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.

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

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 2π 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 ω 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 $|X(\omega)|$. α and γ 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 [ɑ́:] ([a:] rising), [ɑ̀:] ([a:] falling) and [ɑ̄:] (([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]	[ɪ̄]	[ɪ̀]
ʊ	[ʊ́]	[ʊ̄]	[ʊ̀]
ɑː	[ɑ́:]	[ɑ̄:]	[ɑ̀:]
ɛ	[ɛ́]	[ɛ̄]	[ɛ̀]
ɔ	[ɔ́]	[ɔ̄]	[ɔ̀]
ə	-	[ə̄]	-

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:

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

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, μ_i is the average mean for the base phoneme i, $\mu_{\beta,i}$ is the average mean for phoneme i for tone β, $x_\beta^{(i)}$ indicates an instance of the phoneme i for tone β. 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, μ is the Gaussian mean vector and σ^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
[ɪ̄]	0.7292	[ɑ́:]	1.4317	[ɛ̀]	1.1002
[ɪ̀]	0.5993	[ɑ̄:]	1.9577	[ɔ́]	1.6260
[ʊ́]	0.9437	[ɑ̀:]	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)