Abstract of Dissertation

Neural-Network-Based Tone Recognition of Continuous Speech of Standard Chinese Using Tone Nucleus Model

（声調核モデルに基づくニューラルネットワークを用いた標準中国語連続音声の声調認識）

王　暁東

In tonal languages, tones are used to distinguish lexical meaning of words. Meanwhile, tonal information is helpful to detect higher level prosody information. Due to this distinct function of tones and the possible assistance from prosody information which can be further detected using tonal information; tone recognition is desirable when constructing the automatic standard Chinese speech recognition system, thus attracting many researchers in the past decades. These studies can be generally divided into two types: embedded approaches and explicit approaches, i.e. approaches done as an integral part of or in parallel to the existing ASR framework. According to the technologies, efforts of explicit approaches, as the mainstream direction, were continuously made along two lines. One line is to construct appropriate statistical tonal models and classifiers, while the other is to make reasonable prosody models in order to overcome difficulties due to substantial F0 variations.

In this work, we adopted the Tone Nucleus Model, which pointed out that as a portion of syllabic F0 contour tone nucleus contains crucial information for tone perception and recognition, to suppress negative effect for tone recognition from neighboring tones, called tonal co-articulation. This model not only can provide a clear linguistic meaning for the F0 normalization process, but also can show explicit potentials for detecting intonation structure. On the other hand, Multi-Layer Perceptron (MLP), one kind of Neural Network (NN) approaches, was used to easily incorporate heterogeneous features, such as category feature and segment duration, which are important for tone recognition. Via integrating these two efficient methods, our proposal can exploit the above advantages to achieve a better performance in tone recognition of continuous speech of standard Chinese.

To realize this proposal, firstly we present an efficient algorithm to automatically extract tone nucleus. High performance of tone nucleus extraction was confirmed by the inspection on results of 50 utterances. With the assistance of this algorithm, input features were calculated, most of which are related to tone nuclei. As for the MLP tone classifier, one hidden layer was exploited to make the construction clearer and effective enough based on the universal approximation theorem for neural networks.

In order to evaluate the proposed system, comparative experiments were implemented both in the speaker dependent and independent tone recognition. In speaker dependent case, the system with MLP tone classifier and feature extraction from whole syllabic voiced part was constructed as the first baseline. Meanwhile, the second baseline is the system with HMM tone classifier and Tone Nucleus model, i.e. features extraction from tone nucleus, reported in the previous work. The same speech corpus was used in the reported work and thus its results can be directly compared with those of current work. Therefore, among these three systems, comparison was carried out. In speaker independent experiments, 20-fold cross-validation was used to avoid the selection of training and testing sets affecting the result, and Global (denoted as G) Mean/Standard-Deviation (denoted as MSD) feature normalization was preliminarily exploited to reduce the features varying with speakers. Then performance comparison of speaker independent tone recognition was implemented between the first baseline system and proposed system.

From the results of baseline systems and proposed system in speaker dependent and independent experiments we can see,

(1) In speaker dependent experiment, the proposed approach achieved an absolute error reduction of 1.3% compared to the 1st baseline, equal to a relative error reduction of 9.2%. In speaker independent experiment, absolute error reduction of 0.5% was also obtained by the proposed approach. The difference of the two systems lies in whether calculating the features from the whole syllabic voiced part or tone nucleus. The better performance of the proposed approach indicates that tone nuclei do keep important and robust discriminating features for the tone recognition.

(2) In speaker dependent experiment, the proposed approach got an absolute error reduction of 1.7% compared to the 2nd baseline, corresponding to relative improvement of 11.7%. The better performance can be attributed to the use of MLP and two additional features: segmental durations and syllable positions in the sentence. But they are difficult to be exploited in an HMM based approach.

However, some problems still exist in proposed tone recognition system. One of them is the feature normalization in speaker independent case, resulting in relatively large difference in performance (about 10%) between speaker dependent and independent statuses.

To solve this issue, three feature normalization approaches were proposed for speaker independent tone recognition, which are Shifting-Window feature normalization, Cumulative Distribution Function matching based on quantile histogram equalization and normalization inside syllabic voiced part, denoted as SW, CDF and InSyl respectively in this thesis.

With regard to evaluation of these normalization approaches, our previous speaker independent tone recognition based on Global Mean/Standard-Deviation normalization was taken as baseline for comparison, marked as MSD+G. The comparative experiments among this baseline, proposed approaches and hybrid of proposed approaches were implemented. From the results of these comparisons, we can conclude that each of these feature normalization approaches is significantly effective for speaker independent tone recognition. The best performance was achieved by the hybrid approach via combining MSD, SW and InSyl together. Through this hybrid approach, the difference of average error rate between speaker independent and dependent tone recognition was reduced from 10.4% to 3.9%.