ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/221479840

Overlapped di-tone modeling for tone recognition in continuous Cantonese speech.

Conference Paper · January 2003

Source: DBLP

CITATIONS
7

READS
50

3 authors:

Yao Qian
Microsoft
69 PUBLICATIONS 1,439 CITATIONS
SEE PROFILE

Tan Lee
The Chinese University of Hong Kong
210 PUBLICATIONS 1,844 CITATIONS
SEE PROFILE

Yujia Li
The Chinese University of Hong Kong
15 PUBLICATIONS 196 CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project Aphasia speech assessment View project

All content following this page was uploaded by Yao Qian on 13 October 2014.

The user has requested enhancement of the downloaded file.

Overlapped Di-tone Modeling for Tone Recognition in Continuous Cantonese Speech

Yao QIAN, Tan LEE and Yujia LI

Department of Electronic Engineering
The Chinese University of Hong Kong
Shatin, New Territories, Hong Kong
{yqian,tanlee,yjli}@ee.cuhk.edu.hk

Abstract

This paper presents a novel approach to tone recognition in continuous Cantonese speech based on overlapped di-tone Gaussian mixture models (ODGMM). The ODGMM is designed with special consideration on the fact that Cantonese tone identification relies more on the relative pitch level than on the pitch contour. A di-tone unit covers a group of two consecutive tone occurrences. The tone sequence carried by a Cantonese utterance can be considered as the connection of such di-tone units. Adjacent di-tone units overlap with each other by exactly one tone. For each di-tone unit, a GMM is trained with a 10-dimensional feature vector that characterizes the F0 movement within the unit. In particular, the di-tone models capture the relative deviation between the F0 levels of the two tones. Viterbi decoding algorithm is adopted to search for the optimal tone sequence, under the phonological constraints on syllable-tone combination. Experimental results show the ODGMM approach significantly outperforms the previously proposed methods for tone recognition in continuous Cantonese speech.

1. Introduction

The Cantonese dialect is spoken by tens of millions of people in Hong Kong and Southern China. It is one of the so-called tone languages that use the pitch of the speaker's voice to distinguish one lexical entry from the other. Cantonese is known of being very rich in tones and having relatively simple syllable structure. Tone identification plays an important role in automatic speech recognition of Cantonese. Indeed, building statistical language models for Cantonese appears to be difficult and ineffective since there exists little text materials for training. Many words in spoken Cantonese don't have commonly agreed written forms. This makes tone information even more contributive to speech recognition.

There have been many studies on using tone information in Chinese continuous speech recognition. A common approach is to establish tone-dependent acoustic models so that tone recognition is embedded into the process of recognizing basic acoustic units [1-4]. Pitch-related features such as F0 can be added as extra dimensions in the short-time acoustic feature vector. This approach has been shown to be effective for Mandarin tone recognition. However, pitch is a supra-segmental feature, which is spanned across segments and laid on top of a group of voiced segments. For unvoiced speech, F0 interpolation is inevitable to avoid discontinuity between voiced and unvoiced segments.

To better capture the supra-segmental characteristics, tone recognition can be done as a separated process in parallel to the recognition of acoustic units. In our previous work [5,6], an effective approach of explicit tone recognition was used for continuous Cantonese utterances. Each tone occurrence is represented by a pitch-related feature vector. The feature vectors are modeled with context-dependent hidden Markov models. With proper feature normalization, an accuracy of 66.4% was attained for speaker-independent tone recognition. It was also shown that such automatically recognized tones can be used to improve the performance of a large vocabulary continuous speech recognition (LVCSR) system of Cantonese.

As a matter of fact, Cantonese has a more complicated tone system than Mandarin. A total of nine citation tones are used in Cantonese, in comparison to four or five in Mandarin. In this paper, we propose to model Cantonese tones using overlapped di-tone modeling with Gaussian mixture models (ODGMM). This approach aims to model the Cantonese tones more precisely from phonetics and phonology theory point of view. Its effectiveness is also demonstrated by the results of tone recognition experiments.

2. Cantonese tone system

Like Mandarin, spoken Cantonese is seen as a string of monosyllabic sounds. Each Chinese character is pronounced as a single syllable that carries a specific tone. For examples, the syllable [si:] may correspond to the Chinese characters: 詩, 史, 試, 時, 市, or 是, depending on the tones.

Each Cantonese syllable is divided into an Initial part and a Final part. The Initial includes whatever is preceding the main vowel while the Final includes the main vowel and whatever follows it. The tone is manifested by the pitch of the voiced part of the syllable. Tones are used to convey lexical information. Therefore they are referred to as lexical tones. It is an essential feature that determines the meaning of a word.

Cantonese is said to have nine citation tones that are characterized by different pitch patterns as illustrated in Figure 1. The first six tones are carried by syllables that either have no consonant ending or end with /m/, /n/, /ng/. Whilst the syllables that carry the other three tones must end with unreleased stop consonants /p/, /t/, /k/. These tones are called "entering tone". The entering tones are widely regarded as abbreviated counterparts of the non-entering tones 1, 3 and 6 respectively. In many transcription schemes, including the LSHK scheme [7], only six distinctive tones are labeled by numerals 1 to 6.

There is a distinctive difference between the tone systems of Cantonese and Mandarin. Mandarin tone system tends to approximate a CONTOUR (or GLIDING-PITCH) system while Cantonese tone system seems to be close to a REGISTER (or LEVEL-PITCH) system. In a CONTOUR system, tones are characterized by the movement or glide of pitch. The contrast is among the patterns that are described as, for examples, falling, rising and fall-rise. In a REGISTER system, tones are characterized by their distinctive pitch levels. Typically there are two or three such levels and probably never more than four [8]. These levels are defined in a relative sense. Their absolute heights may shift from time to time. A high-level tone, for example, would be perceived as having relatively high pitch in comparison to its neighboring middle-level or low-level tones. Out of the six Cantonese tones, four tones (Tone 1, 3, 4 & 6) have either flat or slightly falling F0 patterns, which can be considered as different level-pitch tones. Perceptual tests revealed that middle-level tones uttered in isolation could be easily confused with high-level or low-level tones.

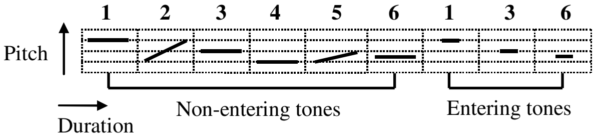

Figure 1: Tones in Cantonese (schematic description)

3. Overlapped di-tone modeling

Tone is a supra-segmental feature that can spread onto neighboring segments and beyond. Its minimal domain is a syllable. As stated in Section 2, the perception of Cantonese tones relies mainly on their relative pitch levels. Therefore, tone is not only a property of syllables, but also a property of cross-syllable and even larger units like words. Although most of the previously proposed methods could capture dynamic change of pitch, they don't reflect such change over beyond syllable boundaries. In particular, the relative deviation between neighboring tones is not taken into account.

To model Cantonese tones properly, a relatively wide window should be used for tone analysis. We propose to use the approach of overlapped di-tone modeling as depicted in Figure 2. The unit to be modeled covers two consecutive tones and there is an overlap between two adjacent di-tone units. In this way, a syllable carrying high, middle or low level tones would be clearly identifiable with the adjacent tone serving as an "anchor". The di-tone model captures the information about the relative deviation between the two tones in a straightforward way. The overlap is needed so that all transitions between adjacent tones are covered.

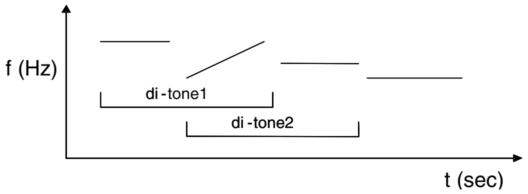

Figure 2: Overlapped di-tone units

The acoustic feature used for di-tone modeling is primarily the fundamental frequency (F0) extracted on a short-time basis from the voiced part of input speech. The number of short-time frames may be different from one di-tone unit to another. Therefore the extracted sequence of F0 values needs to be time-aligned to a fixed length. In this work, each di-tone unit is represented by a feature vector that contains 10 F0 values. Essentially each tone is divided evenly into 5 regions and each region is represented by the median value of linear regression over all F0 values in that region.

For each di-tone unit, a Gaussian mixture model (GMM) is trained with the 10-dimensional feature vector. The details will be given in Section 4.2.

4. Tone recognition with overlapped di-tone GMMs

In continuous speech recognition for Cantonese, Initials and Finals have been used as the fundamental units for acoustic modeling. With these acoustic models, speech segmentation at Initial-Final level is available as a by-product of the Viterbi decoding process. Such segmentation provides useful information for explicit tone recognition [5]. It not only provides the syllable boundaries but also indicates the voiced/unvoiced boundaries within individual syllables.

Using the proposed overlapped di-tone GMM (ODGMM), explicit tone recognition can be done as shown in Figure 3.

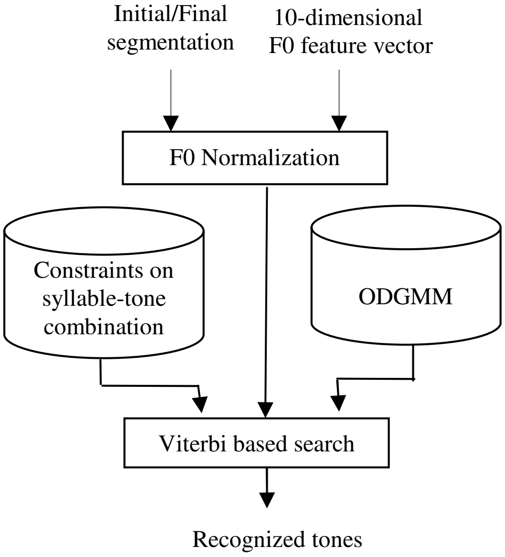
Initial/Final segmentation
10-dimensional F0 feature vector
F0 Normalization
Constraints on syllable-tone combination
ODGMM
Viterbi based search
Recognized tones

Figure 3: The proposed approach of ODGMM based tone recognition

4.1. F0 normalization

F0 is a highly variable acoustic feature. Speaker difference in F0 could be determined by a variety of factors, e.g. age, gender, dialectal background, health condition, education and personal style. Even for the same speaker's voice, the actual range of F0 changes from time to time. For accurate identification of tones, effective F0 normalization is necessary to minimize the undesirable fluctuations. In this research, F0 normalization is carried out as follows:

1. Derive an inter-utterance normalization factor (INF) that represents the underlying F0 level that the speaker is using in a particular utterance;
2. Derive a phrase curve that reflects the instantaneous F0 change over an utterance;
3. Derive an overall normalization factor (ONF) by combining the INF and the phrase curve to represent the underlying F0 level at a particular time instant in a particular utterance.

4. Divide the absolute F0 values by the corresponding ONFs.

In this study, INF is obtained as the mean of F0 over the entire utterance. In [9], it was proposed to describe the F0 movement in a Cantonese intonation phrase by a declining straight line, which is referred to as phrase curve. The slope of a phrase curve depends on the length of the phrase. In our work, we divide the utterances into three categories according to their lengths. The utterances in each category are assumed to share the same phrase curve pattern.

Given an utterance, the phrase curve is first shifted according to the INF. Then the ONF at any particular time in the utterance can be directly obtained from this shifted phrase curve.

4.2. Training of ODGMM

In a GMM, the probability distribution of an observed feature vector **x** is represented by a mixture of Gaussian density functions, i.e.

$$f(\boldsymbol{x}) = \sum_{k=1}^{K} \boldsymbol{c}_k \boldsymbol{N}_k(x; \mu_k, \Sigma_k) \qquad \boldsymbol{c}_k \ge 0 \quad , \quad \sum_{k=1}^{K} \boldsymbol{c}_k = 1$$

where c_k, μ_k and Σ_k are the Gaussian component weight, mean vector and covariance matrix respectively. The expectation-maximization (EM) algorithm is used for GMM parameters estimation. The preliminary estimation for mean vector and covariance matrix of GMM is done by a single iteration k-means clustering for all training data, with equal component weights. The number of components in each GMM is fixed at the beginning. With multiple iterations of EM, the number of components is gradually increased until the relative improvement on output probability becomes less than 1% or the number of iterations is larger than 500.

As stated in Section 3, the feature vector **x** for ODGMM is a 10-dimension vector of time aligned F0 values. There are a total of 42 di-tone units DT_i (including the 6 utterance-initial units). A GMM is trained for each of these units. The number of mixture components varies from 4 to 35, depending on the amount of training data. For tone recognition, the probability of DT_i given observation **x** is given as:

$$\Pr[\boldsymbol{DT}_{\mathrm{i}} \mid \boldsymbol{x}] = \frac{\Pr[\boldsymbol{x} \mid \boldsymbol{DT}_{\mathrm{i}}] \cdot \Pr[\boldsymbol{DT}_{\mathrm{i}}]}{\Pr[\boldsymbol{x}]} = \frac{f_{DT_i}(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{c}, \mu, \Sigma) \cdot \Pr[\boldsymbol{DT}_{\mathrm{i}}]}{\Pr[\boldsymbol{x}]}$$

4.3. Viterbi based search with phonological constraints

In Cantonese, the number of base syllables and tones are 637 and 6 respectively, while the total number of legitimate tonal syllables is 1614. In other words, more than half of the syllable-tone combinations are not allowed. If there is a LVCSR system running in parallel to produce hypotheses of syllables, ODGMM based tone recognition can take advantages of the phonological constraints for better performance. This is done by Viterbi based search to find the optimal path of tone sequence given a sequence of di-tone feature vectors. Details are given below:

- GMM can be seen as one state HMM. There is no state transition within the model. The observation probabilities of the models are used to guide the propagation of paths. If the combination of a tone and a particular syllable is not allowed, the observation probability of the relevant di-tone units will be assigned the minimum value so that it has very little chance of being included in the optimal path;
- Transition scores between models are dynamically assigned. If the first tone in the current di-tone unit has the same identity as the second tone in the preceding unit, the transition score will be assigned a large value. Otherwise, it will be made small;
- In backtracking, if two tones in overlapped region are the same, this tone will pop up. Otherwise, the tone with a higher prior probability will be selected.

5. Experiments and Discussion

The corpus used in our experiments is CUSENT, which contains 20,000 phonetically rich training utterances spoken by 34 male and 34 female speakers. The test data comprises about 1,200 unseen utterances sentences from 6 male and 6 female speakers.

The acoustic models are Initial/Final models with both left and right context dependency. Model training was done with HTK Version 3.1. The acoustic feature vector is composed of 12 Mel-Frequency Cepstral Coefficients (MFCC), Energy, and their first-order and second-order derivatives. Each Initial model is an HMM with three emitting states, while a Final model consists of either three or five emitting states, depending on its phonetic composition. Each emitting state is made up of 16 Gaussian mixture components. Decision-tree based state clustering is used to facilitate parameter sharing. The recognition accuracy for base syllable (tone-independent monosyllabic unit) is 79.52% for the 1,200 test utterances.

Table 1 shows the tone recognition results given by the ODGMM based approach. The overall accuracy is 72.92%. The best accuracy, 87.17%, has been attained for Tone 1, which appears to have the highest percentage of distribution among all tones. Tone 5, which is the least frequent tone, gets the lowest accuracy of 50.72%.

	Recognition Accuracy (%)	% of distribution
Tone1	87.17	24.23
Tone2	75.16	12.96
Tone3	66.93	16.84
Tone4	72.99	17.17
Tone5	50.72	7.36
Tone6	69.10	21.44
Overall	72.92	

Table 1: The result of tone recognition with ODGMM

We have also experimented with the approach of embedded tone recognition using the same training data and test data. In this approach, the units for acoustic modeling are the 20 Initials and 280 Tonal Final models (ITFM). The acoustic feature vector contains F0 and its derivatives, in addition to the MFCC and energy features. In this case, the recognition accuracy of base syllables is 79.55%. That is, the introduction of tonal models gives virtually no improvement on the base syllable accuracy.

Table 2 shows the tone recognition results given by embedded tone recognition approach (ITFM) and the context-dependent uni-tone modeling approach (CDUTM) as described in [5]. It can be seen that ODGMM outperforms ITFM and CDUTM by 4.88% and 6.52% absolutely.

Compared with ITFM, ODGMM improves the accuracy for all tones except tone 2, in which the most significant improvement is 13.39% (Tone 1).

On the other hand, ODGMM significantly outperforms the CDUTM approach by 6.52% absolute difference in overall recognition accuracy. It must be noted that, in the CDUTM approach, tone recognition is done as an independent process without any phonological constraints. The advantage of ODGMM is partially due to the contribution from the phonological constraints. Specifically, the recognition accuracy of middle level tones, namely Tone 6 and Tone 3, are greatly increased by 21% and 11.63% respectively.

A more in-depth examination on the confusion matrices given by the three approaches reveals the following:

- ODGMM largely resolves the confusion among the level tones (Tone 1, 3 and 6). This may be due to that the ODGMM has been designed to fit the characteristics of level tones. This advantage is therefore directly reflected by the recognition accuracy for level tones.
- It was observed that about 67% of the recognition results (including both correct and erroneous recognition) produced by the ITFM and the ODGMM approaches are matched. This suggests that ITFM and ODGMM might be complementary to each other. We may consider using them in a multi-pass recognition process so as to attain further improvement on tone recognition performance.

	ITFM(%)	CDUTM(%)
Tone1	73.78	83.70
Tone2	75.55	74.00
Tone3	60.38	55.30
Tone4	71.89	75.30
Tone5	50.58	56.40
Tone6	66.21	48.10
overall	68.04	66.40

Table 2: Tone recognition results with the approaches of ITFM and CDUTM

In our previous approach (CDUTM) [6], the experiment results showed that the use of normalized energy could lead to an improvement of 0.94% on the tone recognition accuracy. Syllable duration is another feature that can be considered. It is available from the acoustic recognition process. We have attempted to investigate the correlation between tone and syllable duration. The syllable duration are normalized by

$$Dur_{nor} = \frac{Dur_{real}}{Dur_{mean}}$$

where Dur_{mean} is the average syllable duration in a particular sentence. The average normalized duration for each type of tone is showed in Table 3, which indicates that the duration of Tone 5 is the shortest and Tone 2 is the longest.

	Tone1	Tone2	Tone3	Tone4	Tone5	Tone6
Dur	0.9993	1.0653	1.0415	1.0288	0.93	0.9965

Table 3: Average normalized duration of the six tones

With the normalized duration being added as an additional component of the feature vector, the overall tone recognition accuracy with ODGMM is slightly increased to 73.54%.

6. Conclusions

Tone information is extremely important to human perception of natural speech. Recent research has shown that tone information can serve as useful knowledge source for automatic speech recognition of tonal languages. Our previous work also verified that reliable tone information could positively contribute to Cantonese LVCSR.

In this paper, we propose a new approach, namely overlapped di-tone GMM, for tone recognition in Cantonese continuous speech. The ODGMM is designed with special consideration on the distinct characteristics of the Cantonese tones. In addition, an effective method of F0 normalization has been used to alleviate undesirable F0 fluctuations. Viterbi based search algorithm with phonology constraints is employed to search for the best path of tone sequence given a sequence of di-tone feature vectors. Experiment results show that the proposed method significantly outperforms the conventional approaches of embedded and explicit tone recognition.

7. Acknowledgements

This research is partially supported by a Research Grant from the Hong Kong Research Grant Council (Ref: CUHK4206/01E). The authors are especially grateful to Dr. Frank K. Soong for his invaluable advice.

8. References

[1] C.-J. Chen *et al*, "New methods in continuous Mandarin speech recognition", in *Proceedings of 1997 European Conference on Speech Communication and Technology*, pp.1543-1546.

[2] H. Huang and F. Seide, "Pitch tracking and tone features for Mandarin speech recognition", in *Proceedings of 2000 International Conference on Acoustic, Speech and Signal Processing*, pp.1523-1526.

[3] Eric Chang et al, "Large vocabulary Mandarin speech recognition with different Approaches in modeling tones," in *Proceedings of 2000 International Conference on Spoken Language Processing*, pp.983-986.

[4] Y.W. Wong and Eric Chang, "The effect of pitch and tone on different Mandarin speech recognition tasks", in *Proceedings of 2001 European Conference on Speech Communication and Technology*, pp.1517-1521.

[5] Tan Lee *et al*, "Using tone information in Cantonese continuous speech recognition", in *ACM Transactions on Asian Language Information Processing*, Vol.1, No.1, pp.83 - 102, March 2002.

[6] Wai Lau, *Attributes and Extraction of Tone Information for Continuous Cantonese Speech Recognition*. MPhil Thesis, The Chinese University of Hong Kong, 2000.

[7] *Hong Kong Jyut Ping Characters Table* (粵語拼音字表), Linguistic Society of Hong Kong Press (香港語言學會出版), 1997.

[8] Clark, John and Colin Yallop. *An introduction to Phonetic and phonology. Cambridge*, MA: Basil Blackwell, Inc, 1990.

[9] Yujia Li, Tan Lee and Yao Qian, "Acoustical F0 analysis of continuous Cantonese speech", in *Proceedings of the 2002 International Symposium on Chinese Spoken Language Processing*, pp.127 - 130.