prosody is concern with those aspects of speech signal that modulate and enhance its meaning [73]. It makes the human speech natural. It is associated with longer unit of speech such as syllable, words, phrases and sentences. Prosody is acoustically represented by duration, intonation (F0 contour) and energy [74]. Mary and Yegnnarayana [75] analyzed the effectiveness of prosodic features for speaker verification. They observed that shape of the F0 contour reflects certain speaking habits of a person. In order to represent the shape of the F0 contour, tilt parameters have been used [76]. A 7-dimensinal feature vector was proposed, which includes mean value of pitch ($F_{0_\mu}$), peak fundamental frequency ($F_{0_p}$), change of F0 (ΔF0), distance of F0 peak with respect to vowel onset point (VOP) ($D_p$), amplitude tilt ($A_t$), Duration tilt ($D_t$) and change of log energy (ΔE). Prosody plays an important role in the transcript of information in human communication. To utilize the prosodic features in speech recognition, suitable parameterization of the prosodic information is required. Normally, it is represented by fundamental frequency ($F_0$), energy and normalized duration of syllable [77]. In the present study, in order to use only frame-based features, fundamental frequency and energy have been considered for the representation of prosodic information. Fundamental frequency and frame energy are static features, calculated frame by frame. In order to include temporal information,

their first (Δ)- and second (ΔΔ)-order derivatives have been calculated and added to the feature set. Thus, we got a 6-dimensional prosodic feature vector for each frame.

In the present study, each tonal instance of a vowel has been considered as different tonal vowel. For example, the vowel /a/ has three associated tones – rising, falling and level. Thus vowel /a/ give raise to the tonal vowels $\acute{a}$ (/a/ rising), $\grave{a}$ (/a/ falling) and $\bar{a}$ (/a/ level). We referred to these vowels as tonal vowel. 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.

### 3.3.1 F-ratio

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

$$F = \frac{\text{Variance between the tonal vowels for a coefficient}}{\text{Average variance within all the tonal vowels for the coefficient}} \quad \ldots (1)$$

It can be represented as

$$F = \frac{\frac{1}{N}[\sum_{i=1}^{N}(\mu_i - \bar{\mu})^2]}{\frac{1}{N}\sum_{i=1}^{N} S_i} \quad \ldots (2)$$

where N is the number of tonal vowels, $\mu_i$ 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. $S_i$, the within tonal vowel variance is given by,

$$S_i = \frac{1}{M_i}\sum_{j=1}^{M_i}(x_{ij} - \mu_i)^2 \quad \ldots (3)$$

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 i$^{th}$ tonal vowel. Higher F-ratio value for a coefficient indicates that it can be used for good classification.

### 3.3.2 Kullback-Leibler Distance (KLD)

Kullback Leibler distance (KLD) has been used to measure the distance between the features. The KLD provides a natural distance between a probability distribution and a target

probability distribution. In the present study, KL distances have been measured among the features extracted from each tonal instance of the phoneme with other tonal instance of the same base phoneme. If the distance is high, the feature will be able to discriminate among the tonal instances. However, the same feature should also be able to discriminate among the different phonemes. To measure the inter-phoneme discrimination capability the feature, we have computed the KL distance among different base phonemes.

## 3.4 Tonal Speech Recognition System

### 3.4.1 Hidden Markov Model

The concept of HMM was developed by Baum and his colleagues. Baker at CMU and Jelinek and his colleagues at IBM in the late 1960s and early part of 1970s. The fundamental concepts of HMM was published by for speech processing applications in the 1970s [1]. It is a system which can be described by the following components and properties [23, 78]:

- It is a collection of N distinct states, $\{s_1, s_2, \dots s_N\}$. One of these states is considered as the current state of the system, at any discrete time t. At discrete time t,$q_t$ is considered as the state of the system where $q_t = s_i$ and $i$ = 1, 2… N.

- Depending upon the probabilities associated with the states, changes of states are occurred at equal spaced discrete times. The state transition probability is another component of the system which is not dependent of time. In the equation (14), the probability of change in state from $s_i$ to $s_j$ at discrete time t is represented by $a_{s_i s_j}$

$$a_{s_i s_j} = P\left(q_t = s_j \middle| q_{t-1} = s_i\right) \qquad 1 \le i, j \le N \qquad \dots (14)$$

Where

$$a_{s_i s_j} \ge 0 \qquad for\ all\ i, j \qquad \dots (15a)$$

$$\sum_{j=1}^{N} a_{s_i s_j} = 1 \qquad for\ all\ i \qquad \dots (15b)$$

- Equation (16) shows that the probability of the previous state decides the probability of a particular state.

$$P\left(q_t = s_j \middle| q_{t-1} = s_i, q_{t-2} = s_k, \cdots\right) = P\left(q_t = s_j \middle| q_{t-1} = s_i\right) \qquad \dots (16)$$

- The other components are: an initial state ($q_0$ ) and an end state ($q_F$ ).

Fig. 1 shows a four state Markov chain.

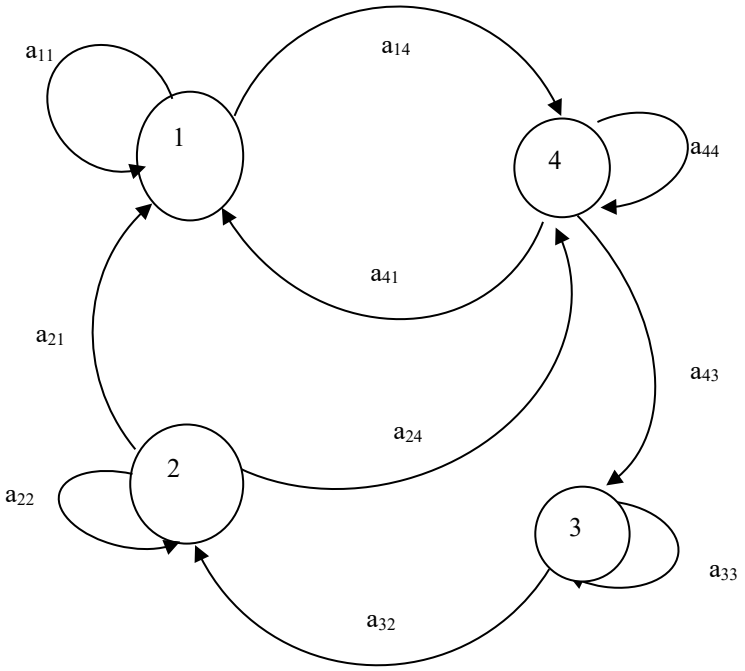


Fig.1: Selected state transitions by a four states Markov chain.

In Markov chain each state is related to an event that can be observed and so it is also known as observable Markov model. An output of a Markov process can be stated as a collection of states at a discrete time instance. When it is required to estimate probability of an output sequence consisting of observable events a Markov chain is useful. Any observation sequence not consisting of directly observable event, Markov chain is not applicable but we need an extension of Markov chain which is specifically known by the name Hidden Markov Model (HMM). HMM allows observing the hidden stochastic process through another collection of stochastic processes which produce the observation sequence.

HMM can be described by the following properties and elements:

- A HMM consist of a collection of N number of states $\{s_1, s_2, \ldots s_N\}$ .At time t, $q_t$ is the state.
- The type of the HMM is responsible for the possible state transitions from one state to another state.
- As mentioned in equation (16), as per Markov chain property, probability of each consecutive state is dependent only on the probability of the previous state. This property is one of the most important elements of HMM.

- Below mentioned equation (17a) represents the state transition probabilities in HMM.

$$A = a_{s_i s_j}\ ,\ a_{s_i s_j} = P\left(q_{t+1} = s_j \middle| q_t = s_i\right)\ , \qquad 1 \le i,j \le N \qquad \ldots\ (17a)$$

$$\sum_{j=1}^{N} a_{s_i s_j} = 1 \qquad for\ all\ i \qquad \ldots\ (17b)$$

When each state can be reached from any other state in a single step as shown in Fig. 2, all $a_{s_i s_j} > 0$.

Otherwise, for one or more than one values of$(i, j)$ as shown in Fig. 3, some $a_{s_i s_j} = 0$.

Each hidden state in a HMM can randomly generates one of the M observation symbols available as collection of M number of observation symbols. The set of observation symbols are denoted as V=$\{v_1, v_2, \ldots, v_M\}$. Probabilities of the sequence of observation symbol are another important element of HMM. It can be stated as the probability of an observation $o_t$ at time t generated from a state $s_i$. In the following equation (18), observation symbol probabilities are denoted by B is given by:

$$B = \left(b_{s_i}(o_t)\right),\ b_{s_i}(o_t) = P(o_t = v_k | q_t = s_i)\ , \qquad \ldots\ (18)$$

$$1 \le k \le M \quad and \quad 1 \le i \le N$$

HMM also has an initial state probability distribution which is denoted by $\pi$ in equation (19) where $\pi_{s_i}$ gives the probability of the Markov chain to start in state $s_i$ and if $\pi_{s_j} = 0$, it means the Markov chain will not start in state $s_j$.

$$\pi = \left(\pi_{s_i}\right)\ ,$$
$$\pi_{s_i} = P(q_1 = s_i)\ , \quad 1 \le i \le N \qquad \ldots\ (19)$$

HMM is used as a generator of the observation sequence, $O = o_1, o_2, \ldots, o_T$ with N and M. Here $o_t$ is one of the observation symbols available in V. T is the total number of observation symbols present in the observation sequence O. The HMM model is represented by $\lambda = (A, B, \pi)$.

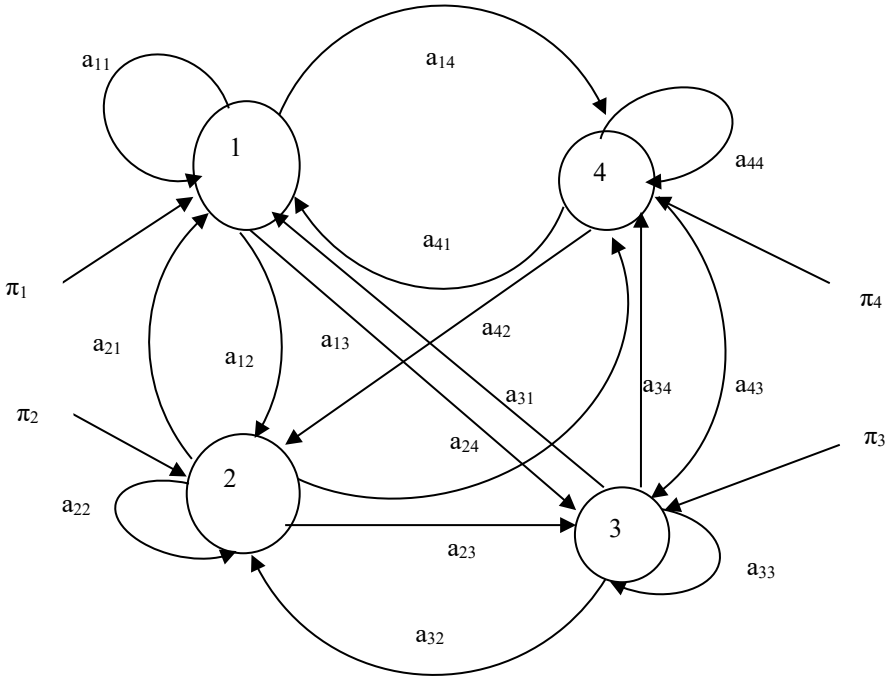


Fig. 2: A 4-state HMM where each state can be reached from any other state in a single step.

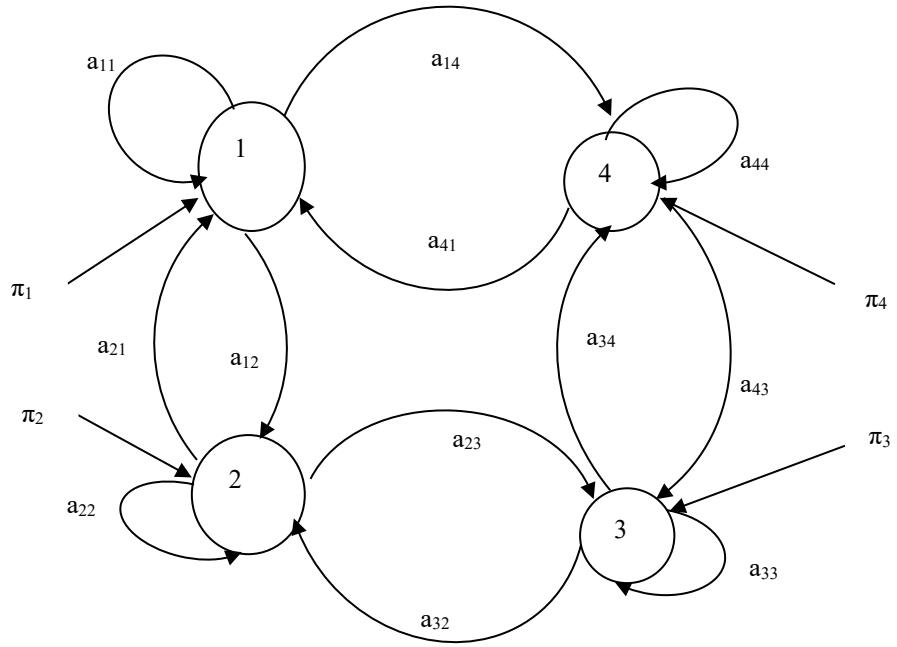


Fig. 3: A 4-state HMM where every state cannot be reached from any other state in a single step.

There are three main issues related to the successful and efficient implementation of HMM in

case of real world applications like speech recognition based on HMM which are stated below:

- First issue: Calculating efficiently the probability to be generated by HMM $\lambda = (A, B, \pi)$ for the observation sequence $O = o_1, o_2, \dots, o_T$ .
- Second issue: Estimating the most likely sequence of hidden states $s_i$ that produced the observation sequence $O = o_1, o_2, \dots, o_T$ with the HMM $\lambda = (A, B, \pi)$.
- Third issue: Determining the HMM parameters $\lambda = (A, B, \pi)$ for some of the training observation sequence, $O = o_1, o_2, \dots, o_T$ by the general structure of HMM hidden and visible states so that the HMM will maximize the probability of the observation sequence.

The above mentioned issues can be solved in the following ways:

- The first issue can be solved using the forward procedure.
- The second issue can be solved by using Viterbi Algorithm which can be used to estimate the single best state sequence that produced the observation sequence with a HMM
- The third issue can be solved by using Baum-Welch method which is an efficient technique to adjust the HMM parameters (A, B, $\pi$) with training observation sequence and the structure of HMM hidden and visible states so that it will maximize the probability of the observation sequence.

### *3.4.2 Forward Algorithm for Probability Estimation*

In HMM, the probability denoted as $P(O|\lambda)$ has to be estimated efficiently where $O = o_1, o_2, \dots, o_T$ is the observation sequence and $\lambda = (A, B, \pi)$ where A denote the state transition probabilities as in equation (17a), B denote the observation probabilities as in equation (18) and $\pi$ denote the initial probabilities (19).

$P(O|\lambda)$can be estimated using the equation (20), considering the state sequence of length T for the observation sequence $O = o_1, o_2, \dots, o_T$ as q= $q_1, q_2, \dots, q_T$.

$$P(O|\lambda) = \sum_{q_1, q_2, \dots\dots, q_T} \pi_{q_1} b_{q_1}(o_1)\, a_{q_1 q_2} b_{q_2}(o_2) \dots . a_{q_{T-1} q_T} b_{q_T}(o_T) \quad \dots (20)$$

This approach cannot be used in practical applications because it requires a large amount of

computations. To obtain the required value of $P(O|\lambda)$ it requires $2\times T\times N^T$ number of calculations with N numbers of HMM states and the observation sequence with length T because for each observation of the observation sequence, N number of possible states are available that means total number of possible state sequence is $N^T$ and for each state sequence require $2\times T$ calculations.

Alternatively, we can use the forward algorithm to estimate $P(O|\lambda)$ efficiently as explained below:

This algorithm considers a forward variable $F_t(s_i)$ which is for the current state $s_i$ at time t, the probability of the observation sequence $o_1 o_2 o_3 \ldots . o_t$ as stated in the equation (21).

$$F_t(s_i) = P(o_1 o_2 o_3 \ldots . o_t, q_t = s_i|\lambda) \quad \ldots (21)$$

Now estimation of $F_t(s_i)$ is performed inductively as mentioned in the following three steps.

Step 1: At first the forward variable is initialized as shown in the equation (22) with joint probabilities of state $s_i$ and initial observation $o_1$.

$$F_1(s_i) = \pi_{s_i} b_{s_i}(o_1) \qquad 1 \leq \mathrm{i} \leq N \quad \ldots (22)$$

Step 2: The second step estimates inductively the probability of the observation sequence $o_1 o_2 o_3 \ldots . o_t, o_{t+1}$ with state transition from $s_i$ to $s_j$ as shown in equation (23)[91]. This step is the most important step in this approach.

$$F_{t+1}(s_j) = \left[\sum_{i=1}^{N} F_t(s_i) a_{s_i s_j}\right] b_{s_j}(o_{t+1}) \qquad 1 \leq t \leq T-1 \quad \ldots (23)$$

$$1 \leq j \leq N$$

Step 3: The third step estimates $P(O|\lambda)$ as shown in equation (24)[1] by summing the estimated terminal forward variables denoted by $F_T(s_i)$, which was obtained in the second step.

$$P(O|\lambda) = \sum_{i=1}^{N} F_T(s_i) \quad \ldots (24)$$

### *3.4.3 Backward Algorithm for Estimation of Probability for Partial Observation Sequence*

Backward algorithm is another way to estimates the probability for partial observation sequence, $o_{t+1}, o_{t+2}, \ldots . o_T$ from the observation sequence $0 = o_1, o_2, \ldots, o_T$ . This Backward calculation is used to solve the following HMM issues:

1) To estimate HMM parameter in an optimal way.

2) To estimate the best state sequence for an observation sequence.

Now the backward variable is computed for the Backward procedure. As shown in equation (25)[70], backward variable is the probability of the partial observation sequence, $o_{t+1}, o_{t+2}, \ldots . o_T$ from the observation sequence $0 = o_1, o_2, \ldots, o_T$.

$$Bk_t(s_i) = P(o_{t+1}o_{t+2} \ldots . o_T | q_t = s_i, \lambda) \quad \ldots (25)$$

At first the backward variable $Bk_T(s_i)$is initialized for all possible N states of the HMM as shown in the equation (26)

$$Bk_T(s_i) = 1, \qquad 1 \leq i \leq N \quad \ldots (26)$$

The second step estimates $Bk_t(s_i)$ inductively for all N states of the HMM as shown in equation (27) [91]. This step is the most important step in this approach.

$$Bk_t(s_i) = \sum_{j=1}^{N} a_{s_i s_j} b_{s_j}(o_{t+1})\, Bk_{t+1}(s_j), \quad t = T-1, T-2, \ldots, 1, \quad 1 \leq i \leq N \quad \ldots (27)$$

### *3.4.4 Baum-Welch Method for HMM Parameter Estimation*

One of the challenges of HMM is optimal estimation of the HMM parameters, $\lambda = (A, B, \pi)$ so that it will maximizes the value of $P(O|\lambda)$[1]. To overcome this challenge, Baum and his colleagues proposed a method known as Baum-Welch method which is stated below. As shown in equation (28), $\xi_t(i, j)$ is the probability of an observation sequence where state $s_i$ is reached at time t and state $s_j$ is reached at time t+1.

$$\xi_t(i, j) = P(q_t = s_i, q_{t+1} = s_j | O, \lambda) \quad \ldots (28)$$

$\xi_t(i, j)$can be represented as shown in the equation (29)[78] taking into consideration the definitions of forward equation (21) and backward equation (25).

$$\xi_t(i, j) = \frac{P(q_t = s_i, q_{t+1} = s_j, O|\lambda)}{P(O|\lambda)} = \frac{F_t(s_i) a_{s_i s_j} b_{s_j}(o_{t+1}) Bk_{t+1}(s_j)}{\sum_{i=1}^{N}\sum_{j=1}^{N} F_t(s_i) a_{s_i s_j} b_{s_j}(o_{t+1}) Bk_{t+1}(s_j)} \quad \ldots (29)$$

Next we need to compute the probability of reaching state $s_i$ at time t and is represented by $\gamma_t(i)$ as shown in equation (30) [70].

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i, j) \quad \ldots (30)$$

The expected number of transitions from state $s_i$ and the expected number of transitions from

state $s_i$ to $s_j$ considering a observation sequence $O = o_1, o_2, \ldots, o_T$ , can be estimated as given in equation (31) and (32)[70].

$$\sum_{t=1}^{T-1} \gamma_t(i) = expected\ numbers\ of\ transition\ from\ state\ s_i\ in\ O \quad \ldots(31)$$

$$\sum_{t=1}^{T-1} \xi_t(i,j) = expected\ numbers\ of\ transition\ from\ state\ s_i\ to\ state\ s_j\ in\ O \quad \ldots(32)$$

The re-estimation of the HMM parameters, $\lambda = (A, B, \pi)$ are processed by equation (33), (34) and (35)[70], using equation (30) and (31) as shown below.

$$\hat{A} = \hat{a}_{s_i s_j} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \quad \ldots(33)$$

$$\hat{B} = \hat{b}_{s_j}(o_t) = \frac{\sum_{\substack{t=1 \\ o_t = v_k}}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)} \quad \ldots(34)$$

$$\hat{\pi} = \hat{\pi}_{s_i} = \gamma_1(i) \quad \ldots(35)$$

If $P(O|\hat{\lambda}) > P(O|\lambda)$, it means re-estimated HMM parameters, $\hat{\lambda} = (\hat{A}, \hat{B}, \hat{\pi})$ provides better probability scores for observation sequence $O = o_1, o_2, \ldots, o_T$ then $\hat{\lambda}$will be considered as $\lambda$ for further iterations of the HMM parameter adjustment process to maximize the value of $P(O|\lambda)$. Until some limiting condition is achieved, this iterative process continues [1].

### *3.4.5 The Viterbi Algorithm for Single Best State Sequence Estimation*

One of the main issues in the testing phase of ASR system while using HMM is to find out the best state sequence for an observation sequence. This issue has several solution, one of which is the Viterbi algorithm where optimality criteria used is to find out the single best state sequence. This is done to maximize $P(q|O, \lambda)$ where q is the single best state sequence for the observation sequence O with HMM $\lambda$[70]. The following describes the Viterbi algorithm.

The first step in the Viterbi algorithm is to find the best probability score along a single path with first t observations of an observation sequence $O = o_1, o_2, \ldots, o_T$ and state $s_i$ at time t as shown in the equation(36)[1].

$$\delta_t(i) = \max_{q_1, q_2, \ldots, q_{t-1}} P(q_1 q_2 \ldots q_{t-1}, q_t = s_i, o_1 o_2 \ldots o_t | \lambda) \quad \ldots(36)$$

Next step is to define $\delta_{t+1}(j)$ by induction as given in the equation (37) [70].

$$\delta_{t+1}(j) = \left[\max_i \delta_t(i) a_{s_i s_j}\right] . b_{s_j}(o_{t+1}) \quad \ldots(37)$$

To obtain the required single best state sequence a tracking process for the parameter which

maximizes equation (37) for each t and j is performed using an array$\varphi_t(j)$. Viterbi algorithm requires the following steps for implementation

Step 1: First step is to define $\hat{\pi}_{s_i}$ , $\hat{b}_{s_i}(o_t)$ and $\hat{a}_{s_i s_j}$as given in the equation (38), (39) and (40)[91].

$$\hat{\pi}_{s_i} = log(\pi_{s_i}) \qquad 1 \le \mathrm{i} \le N \qquad \ldots(38)$$

$$\hat{b}_{s_i}(o_t) = log[b_{s_i}(o_t)] \qquad 1 \le i \le N, 1 \le t \le T \qquad \ldots(39)$$

$$\hat{a}_{s_i s_j} = log\left(a_{s_i s_j}\right) \qquad 1 \le i, j \le N \qquad \ldots(40)$$

Step 2: Second step initializes $\hat{\delta}_1(i)$and $\varphi_1(i)$ as given in equation (41), (42) and (43)[91].

$$\delta_1(i) = \pi_{s_i} b_{s_i}(o_1), \quad 1 \le i \le N \qquad \ldots(41)$$

$$\hat{\delta}_1(i) = log(\delta_1(i)) = \hat{\pi}_{s_i} + \hat{b}_{s_i}(o_t) \qquad 1 \le i \le N \qquad \ldots(42)$$

$$\varphi_1(i) = 0 \qquad 1 \le i \le N \qquad \ldots(43)$$

Step 3: Third step implements a recursive process using equation (44) and array $\varphi_t(j)$ is used as shown in equation (45) to track of the argument that maximized the equation (44)[91].

$$\hat{\delta}_t(j) = log(\delta_t(j)) = \max_{1 \le i \le N}\left[\hat{\delta}_{t-1}(i) + \hat{a}_{s_i s_j}\right] + \hat{b}_{s_j}(o_t) \qquad \ldots(44)$$

$$\varphi_t(j) = \arg\max_{1 \le i \le N}\left[\hat{\delta}_{t-1}(i) + \hat{a}_{s_i s_j}\right] \quad 2 \le t \le T\,, 1 \le j \le N \qquad \ldots(45)$$

Step 4: Forth step implements termination process of the Viterbi algorithm as given in equation (46) and (47) [70].

$$\hat{P}^* = \max_{1 \le i \le N}[\hat{\delta}_T(i)] \qquad \ldots(46)$$

$$q_T^* = \arg\max_{1 \le i \le N}[\hat{\delta}_T(i)] \qquad \ldots(47)$$

Step 5: Fifth step implements a backtracking process as given in equation (48). Backtracking process is used to find out the single best state sequence where $q_t^*$ is the most suitable state at time t to maximize $P(q|O,\lambda)$[70].

$$q_t^* = \varphi_{t+1}(q_{t+1}^*) \quad t = T-1, T-2, \ldots\ldots, 1 \qquad \ldots(48)$$

This implementation of Viterbi algorithm requires about $N^2 \times T$ number of calculations [91].

# 4. Analysis of Features for their Tonal Speech Representation Capability

## 4.1 Introduction

This chapter presents an analysis of the speech features for their tonal speech discrimination capability. Speech features have been evaluated for their tonal base unit (TBU)

discrimination ability even when the TBUs are distinct from each other only by tone or only by base syllabic unit. A detail profiling of the features has been done. Further, different combinations of features have been tried to find the suitable combination for the representation of the tonal base units. A feature extraction method with varying observation window has been proposed for tonal speech recognition and its performance has been evaluated. All the experiments reported in this chapter are carried out using the tonal vowel database.

## 4.2 Tonal vowel database

The Apatani language of Arunachal Pradesh has six vowels and two lexical tones – rising and falling. Beside from these two tones each Apatani vowel has an instance without any associated tone, which we called level tone. Except the vowel [ə], all the vowels have all these three tonal instances. For vowel [ə] only level tone has been observed. 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. 5. Since the vowel [ə] has only one tone, it is not taken into consideration while evaluating the performance of the feature vectors.

Table. 5. 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 |

The vowel database consist of 750 instance for each tonal vowel spoken by 50 speakers in three sessions and each speaker utter the same word 5 times in each session.

## 4.3 Evaluation of the Features for tonal vowel discrimination capability

In this section we have evaluated the features for their tonal vowel discrimination capability. The features are evaluated on the basis of their tonal vowel discrimination capability when

a) Base vowels are same only the tones are different.
b) Tones are same only the base vowels are different.
c) Both tones and base vowels are different.

### 4.3.1 Statistical Evaluation of the Features

The speech signals have been analyzed using a Hamming windows of length 20 ms, frame rate 100 Hz and pre-emphasis factor of 0.97. From each windowed speech signal MFCC, LPCC, RC, LAR and prosodic features have been extracted. Features are extracted from each tonal instance of the vowel separately. Kullback-Leibler distances (KLD) have been computed from each tonal vowel to all the other tonal vowels and their average has been taken. KLD is defined as the relative entropy of two density functions. Higher the value of the KLD indicates that the feature is capable for discriminating among the tonal vowels. The results of the experiments are given in the table-6 (a) and table-6 (b).

Table-6 (a): Table: Average KL Distance from each tonal vowel to all the other vowels (Vowel ɑ:,ɪ and ɔ)

| Feature Type | Vowels | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | [ a̅:] | [ á:] | [ à:] | [ ɪ̄] | [ ɪ́] | [ ɪ̀] | [ ɔ̄ ] | [ ɔ́] | [ ɔ̀] |
| MFCC | 1.3961 | 0.3375 | 0.9665 | 0.7851 | 0.1114 | 0.3104 | 0.3871 | 0.6731 | 0.7513 |
| LPCC | 0.1307 | 0.2220 | 0.5271 | 0.7370 | 0.8135 | 0.6929 | 0.7506 | 0.3960 | 0.0477 |
| RC | 1.1063 | 0.0665 | 0.9483 | 0.3955 | 0.3622 | 0.2413 | 0.0231 | 0.3613 | 0.2004 |
| LAR | 0.1170 | 0.2492 | 0.2749 | 0.2640 | 0.0060 | 0.1262 | 0.1979 | 0.1517 | 0.0505 |
| PROSODIC | 0.0910 | 0.0889 | 0.0658 | 0.0194 | 0.1134 | 0.0038 | 0.1517 | 0.1262 | 0.0488 |

Table-6 (b): Table: Average KL Distance from each tonal vowel to all the other vowels (Vowel ɛ and ʊ)

| Feature Type | Vowels | | | | | |
|---|---|---|---|---|---|---|
| | [ ɛ̄ ] | [ ɛ́] | [ ɛ̀] | [ ʊ̄ ] | [ ʊ́] | [ ʊ̀] |
| MFCC | 1.1161 | 0.9675 | 0.5907 | 0.9076 | 0.6625 | 0.5145 |
| LPCC | 0.3223 | 0.1451 | 0.3293 | 0.2938 | 0.1106 | 0.6640 |
| RC | 0.0720 | 0.2041 | 0.1950 | 0.0737 | 0.0933 | 1.5474 |
| LAR | 0.0762 | 0.0651 | 0.0779 | 0.1828 | 0.1859 | 0.0461 |
| PROSODIC | 0.0762 | 0.1365 | 0.1553 | 0.1258 | 0.1227 | 0.1979 |

The above result shows that all the features, except the prosodic feature exhibit change in entropy due to the change in tone and base-phoneme together. However, significant changes have been recorded only in the case of spectral features MFCC and LPCC only. The average changes in entry for MFCC and LPCC are 0.6985 and 0.4122, which is still significantly low for distinguishing the tonal vowels. The change in entry may be for two reasons – change in base phoneme and change in tone. In the next experiment we have analysed the change in

entropy only due to the change in tone keeping the base phoneme fixed.

Table-7: Average KL Distance among the different tonal instances of the same vowel

| Feature Type | Vowel | | | | |
|---|---|---|---|---|---|
| | ɑ | ɪ | ɔ | ɛ | ʊ |
| MFCC | 0.3884 | 0.2639 | 0.3319 | 0.2146 | 0.1468 |
| LPCC | 0.2068 | 0.1638 | 0.1100 | 0.1056 | 0.1075 |
| RC | 0.0744 | 0.0997 | 0.1818 | 0.0365 | 0.0546 |
| LAR | 0.2469 | 0.0950 | 0.2180 | 0.0245 | 0.0409 |
| PROSODIC | 0.5210 | 0.4610 | 0.7980 | 0.9560 | 0.9100 |

The above result shows that only Prosodic features can discriminate among the different tonal instances of the same phoneme prominently compared to other features. The above result shows that MFCC and LPCC have vowel discrimination capability when the tone is same but the base-phonemes are different and prosodic and RC features shows more vowel discrimination capability when the phonemes are distinct from each other due to tone only. This observation suggest that in tonal vowel recognition problem, the spectral feature can plays an important role in discriminating the base syllable whereas RC and Prosodic features can play important role in discriminating among the tones.

In the next experiment, we have evaluated the F-ratio value for the features. F-ratio gives a measure for intra-class variability to inter-class variability. Class boundaries are appropriately selected to evaluate the features for their phoneme discrimination capability when:

- The base-phoneme of the tonal vowels is same but tone is different.
- The tones of the tonal vowels are same but base phonemes are different.

The F-ratio values have been computed for all the tonal vowels considering the above mentioned class boundaries and their averages have been taken. Higher the F-ratio value indicates that the features are more suitable for discriminating among the phonemes under the test boundary conditions. The results of the experiments are listed in table-8(a) and table-8(b).

Table – 8 (a): Average F-ratio value when the tone is same and base phonemes are different

| Feature | F-ratio value |
|---|---|
| MFCC | 0.3251 |
| LPCC | 0.5323 |
| RC | 1.1333 |
| LAR | 0.3986 |
| PROSODIC | 1.0969 |

Table – 8 (b): Average F-ratio value when the base-phonemes are same and tones are different

| Feature | F-ratio value |
|---|---|
| MFCC | 2.6362 |
| LPCC | 2.1789 |
| RC | 0.9058 |
| LAR | 0.7849 |
| PROSODIC | 0.5760 |

The above result shows that MFCC and LPCC have vowel discrimination capability when the tone is same but the base-phonemes are different and prosodic and RC features shows more vowel discrimination capability when the phonemes are distinct from each other due to tone only. This observation suggest that in tonal vowel recognition problem, the spectral feature can plays an important role in discriminating the base syllable whereas RC and Prosodic features can play important role in discriminating among the tones.

Kolmogorov-Smirnov (KS) test has been conducted to find the maximum distance between two cumulative distributions due the change in tone when the base-phonemes are same, change in base phonemes when the tones are same and change in tone and base-phoneme both. Feature vectors are extracted from different tonal instances of the vowels and their probability distribution functions are converted into a cumulative distribution function (CDF). The maximum difference between the CDFs serves as the test statistics. The results of the experiment are listed below

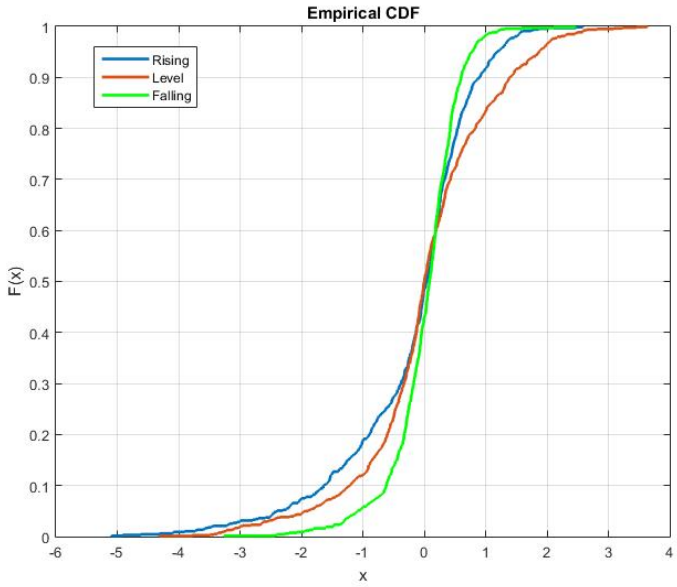


Fig. 4: The CDF of the MFCC Features

Table-9. Average maximum distance among the CDFs for each feature type under different variability conditions

| Feature | Average maximum distance between the CDFs when | | |
|---|---|---|---|
| | **Tones are different base-phonemes are same** | **Tones are same base-phonemes are different** | **Both tone and base-phoneme are different** |
| **MFCC** | 0.2705 | 0.7226 | 0.4392 |
| **LPCC** | 0.2815 | 0.7011 | 0.5904 |
| **RC** | 0.3493 | 0.4949 | 0.3646 |
| **LAR** | 0.0150 | 0.3616 | 0.3957 |
| **PROSODIC** | 0.4723 | 0.0757 | 0.1589 |

The above analysis shows that the spectral features MFCC and LPCC are better in discriminating among the tonal vowels when the base-phonemes are different. However, they fail to distinguish the vowels when the base phonemes are same and only tones are different. Prosodic features are found to be good in discrimination of the phonemes when the base phonemes are same but tones are different.

From the above three experiments, it has been observed that MFCC and LPCC can discriminate among the tonal vowels when the tone is same but the base phonemes are different. Prosodic features are suitable for discriminating among the tones when the base phonemes are same. However, it fails to distinguish among the vowels when the base-phonemes are different. Among the source features, RC shows moderate tone discrimination capability. This observation suggests that combination of the spectral features with prosodic and RC features can improve the tonal vowel recognition accuracy.

### 4.3.2 Model-base Evaluation of the Speech Features

To evaluate the efficiency of the feature set in recognizing the tonal vowels, a Hidden Markov Model based recognizer has been trained. 50% tonal instances of each vowel have been used for training and the remaining 50% for testing the system. The number of HMM states is determined empirically. In the present model, 6 (six) states have been used. The performance of each feature have been evaluated in terms of recognition accuracy, which is the percentage of times the recognizer has been able to recognize the tonal vowel correctly. The error cases have been further in-depth investigated to get an insight into the confusion created at modelling level. Table-10 presents an analysis of the performance of HMM based tonal vowel recognizer.

Table. 10 Performance of the HMM based recognizer for recognizing the tonal vowels training and testing with different feature set

| | Recognition accuracy for training and testing Features (in %) | | | | |
|---|---|---|---|---|---|
| | MFCC | LPCC | RC | LAR | Prosodic Features |
| Correctly recognized the tonal vowel | 59.23 | 54.23 | 40.18 | 34.63 | 23.78 |
| Incorrectly recognized as a tonal vowel with same base phoneme but different tone | 26.46 | 21.45 | 10.71 | 27.56 | 9.78 |
| Incorrectly recognized as a tonal vowel with same tone but different base phoneme | 8.91 | 13.11 | 37.81 | 21.48 | 55.81 |
| Incorrectly recognized as a tonal vowel with different tone and different base phoneme | 5.4 | 11.21 | 11.3 | 16.33 | 10.63 |

The above result shows that MFCC and LPCC have more base-phoneme discrimination capability than tone discrimination capability. MFCC features correctly recognized the tonal vowels 59.23% cases where as in 85.69% cases it recognized the base-phoneme correctly ignoring the tone. The LPCC on the other hand correctly recognized the tonal vowels 54.23% cases where as 75.68% cases it recognized the base-phoneme correctly. RC and Prosodic features on the other hand recognized the vowels correctly in 40.18% and 23.78% cases correctly where as they recognized the tones correctly in 77.99% 79.59% cases. In case of LAR feature no significant biasness towards base-phoneme or tone has been observed.

## 4.4 Feature combination for tonal speech recognition

From the statistical analysis and model-based evaluation it has been observed that MFCC and LPCC features play significant role in identifying the base-phoneme of the tonal vowels whereas Prosodic features and RC features play important role in identifying the tone associated with the base syllable. This observation suggests the need for combination of these two categories of features for tonal vowel recognition. In this section we have analysed the effectiveness of combined evidences from different feature sets for tonal speech recognition. The following combinations are tested for their effectiveness in tonal speech recognition:

a) MFCC and Prosodic features
b) MFCC and RC features
c) LPCC and Prosodic Features
d) LPCC and RC features

The features are extracted separately from each frame and extracted features are frame-wise concatenated to obtain the combined feature set. The performances of the features are evaluated using the same evaluation framework.

Table-11: Table: Average KL Distance from each tonal vowel to all the other vowels for combined features (Vowel ɑ:,ɪ and ɔ)

| Feature Type | Vowels | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | [ a̅:] | [ á:] | [ à:] | [ ɪ̄] | [ ɪ́] | [ ɪ̀] | [ ɔ̄ ] | [ ɔ́] | [ ɔ̀] |
| MFCC + Prosodic Features | 1.2789 | 0.3667 | 0.8878 | 0.6919 | 0.1933 | 0.2702 | 0.4634 | 0.6874 | 0.6881 |
| MFCC + RC | 1.8017 | 0.2909 | 1.3787 | 0.8500 | 0.3410 | 0.3972 | 0.2953 | 0.7448 | 0.6852 |
| LPCC + Prosodic Features | 0.1441 | 0.2021 | 0.3854 | 0.4917 | 0.6025 | 0.4529 | 0.5865 | 0.3394 | 0.0627 |
| LPCC + RC | 0.7546 | 0.1760 | 0.9000 | 0.6908 | 0.7172 | 0.5699 | 0.4720 | 0.4620 | 0.1513 |

Table-12 (b): Table: Average KL Distance from each tonal vowel to all the other vowels for combined features (Vowel ɛ and ʊ)

| Feature Type | Vowels | | | | | |
|---|---|---|---|---|---|---|
| | [ ɛ̄ ] | [ ɛ́] | [ ɛ̀] | [ ʊ̄ ] | [ ʊ́] | [ ʊ̀] |
| MFCC + Prosodic Features | 0.9538 | 0.8832 | 0.5968 | 0.8267 | 0.6282 | 0.5699 |
| MFCC + RC | 0.7129 | 0.7030 | 0.4714 | 0.5888 | 0.4535 | 1.2371 |
| LPCC + Prosodic Features | 0.2790 | 0.1971 | 0.3392 | 0.2937 | 0.1633 | 0.6033 |
| LPCC + RC | 0.2642 | 0.2340 | 0.3513 | 0.2462 | 0.1366 | 1.4816 |

From the above results it has been observed that no significant changes in entropy have been observed as a result of combining the features together.

In the next set of experiments, we have evaluated the intra-class to inter-class variability of the combined features under same class boundary conditions as above. The results of the experiments are given in the table-13(a) and 13(b).

Table – 13 (a): Average F-ratio value when the tone is same and base phonemes are different for combined features

| Feature | F-ratio value |
|---|---|
| MFCC + Prosodic Features | 1.2798 |
| MFCC + RC | 1.2396 |
| LPCC + Prosodic Features | 1.3034 |
| LPCC + RC | 1.2825 |

Table – 13 (b): Average F-ratio value when the base-phonemes are same and tones are different for combined features

| Feature | F-ratio value |
|---|---|
| MFCC + Prosodic Features | 2.2485 |
| MFCC + RC | 2.1252 |
| LPCC + Prosodic Features | 1.7907 |
| LPCC + RC | 1.8508 |

It has been observed that as a result of combining the features, the F-ratio values increases significantly. The combined features are equally good in discriminating among the tonal vowels when the base-phonemes are same and only tone is different as well as when the tones are same and only base phonemes are different. The combined features are evaluated for their average maximum distance among the CDFs obtained from the probability distribution of the features. The results are listed in table-14.

Table-14. Average maximum distance among the CDFs for each combined feature type under different variability conditions

| **Feature** | **Average maximum distance between the CDFs when** | | |
|---|---|---|---|
| | **Tones are different base-phonemes are same** | **Tones are same base-phonemes are different** | **Both tone and base-phoneme are different** |
| **MFCC + Prosodic Features** | 0.2894 | 0.6142 | 0.4831 |
| **MFCC + RC** | 0.2498 | 0.7732 | 0.3777 |
| **LPCC + Prosodic Features** | 0.3012 | 0.5959 | 0.6494 |
| **LPCC + RC** | 0.2599 | 0.7502 | 0.5077 |

It has been observed that due to the combination of the features, no significant changes in the distances among the CDFs obtained from the probability distribution function of the features have been observed.

In the next set of experiments, we have evaluated the combined features for their tonal vowel recognition accuracy using HMM based recognizer. The recognizer is trained with the combined feature vectors. 50% instances of each tonal vowel have been considered for training the model and the remaining have been used for testing the system. The result of the experiment is given in table 15.

Table 15: Performance of the HMM based recognizer for recognizing the tonal vowels training and testing with different combined feature sets

| | Recognition accuracy for training and testing Features (in %) | | | |
|---|---|---|---|---|
| | MFCC + Prosodic Features | MFCC + RC | LPCC + Prosodic Features | LPCC + RC |
| Correctly recognized the tonal vowel | 65.75 | 57.45 | 58.57 | 42.84 |
| Incorrectly recognized as a tonal vowel with same base phoneme but different tone | 23.88 | 22.56 | 19.31 | 23.88 |
| Incorrectly recognized as a tonal vowel with same tone but different base phoneme | 7.13 | 10.04 | 11.90 | 12.73 |
| Incorrectly recognized as a tonal vowel with different tone and different base phoneme | 3.24 | 9.95 | 10.22 | 20.55 |

The above results show that when MFCC and LPCC features are combined with Prosodic features, there is a slight enhancement in the recognition accuracy of the HMM based

system. However, when the spectral features are combined with RC features no such performance enhancement has been observed. Therefore, Spectral features combined with prosodic features may be considered as a viable option for speech parameterization for tonal speech recognition.

## 4.5 Variable Length Feature Combination

It has been observed that MFC and LPCC are better in discriminating among the tonal vowels when the base-phonemes are different. However, they fail to distinguish the vowels when the base phonemes are same and only tones are different. Prosodic features are found to be good in discrimination the phonemes when the base phonemes are same but tones are different. This observation suggests that combination of features can give a better parameterization of the speech signal for the tonal vowel recognition. However, in the frame-based feature combination experiments, it was found that the performance of the system increase only slightly due to the combined effect of spectral and prosodic features. The spectral features are short-term feature which can capture the variability of the speech signal with high resolution only in short-observation window. The prosodic features on the other hand are supra-segmental features. It can capture the variability in the speech signal when the observation window is long. Therefore, when we extract frame-based prosodic features, the acoustical properties which are visible only in long observation window are lost. To overcome this problem, we have proposed a feature combination technique where features from varying observation windows are combined together to generate a single feature set. The block diagram of the feature extraction method is given in the Fig. 5. Here we have proposed a method where the features are extracted with different observation windows and then combined together to take a decision on class boundary of the TBU.

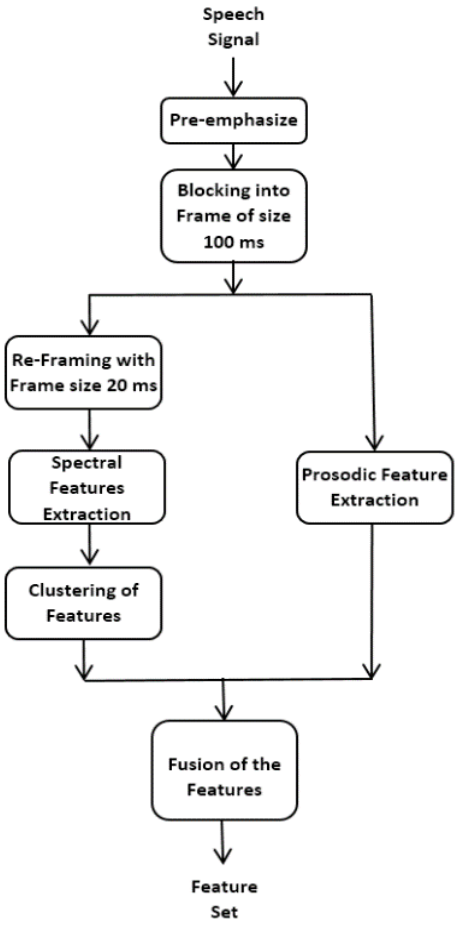


Fig. 5: Block diagram of the hybrid feature extraction system

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 preserves the temporal information of the speech feature. We are calling it temporal k-mean (TKM) algorithm. The algorithm is given below:

### 4.5.1 Temporal K-Mean (TKM) Algorithm

1. Compute the initial value for the i$^{th}$ cluster centroid as follows:

$$c_{ij} = \frac{1}{M}\sum_{1+(i-1)*M}^{i*M} c_j \quad \ldots(49)$$

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 j$^{th}$ coefficient of the feature and $c_{ij}$ is the initial value of the i$^{th}$ cluster for j$^{th}$ 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 near by clusters m and k for j$^{th}$ frame based on proximity index. The cluster with two consecutive proximity index m and k are nearby clusters to j if

$$M * m \leq j \leq k * M \quad \ldots(50)$$

5. Compute the distance of the j$^{th}$ 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.

In order to apply the above mentioned algorithm, the speech signal is first segmented into frame of 100 ms with 50% overlapping. We refer to this as 1$^{st}$ level frame. Each 1$^{st}$ level frame is now passed through two parallel systems. The 1$^{st}$ 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 1$^{st}$ level frame into frame of size 20 ms with 50% overlapping. We refer to this as 2$^{nd}$ level frame. The MFCC and LPCC features are extracted from each 2$^{nd}$ 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 10$^{th}$ dimensional predictor signal

aggregated to a 13-dimentional 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 2nd 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 of 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 these features as High-level MFCC and High-level LPCC features respectively. The performance of the High-level MFCC and LPCC features has been evaluated using the statistical evaluation technique as well as HMM based recognizer.

Table – 16: KL-distance for the High-level features

| Feature vector | KL Distance |
|---|---|
| High-Level MFCC | 0.5754 |
| High-Level LPCC | 0.2958 |

Table – 17: F-ratio value for High-level features

| Feature vector | F-ratio value when the tones are same and base phonemes are different | F-ratio value when the base-phonemes are same and tones are different |
|---|---|---|
| High-Level MFCC | 5.7376 | 3.8250 |
| High-Level LPCC | 4.4236 | 3.5468 |

The statistical values F-ratio and KL-distance indicates that High-level MFCC and LPCC separated from each other in the feature space are having high tonal vowel discrimination capability even when the vowels are separated from each other only by tone or only by base-phoneme.

To find the maximum average distance among the cumulative distribution functions computed from the PDF functions of the high-level MFCC and LPCC feature vectors extracted from the tonal vowels KS-test has been performed. The result of the experiment is given in table-17.

Table-18: Average maximum distance among the CDFs for the high-level MFCC and LPCC feature vectors

| **Feature** | **Average maximum distance between the CDFs when** | | |
|---|---|---|---|
| | **Tones are different base-phonemes are same** | **Tones are same base-phonemes are different** | **Both tone and base-phoneme are different** |
| **High Level MFCC** | 0.8190 | 0.8505 | 0.8696 |
| **High Level LPCC** | 0.7069 | 0.7370 | 0.6799 |

From the above results, it has been observed that high level MFCC and LPCC features are occupying different locations in the feature space. Further, the separation is uniform across when the variability in the tonal phone is due to tone only or base-phoneme only or both. Thus we conclude that High-level MFCC and LPCC features are suitable parameterization technique for tonal vowel recognition. To evaluate the performance of the High-level features in terms of recognition accuracy of the HMM model, Hidden Markov models have been trained separately using the High-level MFCC and LPCC features. The results of the experiments are listed in table-19.

Table 19: Performance of the HMM based recognizer for recognizing the tonal vowels training and testing with High-level MFCC and LPCC features

| | Recognition accuracy for training and testing Features (in %) | |
|---|---|---|
| | High-level MFCC Features | High-level LPCC Features |
| Correctly recognized the tonal vowel | 89.13 | 83.41 |
| Incorrectly recognized as a tonal vowel with same base phoneme but different tone | 5.62 | 8.71 |
| Incorrectly recognized as a tonal vowel with same tone but different base phoneme | 2.33 | 4.12 |
| Incorrectly recognized as a tonal vowel with different tone and different base phoneme | 2.92 | 3.76 |

The recognition accuracies of the HMM ASR are found to be increased considerably due to the use of High-level MFCC and LPCC features. This observation confirms the fact that High-level MFCC and LPCC features are suitable parameterization technique for recognition of tonal vowels.

# 5. Conclusion and Feature Work

The work reported in this report presents a method of developing a automatic speech recognition system for the tonal languages of Arunachal Pradesh. The languages can be broadly categorised into two categories – tonal and non-tonal. Tone plays an important role in distinguishing among the syllables of a tonal language whereas in non-tonal language, tone cannot change the lexical meaning of a syllable. English, Hindi, Assamese etc. are example of non-tonal language and Chinese, Japanese, Apatani, Nyishi etc. are example of tonal language. Due to the active articipation of the tone related information in determining the meaning of a syllable, the tonal speech recognition systems are different from non-tonal speech recognition. In this report we have presented a detail analysis of the most commonly used speech parameters: Mel Frequency Cepstral Coefficient (MFCC), Linear Predictor Cepstral Coefficients (LPCC), Reflection Coefficient (RC), Log Area Ratio (LAR) and Prosodic features for their tonal speech discrimination capability. Analyses of the features have been done using statistical evaluation metrics – KL- distance, F-ratio test and KS test. Further, the features are investigated for their tonal phoneme recognition accuracy using a Hidden Markov Model based recognizer. The tonal phoneme recognition consists of two subtasks – base phoneme recognition and associate tone recognition. Considering the fact that some of the speech features are inherently good in discriminating among the base-phonemes and some other parameters are good in discriminating among the tones, different combinations of the speech features are evaluated for their tonal phoneme discrimination capability. The observation window size plays an important role in extracting meaningful, high resolution information from the speech signal. The spectral features like MFCC and LPCC are extracted from short observation window of duration 20~25 ms. On the other hand prosodic features are supra segmental feature and thus it needs long observation window for the extraction of high resolution information. Since spectral and prosodic features represent different aspects of the speech signal, combination of this information is necessary for many speech based applications. A feature concatenation method has been proposed that combines the features from two different windows size and at the same time preserves the temporal information of the speech signal. These combined features are evaluated for tonal speech recognition.

## 5.1 Major observations

- All the features, except the prosodic feature exhibit change in entropy due to the change in tone and base-phoneme together. However, the major contributor to the change in entropy is the change in base-phoneme. Therefore, the change in tone without the change in base-phoneme remains undetected.
- The prosodic features can capture the change in tone of the Tonal Base Unit (TBU). However, it fails to identify the change in base-syllable itself.

- Combining the features from multiple sources can improve the performance of the tonal speech recognition system.
- The features are broadly classified as segmental and supra-segmental features. The segmental features can be extracted with high resolution only from short observation windows like MFCC, LPCC etc. whereas supra-segmental features like prosodic features can be captured efficiently from long observation window. Therefore, in order to combine the features when a common window size is considered, their combined feature set lose significant information.
- The time-varying property of the speech signal contributes significantly in detection of the sound unit represented by the speech signal. When features from multiple windows size combined together, the temporal information of the smaller observation windows have to be preserved.
- The Hidden Markov Model (HMM) based automatic speech recognition system models the speaker specific information in addition to the phonetic information. Therefore, when normalization techniques are used to minimize the intra-speaker and inter-speaker variability, there speech recognition performance improves.

## 5.2 Future Works

***a) Optimal Feature Selection***

The performance of a speech recognition system depends on the selection of the optimal feature set. The feature selection is the first step of any Automatics Speech Recognition (ASR) System and the errors in this phase are propagated to the subsequent phases. More in-depth analysis of the speech features is required for the development of robust ASR system. Evidence from multiple sources need to be analysed before coming with a feature set which can clearly represents the distinguishing property of tonal base units at noisy ambient conditions.

***b) Development of Large Vocabulary Speech Database***

To develop automatic speech recognition system in Arunachali tonal languages, a large vocabulary speech database need to be developed. The database must be phonetically rich and include all forms of rule based viabilities of the contributing language.

***c) Analysis of Acoustical Cues***

Most of the tonal recognition researches are carried out using the existing feature extraction methods. However, these techniques are inherently biased towards non-tonal languages. Therefore in-depth analysis of the acoustical cues of the speech signal is required to identify the acoustical properties which can contributes significantly to tonal speech recognition but not detected by the existing feature extraction techniques.

### *d) Analysis of the context dependency*

The tonal speech recognition research is mostly centred on the detection of the tonal base unit. However, the context in which the TBU appears also has significant impact on the property of the TBU. Therefore, analysis of the context related information on the TBU need to be investigated.

# References


1. Rabiner, Lawrence, and Biing-Hwang Juang. "Fundamental Of Speech Recognition Prentice-hall International." (1993).
2. Seide, Frank, Gang Li, and Dong Yu. "Conversational Speech Transcription Using Context-Dependent Deep Neural Networks." Interspeech. 2011.
3. Hinton, Geoffrey, et al. "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups." IEEE Signal Processing Magazine 29.6 (2012): 82-97.
4. Dahl, George E., et al. "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition." IEEE Transactions on Audio, Speech, and Language Processing 20.1 (2012): 30-42.
5. Abdel-Hamid, Ossama, et al. "Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition." 2012 IEEE international conference on Acoustics, speech and signal processing (ICASSP).IEEE, 2012
6. Hayes, Monson H. Schaum's Outline of Digital Signal Processing. McGraw-Hill, Inc., 1998.
7. Kondoz, A. M. "Digital speech." Coding for Low Bit Rate Communications (2007).
8. Martin, Thomas B., A. L. Nelson, and H. J. Zadell. SPEECH RECOGNITION BY FEATURE-ABSTRACTION TECHNIQUES. RAYTHEON CO WALTHAM MASS, 1964.
9. Vintsyuk, Taras K. "Speech discrimination by dynamic programming." Cybernetics 4.1 (1968): 52-57.
10. Reddy, D.R., An Approach to Computer Speech Recognition by Direct Analysis of the Speech Wave, Tech. Report No. C549, Computer Science Dept., Stanford Univ., September 1966.
11. Velichko, V.M. and Zagoruyko, N.G., Automatic Recognition of 200 Words, Int. J.Man-Machine Studies, 2: 223, June 1970.
12. Sakoe, H. and Chiba, S., Dynamic Programming Algorithm Optimization for spoken Word Recognition, IEEE Trans.Acoustics,Speech, Signal Proc.,ASSP-26(1): 43-49, February 1978.
13. Karam, M., Khazaal, H. F., Aglan, H., & Cole, C. (2014). Noise removal in speech processing using spectral subtraction. Journal of Signal and Information Processing, 2014.
14. Tappert, C.C., Dixon, N.R., Rabinowitz, A.S. and Chapmam, W.D., Automatic Recognition of Continuous Speech Utilizing Dynamic Segmentation, Dual Classification, Sequential Decoding and Error Recovery, Rome Air Dev. Cen , Rome, NY, Tech. Report TR-71-146 ,1971 .
15. Jelink,F., Bahl, L.R. and Mercer, R.L., Design of a Linguistic Statistical Decoder for the Recognition of Continuous Speech , IEEE Trans. Information Theory , IT-21: 250-256, 1971 .
16. Jelinek,F., The Development of an Experimental Discrete Dictation Recognizer, Proc. IEEE, 73 (11): 1616-1624,1985.
17. Rabiner, L.R., Levinson, S.E., Rosenberg, A.E. and Wilpon, J.G., Speaker Independent Recognition of Isolated Words Using Clustering Techniques, IEEE Trans. Acoustics, Speech, Signal Proc.,and ASSP-27: 336-349, August 1979.
18. Sakoe, Hiroaki. "Two-level DP-matching--A dynamic programming-based pattern matching algorithm for connected word recognition." Acoustics, Speech and Signal Processing, IEEE Transactions on 27.6 (1979): 588-595.

19. Bridle, J. S., & Brown, M. D. (1979, November). Connected word recognition using whole word templates.In Proc. Inst. Acoust. Autumn Conf (pp. 25-28).
20. Myers, C. S., &Rabiner, L. R. (1981). A level building dynamic time warping algorithm for connected word recognition. Acoustics, Speech and Signal Processing, IEEE Transactions on, 29(2), 284-297.
21. Lee, C. H., &Rabiner, L. R. (1989). A frame-synchronous network search algorithm for connected word recognition. Acoustics, Speech and Signal Processing, IEEE Transactions on, 37(11), 1649-1658.
22. Ferguson, Ed., Hidden Markov Models for Speech,IDA,Princeton,NJ,1980.
23. Rabiner, L.R., A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, Proc. IEEE, 77 (2): 257-286, February 1989.
24. Lee, K.F., Hon, H.W. and Reddy, D.R., An Overview of the SPHINX Speech Recognition System, IEEE Trans. Acoustics , Speech , Signal Proc. , 38:600-610 , 1990 .
25. Chow, Y., et al. "BYBLOS: The BBN continuous speech recognition system." ICASSP'87. IEEE International Conference on Acoustics, Speech, and Signal Processing. Vol. 12. IEEE, 1987.
26. Paul, D. B. "The Lincoln robust continuous speech recognizer." International Conference on Acoustics, Speech, and Signal Processing,. IEEE, 1989.
27. Weintraub, Mitch, et al. "Linguistic constraints in hidden Markov model based speech recognition." International Conference on Acoustics, Speech, and Signal Processing,. IEEE, 1989.
28. Zue, Victor, et al. The MIT SUMMIT speech recognition system: A progress report. MASSACHUSETTS INST OF TECH CAMBRIDGE LAB FOR COMPUTER SCIENCE, 1989.
29. Lee, Chin-Hui, et al. "Acoustic modeling for large vocabulary speech recognition." Computer Speech & Language 4.2 (1990): 127-165.
30. Giuseppe Riccardi, Active Learning: Theory and Applications to Automatic Speech Recognition, IEEE Transactions On Speech And Audio Processing, Vol. 13, No. 4, July 2005.
31. Lippmann, R. P. (1987). An introduction to computing with neural nets. ASSP Magazine, IEEE, 4(2), 4-22.
32. Waibel, Alex, et al. "Phoneme recognition using time-delay neural networks." IEEE transactions on acoustics, speech, and signal processing 37.3 (1989): 328-339.
33. Juang, Bing-Hwang, and SadaokiFurui. "Automatic recognition and understanding of spoken language-a first step toward natural human-machine communication." Proceedings of the IEEE 88.8 (2000): 1142-1165.
34. Juang, Biing-Hwang, Wu Hou, and Chin-Hui Lee. "Minimum classification error rate methods for speech recognition." IEEE Transactions on Speech and Audio processing 5.3 (1997): 257-265.
35. Normandin, Yves, Regis Cardin, and Renato De Mori. "High-performance connected digit recognition using maximum mutual information estimation." IEEE Transactions on speech and audio processing 2.2 (1994): 299-311.
36. Watanabe, S., Minami, Y., Nakamura, A., & Ueda, N. (2004). Variational Bayesian estimation and clustering for speech recognition. IEEE Transactions on Speech and Audio Processing, 12(4), 365-381.
37. Frank Wessel and Hermann Ney, Unsupervised Training of Acoustic Models for Large Vocabulary Continuous Speech Recognition, IEEE Transactions On Speech And Audio Processing, Vol. 13, No. 1, January 2005.

38. Nakamura M, Iwano K, Furui S. The effect of spectral space reduction in spontaneous speech on recognition performances. In2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07 2007 Apr 15 (Vol. 4, pp. IV-473). IEEE.
39. Furui, Sadaoki, et al. "Cluster-based modeling for ubiquitous speech recognition." Ninth European Conference on Speech Communication and Technology. 2005.
40. Mathias De-Wachteret.al., Template based continuous speech recognition ,IEEE transactions on Audio, speech and Language processing, Vol.15,No.4, May 2007.
41. Xinwei Li et.al., Solving large HMM Estimation via Semi-definite programming , IEEE Transactions on Audio, speech and Language processing, Vol.15,No.8, December 2007.
42. Hung, Jeih-Weih, and Wei-Yi Tsai. "Constructing modulation frequency domain-based features for robust speech recognition." IEEE transactions on audio, speech, and language processing 16.3 (2008): 563-577.
43. Zweig, Geoffrey, and Patrick Nguyen. "A segmental CRF approach to large vocabulary continuous speech recognition." Automatic Speech Recognition & Understanding, 2009.ASRU 2009.IEEE Workshop on.IEEE, 2009.
44. Rybach, D., Gollan, C., Heigold, G., Hoffmeister, B., Lööf, J., Schlüter, R., & Ney, H. (2009, September). The RWTH aachen university open source speech recognition system. In Interspeech (pp. 2111-2114).
45. Mohamed, Abdel-rahman, Dong Yu, and Li Deng. "Investigation of full-sequence training of deep belief networks for speech recognition."INTERSPEECH.Vol. 10. 2010.
46. Povey, Daniel, et al. "Subspace Gaussian mixture models for speech recognition." 2010 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2010.
47. Soltau, Hagen, George Saon, and Brian Kingsbury. "The IBM Attila speech recognition toolkit." Spoken Language Technology Workshop (SLT), 2010 IEEE.IEEE, 2010.
48. Povey D, Ghoshal A, Boulianne G, Burget L, Glembek O, Goel N, Hannemann M, Motlicek P, Qian Y, Schwarz P, Silovsky J. The Kaldi speech recognition toolkit. In IEEE 2011 workshop on automatic speech recognition and understanding 2011 (No. CONF). IEEE Signal Processing Society.
49. Gemmeke JF, Virtanen T, Hurmalainen A. Exemplar-based sparse representations for noise robust automatic speech recognition. IEEE Transactions on Audio, Speech, and Language Processing. 2011 Feb 7;19(7):2067-80.
50. J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 19, no. 7, pp. 2067–2080, 2011.
51. F. Seide, G. Li and D. Yu, "Conversational speech transcription using context-dependent deep neural networks," in INTERSPEECH, 2011.
52. G. Hinton et al., "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The shared views of four research groups," IEEE Signal Processing Magazine, vol. 29, no. 6, pp.82-97, 2012.
53. G. E. Dahl et al., "Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 20, no. 1, pp. 30–42, 2012.

54. O. Abdel-Hamid and et al, "Applying Convolutional Neural Networks Concepts to Hybrid NN-HMM Model for Speech Recognition," in in IEEE international conference on Acoustics, speech and signal processing (ICASSP), IEEE, 2012.
55. U. Bhattacharjee, "Recognition of the Tonal Words of Bodo Language." International Journal of Recent Technology and Engineering. Volume-1(2013).
56. H. M. Wang, J. L. Shen, Y. J. Yang, C. Y. Tseng, and S. L. Lee, “Complete Chinese dictation system research and development”, Proceedings ICASSP-94, Vol. 1, pp. 59-61.
57. Chen, C.J., Li,H. L. Shen and Fu, G.K., "Recognize tone languages using pitch information on the main vowel of each syllable." Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on. Vol. 1. IEEE, 2001.
58. Sarmah, P., “Tone Systems of Dimasa and Rabha: A Phonetic and Phonological Study”, Doctoral dissertation, University of Florida, 2009.
59. Sun, J. T., “Tani languages”, In The Sino-Tibetan Languages, edited by G. Thurgood and R. LaPolla, pp. 456-466, London and New York: Routledge, 2003.
60. Post, M.W. and Kanno, T., “Apatani Phonology and Lexicon, with a Special Focus on Tone”, Himalayan Linguistics, Vol. 12(1):17-75, 2013.
61. Yip, M., The Tonal Phonology of Chinese, New York: Garland Publishing, 1991.
62. Beach, D.M., “The Science of Tonetics and Its Application to Bantu Languages”, in Bantu Studies, 2nd Series, Vol. 2, PP. 75-106, 1924.
63. Woo, N.H., Prosody and Phonology, Doctoral dissertation, MIT, 1969.
64. Doke, C.M., A Comparative Study in Shona Phonetics, Johannesburg, University of Witwatersrand Press, 1931.
65. Pink, K., “Tone Languages”, Ann Arbor, University of Michigan Press, 1964.
66. Leben, W., “Suprasegmental Phonology”. Ph.D. dissertation, MIT, 1973.
67. Goldsmith, J., “An overview of autosegmental phonology”, Linguistic Analysis, 2(1): 23-68, 1976.
68. Rao, K. Sreenivasa, and Shashidhar G. Koolagudi. Emotion recognition using speech features. Springer Science & Business Media, 2012.
69. Makhoul, J. (1975). Linear prediction: A tutorial review. Proceedings of the IEEE, 63(4), 561-580.
70. Raja, G.S. and Dandapat, S., 2010. Speaker recognition under stressed condition. International Journal of Speech Technology, 13(3), pp.141-161.
71. Atal, B. S., Chang, J. J., Mathews, M. V., & Tukey, J. W. (1978). Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. The Journal of the Acoustical Society of America, 63(5), 1535-1555.
72. Fant, G. (1971). Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations (Vol. 2). Walter de Gruyter.
73. Ververidis, D., & Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. Speech Communication, 48, 1162–1181.
74. Shriberg, L. D., Paul, R., McSweeny, J. L., Klin, A., Cohen, D. J., & Volkmar, F. R. (2001). Speech and prosody characteristics of adolescents and adults with high-functioning autism and Asperger syndrome. Journal of Speech, Language, and Hearing Research, 44(5), 1097-1115.

75. Mary, L. and Yegnanarayana, B., 2006. Prosodic features for speaker verification. In Ninth International Conference on Spoken Language Processing.
76. Taylor, P., “Analysis and synthesis of intonation using the tilt model”, J. Acoust. Soc. Am., Vol. 107, no. 3, p 1697–1714, Mar. 2000.
77. Carey, M.J., Parris, E.S., Lloyd-Thomas, H. and Bennett, S., 1996, October. Robust prosodic features for speaker identification. In Spoken Language, 1996. ICSLP 96. Proceedings, Fourth International Conference on (Vol. 3, pp. 1800-1803). IEEE.
78. Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V. and Woodland, P. (2000). The HTK Book Version 3.0. Cambridge, England, Cambridge University.