A New Approach for Automatic Tone Error Detection in Strong Accented Mandarin Based on Dominant Set

Taotao Zhu, Dengfeng Ke, Zhenbiao Chen, Bo Xu

Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

{taotao, dfke, zbchen, xubo}@hitic.ia.ac.cn

10.21437/Interspeech.2010-283

Abstract

In this paper, we proposed a new approach based on dominant set [1] for tone error detection in strong accented Mandarin. First, the final boundary generated from forced alignment is regulated by the F0 contour in order to locate the final domain more accurately. After that, proper normalization techniques are explored for the tone features. Finally, clustering and classification methods based on dominant set are utilized for the tone error detection. The proposed approach is tested in comparison with the traditional k-means based method, experimental results show that it achieves more satisfying performance with an average Cross-Correlation 0.84 between human and machine, reaches to that between humans, which have verified the effectiveness of the proposed approach. The main advantage of this approach lies in not only the error pronunciation of tone can be well identified, but also the F0 pattern of the tone error can be informatively provided as the feedback.

Index Terms: CALL (Computer Assisted Language Learning), tone error detection, dominant set, forced alignment, F0

1. Introduction

As is known to all, Mandarin is tonal language. The widely used four tones in Mandarin are high, rising, low, falling tones, also denoted by Tones 1 to 4. Tone plays a significant role in the live communication because many words are differentiated solely by tone. However, it is more difficult to be pronounced correctly in comparison with initial and final due to the dialect of a speaker. Most people in China are using their native dialect and Mandarin, their pronunciations are always depending on how well they grasp the language. Therefore, detecting tone errors is an important component in Mandarin CALL systems [2], which aim to help the language learners correct and improve their pronunciations in the whole learning process.

In this paper, a new tone error detection approach based on clustering is proposed. The main idea can be described as: firstly, towards the pronunciations of each tone, the corresponding positive tone clusters and negative tone clusters are obtained via clustering, then the testing tones are assigned to their most correlated cluster using the similarity measure (or distance measure). In general, a straightforward approach to detect tone error in Mandarin is by means of tone recognition, however, in the real circumstances, the pitch variation of tone is not always according to the four canonical Mandarin tones ascribe to the dialectal accent within the pronunciation, this tonal modification leads to the recognition rate of tone is relatively low [3]. Motivated by the desire to determine the pitch variations in different tones more accurately, therefore, the idea based on clustering is presented for tone error detection.

In this work, the traditional model-based approaches such as HMM or GMM are not chosen to train the tone clusters, because these methods require a large number of samples to grasp the feature distribution, and yet the samples among different tone categories in the “tone error space” are not well-proportioned, some categories of the tone errors are incomplete and not easy to be collected. In this task, we use a new dominant set based approach for tone error detection due to its efficiency and directness in clustering and classification.

Dominant set is proposed by Pavan *et al.* [1] [4], and its corresponding clustering technique — dominant set clustering (DSC) has been widely used in the fields of image segmentation and video processing due to its intuitiveness, inherent hierarchical nature and superiority in clustering accuracy and stability, different from traditional clustering algorithms, it can automatically determine the number of clusters with low computational cost. Furthermore, dominant set can be used for classification [5]. Therefore, We introduce its application to the task of tone error detection.

The outline of this paper is organized as follows: Related works are discussed in Section 2; Section 3 formulates the dominant set clustering and classification algorithms; In Section 4, we describe our tone error detection approach; Experimental results and analysis are given in Section 5; Followed by conclusion and future work in Section 6.

2. Related works

In the last decade, great achievements have been made in CALL systems. Following the GOP (goodness of pronunciation) score used by Witt [2], lots of studies have been investigated, and the majority of them are based on posterior probability [6] and pronunciation rules [7] derive from the state-of-the-art speech recognition. In Franco *et al.* [6], posterior-based methods with native models are preferred in detection tasks. Ito *et al.* [7] introduce decision tree based error clustering with multi-thresholds. These works mainly focus on the segmental pronunciation error. By contrast, tone has drawn much less attention in the literature of CALL. Pan *et al.* [3] use the posterior probabilities generated by GMM for tone assessment on strong accented Mandarin speech. Zhang *et al.* [8] use log-posterior probabilities as the GOP score for tone mispronunciation detection under an MSD-HMM framework. Wei *et al.* [9] utlize HMM to detect tone errors, and the F0 after a CDF-matching normalization is used as the feature for tone model. The aforementioned methods for the task of tone error detection are based on posterior probabilities. Despite the fact these methods can perform well, both of them are heavily threshold-dependent. Comparing to the previous work, our approach avoids this problem to reach a high average CC between human and machine.

This work was supported by a grant from the National Natural Science Foundation of China No. 90820303

Algorithm 1: Dominant set clustering algorithm

```
Input: $G = (V, E)$, affinity matrix $A$
  1. Initialize k=1, $A^k = A$, $G^k = G$
  2. Solve equation (1): $\mathrm{x}^k$ and $f(\mathrm{x}^k)$
  3. Find the dominant set $S^k = \sigma(\mathrm{x}^k)$
  4. Remove $S^k$ from $V^k$ and its affinity relation in $G^k$
  5. Update $G^{k+1}$ and the affinity matrix $A^{k+1}$
  6. If $V^{k+1} \neq \emptyset$, $k \leftarrow k+1$, goto step 2; else exit
Output: $\bigcup_{k=1}^{K}\{S^k, \mathrm{x}^k, f(\mathrm{x}^k)\}$
```

Algorithm 2: Dominant set fast assignment algorithm

```
Input: affinity vector $\alpha \in \mathbb{R}^n, \bigcup_{k=1}^{K}\{S^k, \mathrm{x}^k, f(\mathrm{x}^k)\}$
1. $g^k = \frac{|S^k|-1}{|S^k|+1}(\frac{\sum_{h\in S^k}\alpha_h x_h^k}{f(\mathrm{X}^k)} - 1), k \in \{1, \cdots, K\}$
2. Get $k^* = \arg\max_k g^k$
3. If $g^{k^*} > 0, \hat{k} = k^*$; else $\hat{k}$=0
Output: $\hat{k}$
```

3. Dominant set clustering and classification

Dominant set is a graph-theoretic concept first accurately defined by Pavan *et al.* [1]. Motivated by the analogies between intuitive of a cluster and that of a dominant set of vertices, it generalizes the notion of a maximal clique in context of edge-weighted graphs. Pavan *et al.* [1] also have established an intriguing connection between dominant set and the extrema of a quadratic program over the standard simplex. Let $G = (V, E)$ be the edge-weighted graph, V is the vertex set, E is the edge set. Then finding dominant set can be transformed into solving the quadratic program:

$$\text{maxmize}: f(\mathrm{x}) = \frac{1}{2}\mathrm{x}^T A\mathrm{x} \quad \text{s.t.} \quad \mathrm{x} \in \Delta \tag{1}$$

$$\text{Where} \quad \Delta = \{\mathrm{x} \in \mathbb{R}^n : x_i \geq 0 \quad \text{and} \quad \sum_{i=1}^{n} x_i = 1 \quad i \in V\}$$

and A is the symmetric affinity matrix. Here, $f(\mathrm{x})$ measures the cohesiveness of the corresponding cluster. Thus, a maximal cohesive of cluster corresponds to the solution of program (1), and the support set of vector x is defined as the set of indices of its positive components, i.e. $\sigma(\mathrm{x}) = \{i \in V : x_i > 0\}$, which is equivalent to a dominant set. The so-called replicator equation which is originated from evolutionary game theory can be used to solve program (1):

$$x_i(t+1) = x_i(t)\frac{(A\mathrm{x}(t))_i}{\mathrm{x}(t)^T A\mathrm{x}(t)} \tag{2}$$

Dominant set clustering (DSC) is a pairwise clustering framework centered around the notion of dominant set. The samples for clustering can be deemed as the vertices in an undirected edge-weighted graph with no self-loops, and the weights represent the similarities between the linked vertices. The process of DSC is shown in Algorithm 1. The procedure iteratively finds a dominant set and then removes it from the edge-weighted graph G until all the vertices have been assigned to the clusters.

Dominant set can be used for classification after the clustering process. Pavan *et al.* [5] formulate a dominant set fast assignment approach to classify the out-of-sample instances as shown in Algorithm 2. Where $\bigcup_{k=1}^{K}\{S^k, \mathrm{x}^k, f(\mathrm{x}^k)\}$ is the output of DSC, and the affinity vector α denotes the similarities between a new sample T^{new} and n grouped samples. The output $\hat{k}$ stands for the "nearest" cluster which T^{new} belongs to.

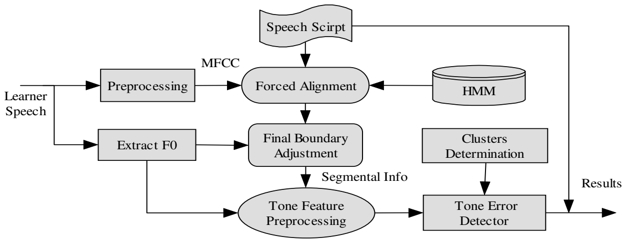

Figure 1: *Block-diagram of the tone error detection approach.*

4. Tone error detection via dominant set

Overview of our proposed tone error detection approach is illustrated in Figure 1. The detection procedure starts from the F0 extraction and viterbi decoding (forced alignment) with HMM-based recognizer to find out the final boundary. And then the final boundary is adjusted by the local F0 contour. Afterwards, F0 postprocessing and tone feature normalization are implemented. At last, the tone features are fed into the dominant set based tone error detector to get the detection results.

4.1. Final boundary adjustment

In real conditions, due to the acoustic differences between the strong accented database and the HMM model trained by general Mandarin database, the forced alignment can not scale well on the final domain. Figure 2 illustrates the primary warps of the forced alignment: the start point of the final is located at B_1 or B_2, and the end point of the final is located at C_1 or C_2. It is obvious that these warps can be modulated by the local F0 contour. In order to locate the final domain more accurately, we use the rule-based method as follows: 1) If the start point is at B_1, shift B_1 to B until the beginning of a voiced segment; 2) If the start point is at B_2, shift B_2 to B m frames within the voiced portion, we set $m = 3$; 3) If the end point is at C_1, shift C_1 to C until the last frame of the voiced segment; 4) If the end point is at C_2, shift C_2 to C until the last frame of a voiced segment. By using these rules, we can approximately locate the final domain of a syllable, as interval (B, C).

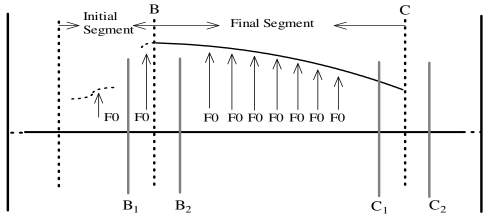

Figure 2: *Final boundary adjustment and F0 postprocessing.*

It is notable to mention that some of the voiced initials such as: $/l/, /m/, /n/$ can provide effects on the tones. As can be seen by the dotted pitch curve in Figure 2. Here, an F0 contour postprocessing is employed to deal with these cases. The postprocessing includes: removing the octave jumps and curve smoothing, also, a second-order polynomial $y = ax^2 + bx + c$

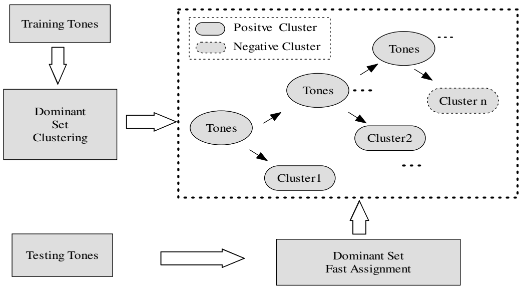

Figure 3: *Dominant set based tone error detector.*

is chosen to fit the F0 contour on the final domain using the least square criteria.

4.2. Tone feature normalization

After the postprocessing, F0 in Hertz is transformed into semitone scale, i.e. $F0_{semitone} = 50+12log_2(F0_{hertz}/320)$. The logarithm attempts to represent F0 in conformity with the human-like perception scale. The F0 contour on the final domain is sampled into k points, and we set $k = 10$. To reduce the pitch range differences among speakers, two methods: mean normalization (MN), mean and variance normalization (MVN) are explored:

$$\text{MN} : \hat{\text{f}} = \text{f} - \bar{\text{f}} \tag{3}$$

$$\text{MVN} : \hat{\text{f}} = \frac{\text{f} - \bar{\text{f}}}{s} \tag{4}$$

where $\bar{\text{f}}$ and s denote the mean and standard derivation of F0 respectively.

4.3. Tone error detector based on dominant set

The tone error detector consists of two parts as illustrated in Figure 3: 1) For each canonical tone, using the DSC to produce the positive and negative tone clusters, where the positive clusters stand for the tone pronunciations are right and the negative clusters stand for the tone pronunciations are wrong; 2) The testing tones are classified to their "nearest" cluster by using the dominant set fast assignment (DSFA) according to Algorithm 2. In this work, four tone error detectors are constructed corresponding to the four tones in Mandarin.

The DSC procedure starts from the calculation of similarity matrix. Here, similarity matrix is defined as $A = (a_{ij})$, and $a_{ij} = e^{-d_{ij}^2/\sigma}$, where d_{ij} represents the l_2 distance between tone i and tone j, and $\sigma > 0$ is a scaling factor which affects the decreasing rate of a_{ij}, we set $\sigma = 1.25$ as an empirical value. In the clustering process, DSC peels off a tone cluster from the similarity graph at each iteration. This peeling strategy generates tone clusters increasingly and induces that the similarity matrix becomes smaller and smaller. To avoid producing tiny clusters, we choose to terminate the clustering process when more than 95% tones are clustered, then the unprocessed tones are grouped to the formed clusters via DSFA.

By clustering, different styles of the tone errors can be collected, which facilitate the tone error detection. It is notable to say that the consistency inside a cluster is expected to be high enough so that the positive or negative tone clusters can be exactly determined, which also guarantee the promising performance on the tone error detection, and that DSC is appropriate for this application because this approach simultaneously emphasizes on internal homogeneity and external inhomogeneity.

5. Experimental results and analysis

5.1. Database and evaluation measures

The following experiments are performed on the database which consists of 7000 single syllable words pronounced by 70 persons (35 male and 35 female) from Xinjiang Uyghur Autonomous Region. The database is carefully designed in order to be consistent with Putonghua-Shuiping-Ceshi (PSC) — a national test to evaluate the proficiency of spoken Mandarin. We split the database into the training dataset and testing dataset, 60% for training and 40% for testing. Two native human experts are invited to label the tone errors in the dataset. To evaluate the performance of clustering, we utilize the average cluster consistency (ACC) as equation:

$$ACC = \frac{\sum_{i=1}^{K} N_i}{N_T} \tag{5}$$

where K and N_T denote the number of clusters and total samples, N_i is the number of right or wrong samples judged by human expert belong to a corresponding positive or negative cluster i. Meanwhile, to measure the performance of tone error detection, we utilize the Cross-Correlation (CC) between two detection results. CC is a very suitable criteria to validate the tone error detection algorithm [9]. It can be calculated by equation:

$$CC_{j1,j2} = \frac{\text{y}_{j1}^T \text{y}_{j2}}{\|\text{y}_{j1}\|_E \|\text{y}_{j2}\|_E} \tag{6}$$

Here, $\|\text{y}\|_E = \sqrt{\sum_{i=0}^{N-1} y(i)^2}$ is the standard Euclidean distance. The elements in judgement vector y_{j1} and y_{j2} are 0 or 1, where 0 denotes the tone pronunciation is right and 1 for wrong. If both of the judgements for a segment are right, they are ignored in order to put emphasis on error pronunciation [9].

5.2. Clustering performance analysis

The experiment is carried out on the training dataset to analysis the consistency by DSC. By reason of the dynamic pitch ranges among speakers differ greatly, tone normalization on three levels are taken into consideration including: by gender, by person, by syllable. Also, the mean normalization (MN), mean and variance normalization (MVN) are explored. Towards the local F0 fluctuation, we only take use of the MN. Upon that, five normalization methods are introduced. The consistency between human experts (HM VS HM) and the ACC of DSC are shown in Figure 4. Several interesting observations can be made: 1) MVN is not always better than MN, it can be seen from the figure that MN achieves higher consistency than MVN in the experiment; 2) MN by syllable provides better ACC than by gender and by person, which indicates that the local F0 fluctuation is more beneficial for normalization; 3) It can be seen that

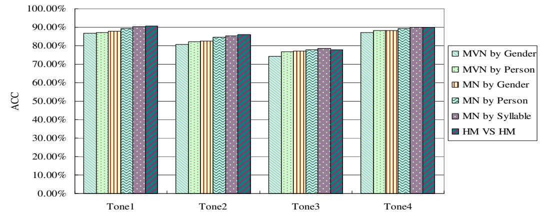

Figure 4: *Comparison of ACC with five normalization methods.*

Table 1: *CC between two human experts.*

CC	Tone1	Tone2	Tone3	Tone4	Average
HM VS HM	0.88	0.84	0.76	0.89	0.84

Table 2: *CC between human expert and machine, I stands for the approaches without final boundary adjustment and II stands for the approaches using final boundary adjustment*

	CC	Tone1	Tone2	Tone3	Tone4	Average
I	K-means VS HM	0.83	0.76	0.68	0.80	0.77
I	DS VS HM	0.86	0.81	0.74	0.86	0.82
II	K-means VS HM	0.85	0.79	0.72	0.82	0.80
II	DS VS HM	0.87	0.83	0.77	0.89	0.84

the ACC of clustering reaches to the consistency between two human experts, and this suggests that the proposed scheme can be accurate enough for practical applications.

5.3. Experimental results of tone error detection

Table 1 shows the CC between professional evaluators. It can be seen that some of the tones are hard to differentiate even by human experts. The CC between human experts is used as the reference for tone error detection approaches by machine.

We introduce the classical k-means and the proposed dominant set (DS) based approach to the tone error detection experiment. Here, MN by syllable is utilized for normalization. It is worth noting that one shortcoming of the k-means is that the number of clusters must be specified before clustering. For the sake of comparison between different methods, we choose the number of cluster for k-means: $N_k = N_d$, and N_d is the cluster number automatically determined by DSC. The CCs between tone error detection approaches and human expert are listed in Table 2. It shows that the final boundary adjustment can improve the performance in the process of detecting tone errors, the CCs increase 0.02 and 0.03 respectively in DS and k-means based approach. Obviously, DS based approach exhibits its efficiency in the experiments. By comparison with k-means, it yields more satisfying performance, which means that it is more applicable for acquiring of the data distribution in the "tone error space". With an average CC as high as 0.84, DS based approach reaches to the average CC between humans.

5.4. Experimental results analysis and discussion

The proposed dominant set based approach has several advantages. In practice, to determinate whether a cluster is positive or negative, it just requires a small number of samples due to the reliable consistency in a cluster, then can be further validated by F0 pattern of the cluster center, therefore, this strategy can save lots of tone annotations which are time-consuming and costly. Moreover, it yields better performance compared with the classical k-means based approach, and the number of clusters can be dynamically determined during the clustering, which is more applicable to tone error detection. Finally, by means of clustering, different categories of the F0 contours are collected. Some of the typical F0 modifications in the four standard tones are as follows: 1) In tone-1, there exists a small rising or falling in the forepart of the F0 contour; 2) In tone-2, there exists a flat or small falling in the forepart of the F0 contour; 3) In tone-3, the beginning of F0 is too high or the F0 rising in the end part is not enough; 4) In tone-4, there exists a flat or rising in the forepart of the F0 contour or a small rising at the end of the F0 contour.

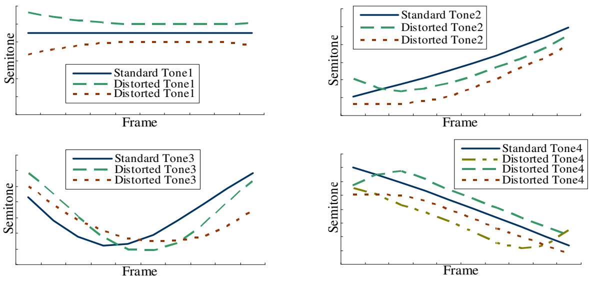

Figure 5: *Special F0 contours of the four tones in the database.*

The special F0 contours are illustrated in Figure 5. In general, these distortions can be tolerated by the perception of human listeners. However, using the model based approaches such as GMM or HMM, the tone categories of these special F0 contours are not easy to be determined, and a normal approach is with the help of using probability score and setting threshold, but the performance is more likely to be threshold-dependent. By utilizing our approach, these special F0 contours can be well partitioned into different clusters, which facilitate the exact identification of the F0 pattern and provide more informative feedback as a tutor.

6. Conclusions

In this paper, we propose a new approach based on dominant set in the application of tone error detection in strong accented Mandarin. The proposed scheme is computationally simple, and achieves satisfying performance with an average CC between human and machine as high as 0.84, attains that between humans. The success of our proposed scheme on isolated syllables has inspired exploration of tone mispronunciation detection in strong accented continuous speech as a topic of future work.

7. References

[1] M. Pavan and M. Pelillo, "A new graph-theoretic approach to clustering and segmentation", in Proc. CVPR, pp. 3895-3900, 2003.

[2] Witt, S., "Use of Speech Recognition in Computer-Assisted Language Learning", PhD thesis, Cambridge University Engineering Department, Cambridge, UK, 1999.

[3] F. Pan, Q. Zhao, and Y. Yan, "Improvements in Tone Pronunciation Scoring for Strongly Accented Mandarin Speech", Proc. ISCSLP, 603-608, 2006.

[4] M. Pavan and M. Pelillo, "Dominant Sets and Pairwise Clustering", IEEE Trans PAMI, VOL 29, pp. 167-172, 2007.

[5] M. Pavan and M. Pelillo, "Efficient Out-of-Sample Extension of Dominant-Set Clusters", Advances in Neural Information Processing Systems, vol. 17,pp. 1057-1064, 2005.

[6] Franco, H., Neumeyer, L., Kim, Y., Ronen, O., Bratt, H., "Automatic Detection of phone-level mispronunciation for language learning", in Proc. Eurospeech, pp. 851-854, 1999.

[7] Ito, A., Lim, Y., Suzuki, M., Makino, S., "Pronunciation Error Detection Method based on Error Rule Clustering using a Decision Tree", in Proc. EuroSpeech, pp. 173-176, 2005.

[8] Zhang, L. Huang, C. Chu, M. Soong, F. Zhang, X. & Chen, Y. "Automatic Detection of Tone Mispronunciation in Mandarin", Proc. ISCSLP, 590-601, 2006.

[9] Si Wei, Hai-Kun Wang, Qing-Sheng Liu, Ren-Hua Wang "CDF-matching for automatic tone error detection in mandarin call system".proc. ICASSP, 205-208, 2007.