TONE VARIATION MODELING FOR FLUENT MANDARIN TONE RECOGNITION BASED ON CLUSTERING

Wan-Yi Lin

Graduate Institute of Communication Engineering, National Taiwan University
Taipei, Taiwan, Republic of China
mfan@speech.ee.ntu.edu.tw

ABSTRACT

Tone recognition for fluent Mandarin speech has always been a very difficult problem, because the complicated tone behavior is difficult to analyze. In this paper, a new method of modeling tone variation for fluent Mandarin tone recognition by clustering training data into few subsets and weighting the likelihood computed by inter-syllabic features [6] is proposed. Experimental results indicate that the tone recognition accuracy can be improved significantly by this method and one modification of the method is robust and have smaller computing. Our tone variation modeling method is shown to improve the recognition rate from 91.3% to 95.2%.

1. INTRODUCTION

Chinese is a typical tonal language. Each Chinese character is pronounced as a monosyllabe. There is a total of about 416 phonologically allowed syllables in Mandarin Chinese, if the difference in tones are disregarded. But if the differences in tones are considered, there are about 1345 syllables. Accurate tone recognition is greatly helpful for Chinese recognition systems, because the tones in syllables have lexical meanings in Chinese. There exist four lexical tones (tone 1 to tone 4) plus a neutral tone (tone 5) in Mandarin. These numbers are slightly different for different dialects.

Many substantial efforts have made in tone recognition for Chinese syllables, even including that for Cantonese, [1], although the major efforts were for Mandarin. For Mandarin, tone 5, the neutral tone, appears slightly less frequent than the four lexical tones and is often seriously confused with the four lexical tones. Some of the studies only tried to recognize the four lexical tones. Vector quantization, HMM, and neural networks have been applied for tone recognition over since roughly fifteen years ago [2, 3]. In recent years, the context dependency [4] as well as the tone nuclei have also been introduced [5].

One of the biggest difficulties of tone recognition is that tone varies seriously, and the reasons that cause tone variations are too many and too complicated to solve, such as speakers and contexts. Even spoken by the same speaker, syllables with the same tone usually don't have similar characteristics. Thus a good method to model tone variations is necessary.

In the previous work [6], we proposed a pitch extraction method and a set of four inter-syllabic features for tone recognition. In this paper, we proposed a method for modeling tone variation based on these features and the weighting of inter-syllabic features. It is easy to think that if we find some features that the distribution of all syllables with the same tone on these features is focused, then these features can model tone variations well. But in fact, there are no such features recently, so we model tone variations by clustering the training data into few subsets. Each subset is used to train a tone model, which means that this model represents a kind of variation of the tone.

We proposed an improved clustering algorithm based on Partition around Medoids (PAM) [7]. By this algorithm, the number of subsets after clustering is data-driven, which is proper for our application. In fact, we don't know how many subsets the training data should be divided into at the beginning, because the degree that every tone varies at is not the same. The algorithm is used to divide the training data of a tone into few subsets based on inter-syllabic features. After this, these subsets are refined by the intra-syllabic features. We proposed two methods of refining subsets of training data by clustering based on the intra-syllabic features to refine the subsets derived above. One is the maximum HMM likelihood, which uses all the 42-dim intra-syllabic to cluster. The other is curve fitting, which uses only logarithms of f0, Δf0, and delta Δf0 to cluster.

The experimental results show that the clustering procedure can divide the training data properly, and the tone variation modeling method by dividing the training data can significantly improve the recognition rate, especially the tones that originally have low recognition rate. This means that the tone variation modeling method can reduce the tone confusions significantly.

The organization of the rest of this paper is as follows. In section 2 we present the tone variation modeling method by clustering training data and weighting the likelihood computed by inter-syllabic features. While the experimental results are then described and analyzed in section 3, and the conclusions are made in section 4.

2. TONE VARIATION MODELING

Because tone features vary so great, it is more efficient to model tone variations based on the existing features than to use new features or transformations. We model tone variation by clustering training data based on tone features, that is, clustering training data of a tone into few groups, each represents a kind of variation of the tone, and then every subset is used to train a model. The flowchart of training the proposed tone variation models is in Fig. 1. We use two categories of tone features: 42 intra-syllabic features (39-dim MFCC and logarithms of f0, Δf0, and delta Δf0) and four inter-syllabic features [6]. We first cluster training data based on inter-syllabic features and then refine the subsets by intra-syllabic features. This procedure of

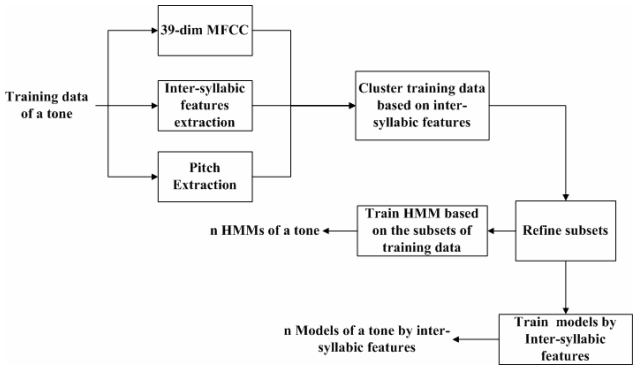

Fig. 1: Flowchart of training models.

clustering will save lots of computing. We proposed two methods of refining: maximum likelihood for HMM models and f0, Δf0, and delta Δf0 curve fitting.

2.1. Tone features and feature weighting

In the previous work [6] we proposed four inter-syllabic features for tone recognition. The four features are the ratio of duration of adjacent syllables, average |Δf0|, |f0|, and maximum f0 difference of a syllable. These features are used to model neutral tone and tone sandhi.

We add the likelihood computed by the inter-syllabic features between two syllables in the Viterbi search (that is, change the transition probability between two HMMs, which is 1 by default). To make the influence of inter-syllabic features the same as intra-syllabic features, the likelihood should be properly weighted, because the logarithm of likelihood of a syllable computed by the intra-syllabic features is added along the frames, but that computed by the inter-syllabic features is only added one time per syllable.

2.2. Cluster training data by inter-syllabic features

Because clustering time sequence like the likelihood of every frame of a syllable computed by intra-syllabic features needs huge computing and the optimal algorithm hasn't been developed, we start with clustering inter-syllabic features.

Considering a syllable's likelihood computed by the four inter-syllabic features as a point in a four-dimension space, we divide the points of one tone into few subsets by Partition around Medoids (PAM) [7]. We choose PAM because it is more robust than k-means in the presence of outliers, which usually occur in the distribution of tone features. Traditional PAM can't work without giving the number of subsets at the beginning, but in this case, we don't know what the proper number of subsets is. We solve this problem as follows:

n is the number of subsets, D_i is the average Euclidean distance between every point in the ith subset and the medoid of this subset: Assume

$$n = 2^m + k \quad m, k \in N \cup \{0\} \quad (1)$$

$$\mathrm{Dav_n} = \frac{1}{n}\sum_{i=1}^{n} D_i \quad (2)$$

Where $\mathrm{Dav_n}$ is the average of D_i, we can consider $\mathrm{Dav_n}$ as the reciprocal of the "average concentration degree" of the subsets. To avoid dividing a subset into two subsets that shouldn't be further divided, we set a criterion: If we divide the points into n subsets, and the average concentration degree of these n subsets is more than n times that of only one subset, we call this division proper. It can be easily inferred that if $n=n_1$ is proper and $n_2>n_1$, the criterion of $n=n_2$ being proper is:

$$\frac{Dav_{n_2}}{Dav_{n_1}} \le \frac{n_1}{n_2} \quad (3)$$

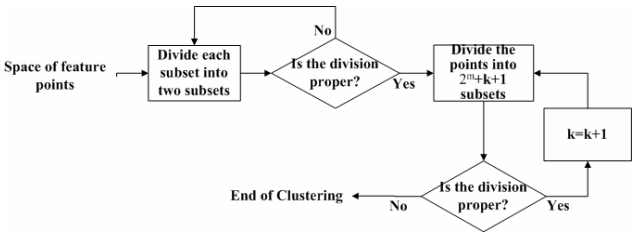

Fig. 2: Flowchart of clustering syllable distribution based on inter-syllabic features.

Three steps for clustering the feature points were developed, and the flowchart of this procedure is shown in Fig. 2:

Step 1: Divide each subset into two subsets. (If this is the first iteration, divide the points into two subsets.)

Step 2: Check whether the division is proper by (3). If yes, go back to Step 1; if not, it means we find m in (2) and go to Step 3. That is, the number of subsets we want to find is between 2^m and 2^{m+1}.

Step 3: Now we have to find k in (1). First we assume k=0, then divide the points into 2^m+k+1 subsets, check if the division is proper by (3). If yes, $k=k+1$; if not, the current k is the answer.

2.3 Refine subsets of training data by maximum HMM likelihood

We refine the subsets derived form Sec.2.2 by the whole 42-dim intra-syllabic features, based on maximizing the likelihood of syllables in the subsets of HMM models. It is clear that if every syllable used to train a HMM has maximum likelihood with this HMM, the subsets are the most fit to the tone variations of training data.

Now, after clustering in Sec. 2.2, the syllables with same tone are divided into few subsets. We train each subset as a HMM, which means that one tone has few HMMs, and do inside-test. After inside-test, some syllables are mis-recognized (that is, the syllable used to train a HMM being recognized as another model after inside-test), then re-classify these syllables into HMM models which maximize the likelihood of these syllables, which means that the later HMM is more similar with the syllable than the former one.

Do the above steps iteratively until the number of mis-recognized syllables converges. The final subsets are optimal for representing tone variations of training data.

2.4 Refine subsets of training data by curve fitting

Besides the above clustering method, we also proposed another method by curve fitting. The main idea is that although MFCC has some tone information, f0 carries more information about tone and the distribution of f0 is more correlated with tone. So refining the subsets by f0 curve fitting can avoid the disturbance

caused by MFCC and noise, for MFCC is less robust than f0 we derived in [6].

Consider a four-dimension space, and the four axes are f0, Δf0, delta Δf0, and time, respectively. Every frame of a syllable is a point in this space, and, after putting all the points of the syllables in the space, the syllable can be considered as a curve segment in the four dimensional space. Start clustering with the subsets and medoids derived after Sec. 2.2, three steps of refining the subsets were developed:

Let C_j^i be the *j*th curve segment in the *i*th subset of a tone, $P_j^i(m)$ be the *m*th point of C_j^i, n_j^i be the number of points that C_j^i consist of, and C_{med}^i be the curve segment of the medoid of the *i*th subset.

Step 1: For the *i*th subset, normalize the length of every curve segment (assume C_j^i) to the length of its medoid C_{med}^i. Keep the height of these points: $P_j^i(1)$ (the first point of C_j^i), $P_j^i(n_j^i)$ (the last point of C_j^i) and the point with extreme value other than $P_j^i(1)$ and $P_j^i(n_j^i)$. (If there is no extreme value in C_j^i, keep the height of $P_j^i(1)$ and $P_j^i(n_j^i)$ only) Then adjust n_j^i to n_{med}^i.

Step 2: Compute the Euclidean distance D_j^i between C_j^i and C_{med}^i, let $D^i = \{D_j^i \mid \forall j\}$

$$D_j^i = \frac{1}{n_{med}^i} \sum_{k=1}^{n_{med}^i} d(P_j^i(k), P_{med}^i(k)) \qquad (4)$$

Step 3: Compute the mean and variance of D^i for all *i*. If a subset has the largest mean and it is larger than two times the average of other means, the medoid of this subset should be recomputed. Otherwise, re-classify the points in the *i*th subset that D_j^i >mean of D^i plus variance of D^i for all *i*.

Iteratively do the above steps, until the means of all D^i converges.

3. EXPERIMENTAL RESULTS

We use the 30-hour HUB4 (Train) as the training corpus and the one-hour HUB4 (Evltest) as the test corpus, which is a Mandarin broadcast news database. At the beginning, we used the 13 MFCC parameters plus the logarithm of pitch values and their first and second derivatives, thus a total of 42 features to train the tone models with different context conditions (175 context dependent HMM models) and the four inter-syllabic features, each used to train a four-mixture Gaussian distribution. Then we use the tone variation modeling method proposed in sec 2, that is, cluster the training data of each context dependent model into few subsets, and each subset is used to train a new model. The flowchart of training models is showed in Fig. 1. For example, the training data of a context dependent tone model 1-1-2 (tone sequence 1-1-2) is divided into three subsets, and each is used to train a new HMM and an inter-syllabic model. After all, there will be three HMMs for the context dependent tone 1-1-2. The number "three" in the above example is determined automatically by the algorithm proposed in Sec. 2.2. In our experiments, the number of the models of a tone decided by the algorithm is between one and five, and tones that vary seriously, for example, 2-3-3 and 1-3-4, have bigger numbers. Another interesting phenomenon is that the training data of neutral tone, which only has few significant enough features and varies too great and irregularly, was usually divided into only two subsets, or even can't be divided by the proposed algorithm. The result is reasonable because the variations are not meaningful enough to train a new model. These phenomena show that the proposed auto-terminated PAM works properly.

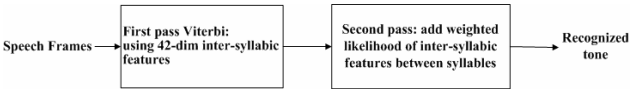

Fig. 3: Flowchart of tone recognition.

The flowchart of our tone recognition method is in Fig. 3. For each input speech sentence, we first extract the 42-dim intra-syllabic features and run the first-pass recognition. After that, the syllable boundaries are also derived, so we can extract inter-syllabic features. Then do the second-pass recognition by adding the weighted likelihood of the syllables computed by these inter-syllabic features between adjacent syllables.

Fig. 4 shows the weighting of likelihood computed by inter-syllabic features versus recognition rate for the four conditions: model tone variation by clustering training data without refining, refined by maximum HMM likelihood, refined by curve fitting, and without tone variations. The baseline (without tone variations) has the highest recognition rate in comparison to itself when the weighting is 0.6 times frame number but lower recognition rate when weighted by frame number. Because the likelihood of inter-syllabic feature is a kind of compensation for the variation of intra-syllabic tone features, it should be emphasized to be as influential as intra-syllabic features, but when over-emphasized, the recognition rate will be reduced.

Tone variation modeling without refining and refined with maximum HMM likelihood show similar trend versus the weighting, but the weighting with that the former method reaches the highest recognition rate in comparison to itself is less than that of the later. This is because the tone variation modeling without refining is only based on inter-syllabic features, so the weighting of inter-syllabic features should be reduced. Tone variation modeling refined by curve fitting reaches the highest recognition rate when the weighting equals to 3, and the recognition rate varies most seriously among the four methods. Because both inter-syllabic features and curve fitting are derived from f0 values, the variation they model is similar, which make the optimal weighting smaller, and if over-weighted, the recognition rate degrades seriously.

Fig. 5 shows the robustness of the four tone recognition method, and the noise we used is Gaussian noise, choosing clean condition and SNR= 15, 5, 2. It is clear that modeling tone variations refined by curve fitting is the most robust among the four methods, which is because that the MFCC is affected by noise more seriously than the pitch we extract. This also shows that although tone variation modeling refined by curve fitting performs a little worse than by maximum HMM likelihood, it is more robust, and the computation is much less. Modeling tone variations refined by HMM likelihood has the highest recognition rate in clean condition, but when SNR is lower than 15, its recognition rate is less than without refining and refined

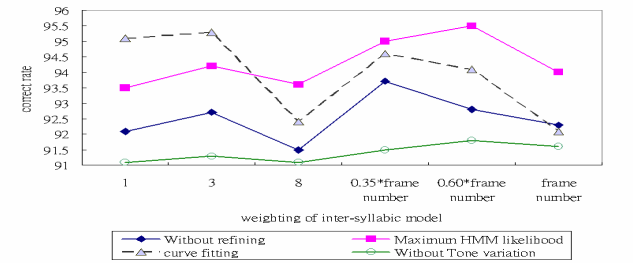

Fig. 4: Weighting of likelihood computed by inter-syllabic models versus recognition rate.

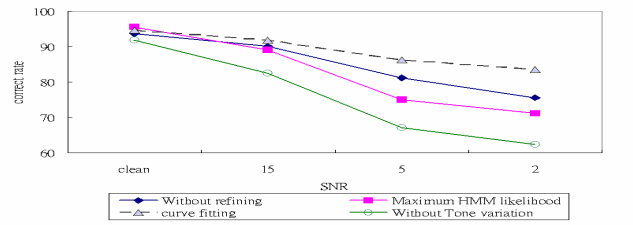

Fig. 5: Recognition rates of the four methods under Gaussian noise.

by curve fitting. This is because that this refining method depends much more on MFCC than f0, and the MFCC is less robust than f0.

Table 1 is the confusion table of baseline and tone variation modeling refined by HMM likelihood, both with the weighting of 0.6 times frame number. The tone variation modeling method works best when the confusion rate is high. For example, the rate that tone 1 be recognized as tone 5 is reduced from 3.01% to 1.29%. The tone variation method works especially in the case tone 3 recognized as tone 2, and the confusion rate is reduced from 5.8% to 1.7%; this is because that in this case, the variation happens frequently and regularly.

We also compare the proposed method with other tone recognition methods: VQ [2], HMM [3], Tone Nuclei [5], and our proposed method. The Experiment is five-tone recognition and the result is shown in Fig. 6. Although our proposed method has the highest recognition rate, the other three methods are focused on the four lexical tones in Mandarin, so their recognition rate of tone 5 is low. But we also do some efforts to solve tone 5, and the portion of syllables of tone 5 in HUB4 (about 5%) is larger than in other corpuses, so this comparison is somewhat unfair. So Fig. 6 is only a reference.

4. CONCLUSIONS

In this paper, we present a method of modeling tone variations to recognize the tones in continuous Mandarin speech. Experimental results verified that the two modifications of the method work very well, and one of them also performs well under noisy conditions.

5. ACKNOWLEDGEMENTS

This paper would not be possible without the inspiring comments and solid education from my advisor, Dr. Lin-shan Lee.

(a)

	Recognized Tone				
	Tone 1	Tone 2	Tone 3	Tone 4	Tone 5
Tone 1	92.5	2.13	1.36	1	3.01
Tone 2	1.1	95.8	1.5	0.5	1.1
Tone 3	0.8	5.8	91.2	1.9	0.3
Tone 4	2.5	0.6	1.7	94.8	0.4
Tone 5	9.5	6.5	6.2	5.1	79.2

(b)

	Recognized Tone				
	Tone 1	Tone 2	Tone 3	Tone 4	Tone 5
Tone 1	95.5	1.2	1.11	0.9	1.29
Tone 2	0.8	96.8	1	0.4	1
Tone 3	0.6	1.7	96.4	1.1	0.2
Tone 4	2.1	0.6	1.5	94.9	0.9
Tone 5	4.2	5.7	5.4	4.2	86.2

Table 1: Confusion matrix of (a) without tone variations, (b) tone variation modeling refined by maximum HMM likelihood.

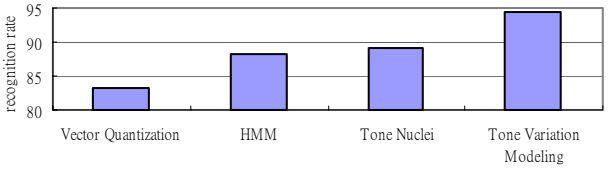

Fig. 6: Recognition rates of several tone recognition methods.

6. REFERENCE

[1] T. Lee, P.-C. Ching, L.-W. Chan, Y.-H. Cheng, and B. Mak "Tone recognition of isolated Cantonese syllables," *IEEE Tras. Acous Speech Audio Processing,* Volume: 3 Issue: 3 , May 1995, pp. 204 -209

[2] S. H. Chen and Y. R.Wang, "Vector Quantization of Pitch Information in Mandarin Speech," *IEEE Transactions on Communications*, Vol.38, pp. 1317-1320, 1990.

[3] W.-J. Yang, J.-C. Lee, Y.-C. Chang and H.-C. Wang, "Hidden Markov Model for Mandarin Lexical Recognition", *IEEE Trans. on ASSP,* Vol. 36, No7, July 1988, pp.988-992.

[4] L.-S. Lee **"**Voice dictation of Mandarin Chinese,**"** *IEEE Signal Processing Magazine,* Volume: 14 Issue: 4 , July 1997, Page(s): 63 -101

[5] J. S. Zhang and K. Hirose, "Tone Recognition of Chinese Continuous Speech Using Tone Critical Segments", *Eurospeech*, Budapest, Hungary, Sept. 1999, pp.879-882.

[6] W.-y. Lin and L.-s. Lee," Improved Tone Recognition for Fluent Mandarin Speech Based On New Inter-Syllabic Features and Robust Pitch Extraction", *ASRU 2003*

[7] J. Han and M. Kamber, "Data Mining: concepts and techniques", *San Francisco: Morgan Kaufmann Publishers*, 2001