Role of pitch slope and duration in synthesized Mizo tones

D. Govind[1]*, Priyankoo Sarmah*[2] *and S. R. Mahadeva Prasanna*[1]

[1]Department of Electronics & Electrical Engineering
[2]Department of Humanities and Social Sciences
Indian Institute of Technology Guwahati, Assam, India
govinddmenon@gmail.com, {priyankoo, prasanna} @ iitg.ernet.in

Abstract

This paper reports the results of an attempt to synthesize the lexical tones of the Mizo language. Firstly, the study reported in this paper attempts to confirm the findings of previous acoustic studies on Mizo tones. Secondly, using the parameters defined in the previous acoustic studies, the work reported in this paper synthesized Mizo tones and then confirmed the acceptability of the synthesized tones from native speakers of Mizo. The work reported in this paper confirms that (a) mean fundamental frequency (F_0) alone cannot be a parameter to recognize Mizo tones; (b) mean F_0 and tone slope (F_d) information integrated into synthesized Mizo tones elicit better identification and acceptance and (c) durational information is important for correct identification of rising tones in Mizo.

Index Terms: Tone language, contour tone, Mizo, tone synthesis, epoch based prosody modification

1. Introduction

Processing of lexical tones in tone languages have been of interest to both the linguistic and speech processing communities. Even though a lot of progress have been made in the domain of speech synthesis and recognition in tone languages like Mandarin Chinese, tones still pose a challenge in correct recognition of speech. For example, [1] report that in Mandarin Chinese automatic speech recognition systems 48.3% of the errors are caused due to misrecognition of tones.

While the Indian subcontinent is home to a lot of tone languages, they have not received much attention from the linguistic or the speech processing communities. The language of the current study, Mizo, is a Tibeto-Burman tone language spoken by about 674,756 native speakers in the province of Mizoram in North-East India [1]. Mizo tones have been described in several works [2–5] ; however, only [5] conducts acoustic study on Mizo tones. As reported in [5], Mizo has four salient lexical tones namely, high, low, falling and rising. Figure 1 shows the pitch tracks of four Mizo tones extracted from an adult, female Mizo speaker in [5].

Based on the acoustic study on Mizo tones [5], in this study, we synthesized a set of stimuli from the low toned Mizo [pa:] syllable as the source and the falling and rising tones as targets. The [pa:] syllable can mean *father, mushroom* and *male*, when it is produced with a falling, rising and a low tone, respectively. In the synthesis of the target tones based on the [pa:] syllables, three parameters were manipulated namely, the average F_0, the slope of the pitch (F_d) and the duration of the syllable. The duration of the syllable was manipulated only for the rising tone as it is seen that the rising tones in Mizo are comparatively longer than the other three tones [5]. Hence, in the synthesis of the rising tones, three different durational parameters were used- firstly, the original duration of the source was maintained; secondly, the synthesized rising tone was elongated by a factor of 1.5 and thirdly, the synthesized tone was elongated by a factor of 2.0. As the falling tone in Mizo is comparatively shorter than the other three tones, no durational manipulation was deemed necessary for the synthesized falling tone. After the stimuli were synthesized, perception tests were conducted where ten Mizo speakers identified and rated the acceptability of the synthesized tones.

[1] *http://www.censusindia.net/* accessed on 30/11/2011

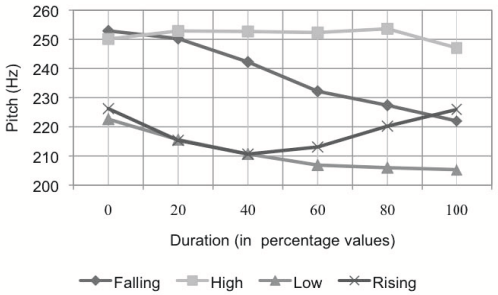

Figure 1: Pitch tracks for four Mizo tones [5]

The results of the perception test confirmed that, while identifying Mizo tones, native speakers depend significantly on the slope of the F_0 track. However, providing them only with average F_0 information resulted in incorrect identification of the tones. It was also seen that for rising tones, Mizo speakers also depended on the duration of the tone. A short tone duration resulted in wrong identification of the tones, while a longer duration resulted in near-perfect identification.

This paper is organized as follows: Section 2 describes the procedure of prosody modification for synthesizing the Mizo tone stimuli. Section 3 describes the perception test conducted as part of the current study. Section 4 describes the results of the perception experiment while Section 5 summarizes this study.

2. Synthesis of stimuli

The synthesis in the current study is based on original sound files recorded from a female Mizo speaker, aged 25 at the time of recording. The sounds were recorded in a sentence frame.

2.1. Epoch based prosody modification

To reduce the distortion in the prosody modification, we have used epoch based prosody modification method for manipulat-

ing the pitch and duration to convert one Mizo tone to another. Epoch based prosody modification proposed in [6] consists of the following steps:

- Estimating the accurate epoch locations
- Deriving the modified epoch locations according to the desired prosody
- Reconstructing the waveform to obtain the prosody modified speech

2.2. Estimating the accurate epoch locations

Epochs are the instants of glottal closure in case of voiced speech and onset of bursts or frication incase of unvoiced speech [6, 7]. These epoch locations are used as the analysis pitch marks for the prosody modification [8]. The accurate epoch locations are estimated by the zero frequency filtering (ZFF) method [6, 7].

2.3. Deriving the modified epochs locations

The modified epoch locations have to be derived according to the desired prosody for generating the synthesis pitch marks. The synthesis pitch marks can be generated by scaling the epoch intervals, which is the difference between successive epoch locations for pitch modification. In case of duration modification, the epoch interval plot is resampled according to the desired duration modification factors. Epoch locations stating from a point in the modified and interpolated epoch interval plot gives the locations of the synthesis pitch marks.

2.4. Reconstructing the prosody modified speech waveform

For synthesizing the speech waveform, the analysis pitch marks that are nearest to the synthesis pitch marks are found. The speech samples of the original epoch interval starting from the analysis pitch mark are copied to the synthesis pitch mark locations. The resulting sequence will be the speech waveform with desired prosody by processing all the epochs.

To further improve the naturalness of the prosody modification, the desired prosody of the glottal activity (GA) regions are retained by the accurate ZFF based GA detection [9].

Figure 2 plots the waveforms and F_0 contours of original [pa:] low tone, synthesized rising tone and original rising tone. The rising pattern in the synthesized tone contour can be observed from the Figure 2(e) after incorporating the pitch slope of the desired rising tone. Figure 3 plots the waveforms and F_0 contours of the original [pa:] low tone, the synthesized falling tone and the original falling tone. The falling nature of the contour can be identified in the synthesized tone as given in Figure 3(e) unlike the F_0 contour of the original low tone shown in Figure 3(d).

Table 1 shows the parameters on which the low toned [pa:] syllable was modified into other tones. The stimuli where the F_0Avg is modified, only the average F_0 values of the target tone is embedded in the synthesized stimuli and the slope of the tone remained unmodified. The stimuli where the slope is modified, the pitch slope values are also incorporated. The stimuli where the duration is modified, their total duration is increased by a factor of either 1.5 or 2.

3. Perception test

The synthetic stimuli along with the original sounds recorded from a 25 year old, female Mizo speaker were presented to ten native speakers of Mizo (9 male and 1 female). They all were between 25 and 35 years in age. Each of the ten Mizo speakers were asked to perform an identification and a goodness judgment task on the stimuli. In total there were 12 synthesized stimuli (6 conditions x 2 different source syllables) and 6 real speech stimuli (3 tone types x 2 separate iterations). The 18 stimuli (12 synthesized + 6 real speech) were repeated 5 times each resulting in 90 total stimuli.

Table 1: Table of synthesized stimuli indicating the manipulated parameters

	Condition	Avg.F_0	Slope	Dur.
[pa:] Falling	F_0Avg	✓	×	×
[pa:] Falling	F_0Avg+ F_d	✓	✓	×
[pa:] Rising	F_0Avg	✓	×	×
[pa:] Rising	F_0Avg+ F_d	✓	✓	×
[pa:] Rising	F_0Avg+ F_d + Dur. 1.5	✓	✓	✓
[pa:] Rising	F_0Avg+ F_d + Dur. 2	✓	✓	✓

The stimuli were presented using Praat MFC test interface [10] on a desktop computer and the subjects heard the sounds though a pair of headphones. The stimuli were presented on a desktop computer in a noise-free environment. The speakers were presented with clickable choices on a computer screen and depending on what tone they heard, the speakers were required to click on one of the three meanings associated with the syllable [pa:]. A null option was also provided as *none*, if the presented stimulus did not evoke any meaning reference in the speakers or if the speakers considered the tone of the stimulus to be completely unacceptable. After choosing a meaning, the speakers had to judge the quality of the tone of the stimulus on a Likert scale ranging from 1 to 5, 1 being *poor* and 5 being *good*. The speakers were allowed to replay the sound twice if he needed. Each new stimulus was programed to be played after an initial silence of 1500 milliseconds. Each stimulus was repeated 5 times, resulting in a total of 90 stimuli that were randomly permuted with no doublets. The subjects completed the experiment in 40 minutes and their results were saved into a spreadsheet for further analysis.

4. Results

The results of the tone identification and goodness judgment tasks performed by the Mizo speaker are detailed in the sections below.

4.1. Tone identification

The results of the perception study conducted with the synthesized Mizo tones revealed that when the stimuli incorporates only the average F_0 information, identification of the tones is comparatively low. In case of the synthesized [pa:] with a falling tone that incorporates only the average F_0 information, the correct identification percentage is 67 (see Table 2). In the same condition the percentage of correct identification for a rising tone is 5. When the stimuli also incorporates slope information (F_d) with the average F_0 information, correct identification increases to 69% for the falling tone. While for the rising tone, in the same condition, correct identification marginally improves to 29%. As the duration of the rising tone in Mizo is longer than the other three tones, we decided to elongate the tones 1.5 and 2 times the total duration of the source [pa:] syllable. It was noticed that the perception of the rising tones with average F_0 and F_d information was significantly augmented

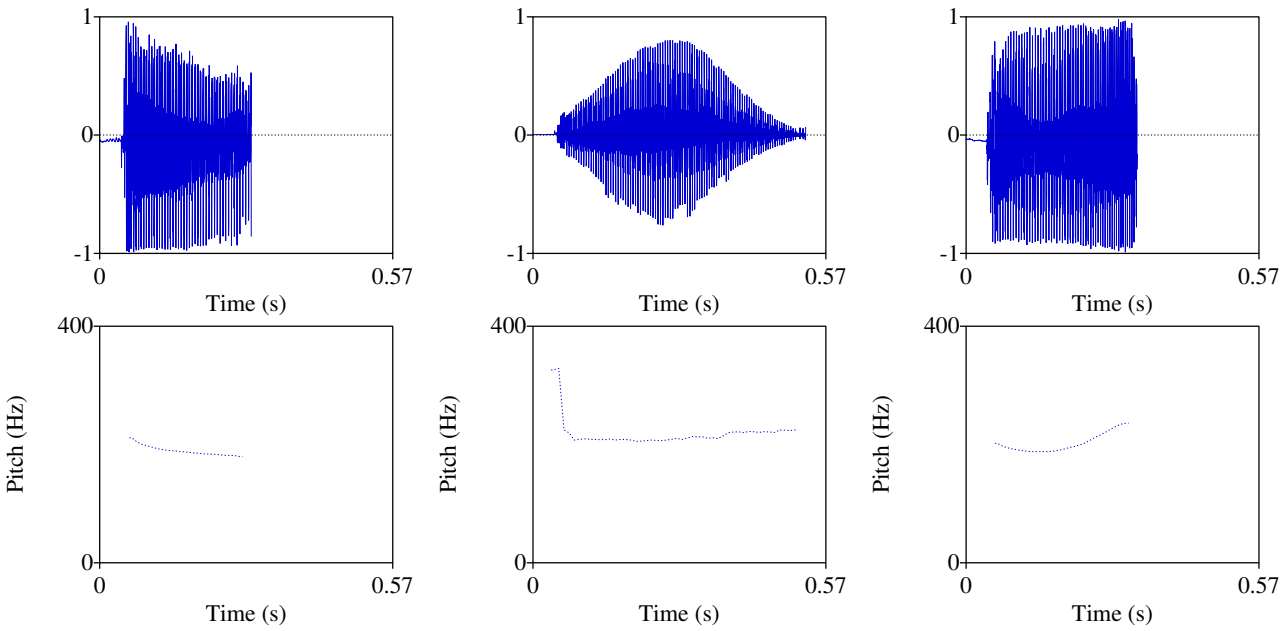

Figure 2: Synthesized [pa:]-R from [pa:]-L. Top left to top right: Speech waveforms of [pa:] L tone, synthesized [pa:] R tone, original [pa:] R tone. Bottom left to bottom right: F_0 contours of [pa:] L tone, synthesized [pa:] R tone, original [pa:] R tone.

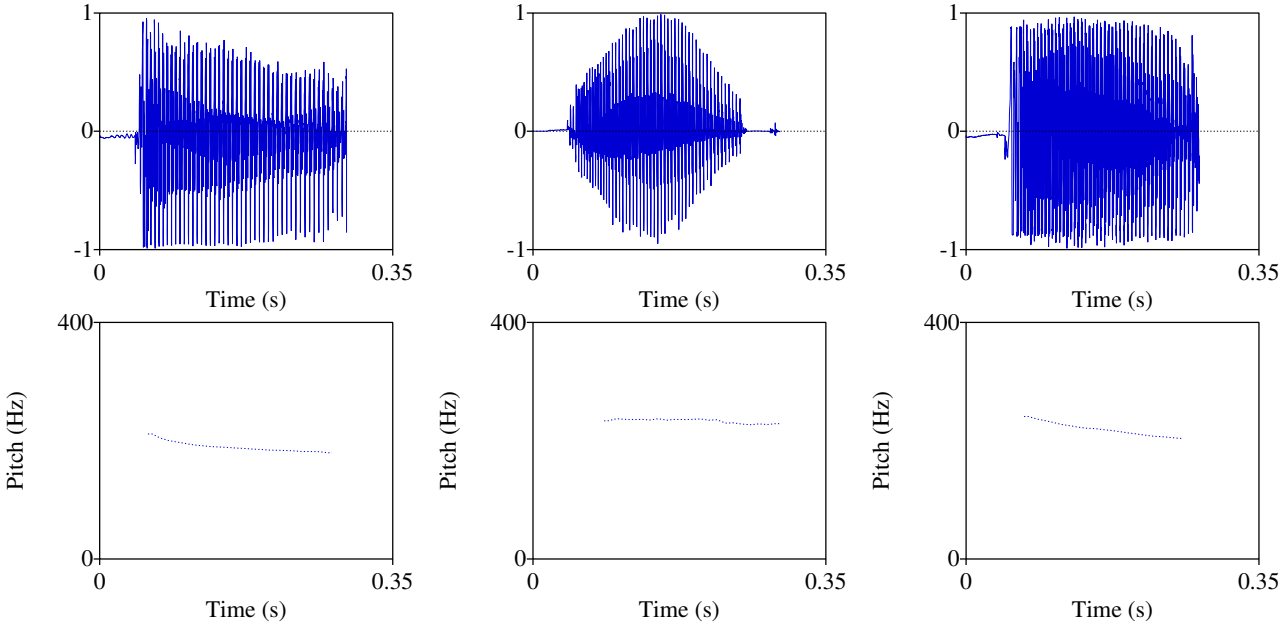

Figure 3: Synthesized [pa:]-F from [pa:]-L. Top left to top right: Speech waveforms of [pa:] L tone, synthesized [pa:] F tone, original [pa:] F tone. Bottom left to bottom right: F_0 contours of [pa:] L tone, synthesized [pa:] F tone, (c) original [pa:] F tone.

when the total durations of the tones were increased. When the stimuli were elongated by a factor of 1.5, the average correct perception of the synthesized rising tones increased to 72%. In the condition where the total duration was elongated by a factor of 2, the average correct perception of the rising tones increased to 84%. In order to ascertain that the subjects of the current study had native competence in identifying Mizo tones, we also subjected them to perceive 6 real speech stimuli as controls. The real speech stimuli were produced by a single Mizo speaker. As seen in Table 3, the identification of the tones in real speech was near perfect.

Table 2: Correctness in recognizing synthesized speech tones on [pa:]

Tone	Conditions	Correctness
Falling	F_0	67%
Falling	F_0 + F_d	69%
Rising	F_0	5%
Rising	F_0 + F_d	29%
Rising	F_0 + F_d + dur1.5	72%
Rising	F_0 + F_d + dur2.0	84%

Table 3: Correctness in recognizing real speech tones on [pa:]

Tone	Correctness
Falling	84%
Low	63%
Rising	98%

4.2. Goodness judgment

The synthesized stimuli were also rated for goodness by ten native speakers on a Likert scale of 5 levels with 1 being *poor* and 5 being *good*. As depicted in Table 4 the goodness rating for the synthesized falling tones is the best in the condition where the stimuli has only the F_0 information incorporated. In case of the rising tone, the stimuli with F_0 and F_d information with elongation with a factor of 2 are considered to be the best by the native speakers.

Table 4: Average goodness scores of correctly recognized synthesized speech tones on [pa:]

Tone	Conditions	Goodness	Std dev.
Falling	F_0	3.50	1.19
Falling	F_0 + F_d	3.02	1.39
Rising	F_0 + F_d	3.03	1.18
Rising	F_0 + F_d + dur1.5	3.31	1.31
Rising	F_0 + F_d + dur2.0	3.65	1.15

We also required the native speakers to rate the goodness of the tones in real speech. As depicted in Table 5, the native speakers rated the falling and the low tone as of equal goodness and rated the rising tone as near perfect.

5. Summary and conclusion

The results of the current study can be viewed from two perspectives. Firstly, it reinforces the findings of [5] that, for correct identification of Mizo tones, information about pitch slope is crucial. This study also confirms that, in case of the rising tone, apart from the pitch slope information, durational cues are also important. The perception of synthesized rising tones improve with the increase in the total duration of the [pa:] syllable. Secondly, the results of this study also confirm that the use of zero frequency filtering based prosody modification reduces the perceptual distortions of the modified speech. As the pitch slope modification was performed using the epoch based prosody modification method; the success of the participants of this study in identifying the synthesized tones confirms the effectiveness of the method in synthesizing one tone from the other.

Table 5: Average goodness scores of correctly recognized real speech tones on [pa:]

Tone	Goodness	Std dev.
Falling	4.16	1.10
Low	3.89	1.33
Rising	4.61	0.74

6. Acknowledgements

This work is a part of ongoing UKIERI project (2007-2011) titled, Study of *Source Features for Speech Synthesis and Speaker Recognition* between IIT Guwahati, IIIT Hyderabad and CSTR, University of Edinburgh, UK. The authors would also like to express their gratitude to the three anonymous reviewers, whose comments helped to improve this paper. The authors would like to acknowledge the help of Mr. Biswajit Dev Sarma in conducting the perception tests in this study. The authors are grateful to Mizo speakers, Mr. H. Lalhmingsanga and Mr. Lalhriatzuala for their assistance.

7. References

[1] H. Hao and Z. Jie, "Discriminative tonal feature extraction method in mandarin speech recognition," *J. China Univ. of Posts and Telecommun.*, vol. 14, no. 4, pp. 126–130, 2007.

[2] L. Chhangte, "A preliminary grammar of the Mizo language," Master's thesis, University of Texas, Arlington, 1986.

[3] L. Fanai, *Some aspects of autosegmental phonology of English and Mizo*, M. Litt. dissertation,CIEFL, Hyderabad, India, 1989.

[4] ——, *Some aspects of the lexical phonology of Mizo and English: an autosegmental approach*, Ph. D. dissertation,CIEFL, Hyderabad, India, 1992.

[5] P. Sarmah and C. R. Wiltshire, "A preliminary acoustic study of mizo vowels and tones," *J. Acoust. Soc. Ind.*, vol. 37, no. 3, pp. 121–129, 2010.

[6] S. R. M. Prasanna, D. Govind, K. S. Rao, and B. Yenanarayana, "Fast prosody modification using instants of significant excitation," in *Proc Speech Prosody*, May 2010.

[7] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. Audio, Speech and Language Process.*, vol. 16, no. 8, pp. 1602–1614, Nov. 2008.

[8] K. S. Rao and B. Yegnanarayana, "Prosody modification using instants of significant excitation," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, pp. 972–980, May 2006.

[9] N. Dhananjaya and B. Yegnanarayana, "Voiced/nonvoiced detection based on robustness of voiced epochs," *IEEE Signal Processing Letters*, vol. 17, no. 3, pp. 273–276, 2010.

[10] P. Boersma and D. Weenink, "Praat 5.2.22: A software for doing phonetics," in *http://www.fon.hum.uva.nl/praat/*, 2011.