Exploring Tonal Information for Lhasa Dialect Acoustic Modeling

Jian Li, Hongcui Wang, Longbiao Wang, Jianwu Dang , Kuntharrgyal khuru, Gyaltsen Lobsang

Tianjin key Laboratory of Cognitive Computing and Application, Tianjin University, China

Abstract: Detailed analysis of tonal features for Tibetan Lhasa dialect is an important task for Tibetan automatic speech recognition (ASR) applications. However, it is difficult to utilize tonal information because it remains controversial how many tonal patterns the Lhasa dialect has. Therefore, few studies have focused on modeling the tonal information of the Lhasa dialect for speech recognition purpose. For this reason, we investigated influences of the tonal information on the performance of Lhasa Tibetan speech recognition. Since Lhasa Tibetan has no conclusive tonal pattern yet, in this study, we used a four-tone pattern and designed a phone set based on the four contour contrasts scheme. Speech recognition performance was examined using the acoustic model with and without the pitch-related features. The experimental results showed that the character error rate (CER) was improved 11% after applying the tone based phone set and pitch-related features to DNN-HMM based speech recognition by comparing to that without tonal information. This preliminary study revealed that the tonal information plays an important role in speech recognition of Tibetan Lhasa dialect.

Experiment and Method

Phone set: There are totally 29 initial consonants and 48 final units without considering the tones.

Four tone patterns:

55: highest tone with flat pitch contour
51: high tone with a falling contour
53: low tone with increasing contour
132: firstly rising to a medium and then falling down

Initials	p ph t th c ch k kh ts tsh tʂ tʂh tɕ tɕh m n ȵ ŋ l f s ɬ ʂ ɕ çh ɹ w j
Finals	i e a o u ø i: e: a: ə: o: u: y: ø: ɛ: ao im em am om um in en an on yn iŋ eŋ aŋ oŋ uŋ ip ep ap əp op up i? e? a? o? u? ɛ? ir er ar or ur

Table 1. Lhasa dialect phone set

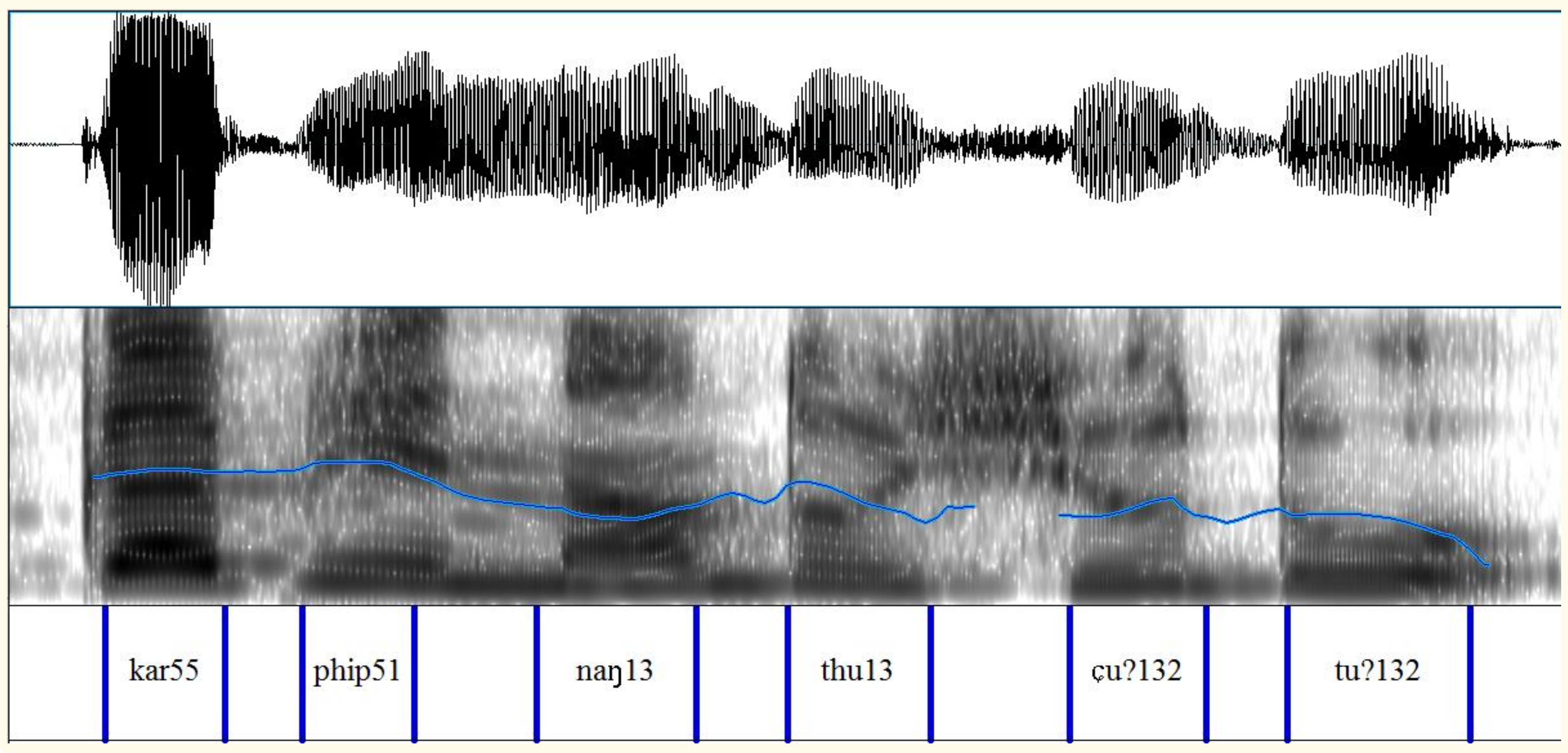

Figure 1. Part of one utterance spoken by a female speaker

Pitch tracking:

Kaldi pitch tracker operates on the speech signal to generate a set of candidate pitch values by calculating the normalized cross-correlation function (NCCF). After getting those estimated candidates, the dynamic programming (DP) is used to determine the best F0 at each frame based on a combination of local and contextual evidences.

Pitch related features:

- Log of F0
- Probability of voicing
- Delta features using ±2 frames of context
- 39-dimension MFCC

Acoustic modeling:

- CD-GMM-HMM:
  - Cepstral Mean Normalization
  - Linear Discriminated Analysis
  - Maximum Linear Likelihood Transform
- CD-DNN-HMM:
  - Pre-training DBN
  - Stochastic gradient descent
  - Window length is 11 frames
  - 6 hidden layers with 2,048 nodes per layer
  - Speaker normalization by feature space maximum likelihood linear regression (fMLLR)

Results and Discussions

For the baseline system, we use the non-tone related phone set and the input features are the traditional MFCC features. Then we change the phone set to be tone-related one and the input features are also MFCC features. For the third experiment, the input features are changed to the pitch related features.

	GMM	*DNN*
Non-tonal Phone + MFCC	*40.91%*	*33.97%*
Tonal Phone + MFCC	*39.80%*	*31.91%*
Tonal Phone + Pitch-related features	*37.75%*	*30.20%*

Table 2. Character error rate for different configurations

In this study, we explored the influence of the tonal information on the performance of Lhasa Tibetan speech recognition. We developed a tone based phone set for Lhasa dialect. The expended phone set included 221 phone units. To the best of our knowledge, it was the first study to discuss how to use tonal information to a Lhasa dialect ASR system. The performance evaluation on the Lhasa dialect corpus showed that the tone-related phone set can obtain better result compared with the non-tonal one. It suggests that the expanded phone set can help to distinguish the homophones, thus reducing the mismatching character error rate. After adding the pitch-related features, the performance got better. By using the tone-related phone set and pitch-related features, we can achieve the relative error reduction rate of 7.7 % on GMM-HMM and 11.1 % on DNN-HMM.