Use of Tone Information in Continuous Cantonese Speech Recognition

Yao Qian[1], Tan Lee[1] and Frank K. Soong[2]

[1]Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong
[2]Spoken Language Translation Labs, ATR, Kyoto, Japan
{yqian,tanlee}@ee.cuhk.edu.hk, frank.soong@atr.co.jp

Abstract

Cantonese, a syllabically paced, southern Chinese dialect, is also a tonal language where tones carry important lexical information. It is rich in tonal variations and each syllable can have up to 9 different tone patterns. In this paper we investigate how to incorporate the tone information into a large vocabulary continuous speech recognition system. A two-pass, post-processing scheme is proposed to utilize the recognized tones in rescoring the recognized N-best strings. Utterance level confidence measures of the N-best hypotheses are used in the rescoring process. It has been found from our experiments that weighted tone information can yield 8% relative improvement of the Chinese character error rate.

1. Introduction

Cantonese, a syllabically paced Chinese dialect, is the mother tongue of tens of millions of people living in Southern China, Hong Kong and overseas. Like Mandarin (Putonghua), the official standard of spoken Chinese, Cantonese is a tonal language. The basic written unit of Cantonese is the Chinese character and almost every Chinese character is pronounced as a tonalized monosyllable. Compared with English, Cantonese has a simpler and more restricted syllabic structure. Each syllable has a typical form of (consonant)-vowel-(consonant), where only the vowel nucleus is an obligatory element. Each Cantonese syllable can be divided into syllable initial and final which we shall refer to as Initial and Final from now on. The Initial includes what precedes the vowel while the Final includes the vowel and what follows it.

Cantonese is known of being very rich in tones. It is said to have nine citation tones that are characterized by the stylized pitch patterns as illustrated in Figure 1. The first six tones are carried by syllables that either have no ending consonants or end with /m/, /n/, /ng/. Syllables carrying the so-called entering tones must end with unreleased stop consonants /p/, /t/, /k/. Each of the entering tones is often considered as the shortened counterparts of a particular non-entering tone, in terms of the relative time duration. Therefore in many Cantonese transcription systems, e.g. the LSHK system [1], only six distinctive tone categories, labeled as T1 to T6 in Figure 1, are defined.

It is seen from Figure 1 that the tone system of Cantonese is close to a REGISTER (or LEVEL-PITCH) system [2,3]. Four of the six tones, i.e. T1, T3, T4 and T6, have either flat or slightly falling pitch patterns, which can be viewed as level-pitch tones at different levels. The six Cantonese tones can be divided into two levels or registers of HIGH and LOW, according to their pitch ranges. The HIGH group includes T1, T2 and T3, while the LOW group includes T4, T5 and T6.

There have been many studies on how to incorporate the tone information into continuous Chinese speech recognition. Approaches can be roughly divided into two major categories: embedded tone modeling and explicit tone recognition [4]. In the embedded tone modeling pitch-related features such as F0 are included as extra components in the short-time acoustic feature vector and tone explicit or dependent acoustic models are trained accordingly [5-7]. On the other hand, in explicit tone recognition tone recognition is handled in a process parallel to the process of phonetic recognition. The result of tone recognition is then incorporated with the phonetic recognition results in a post-processing stage [4,8] or integrated back into a global search process with the phonetic information [9,10].

Since pitch is a supra-segmental feature which can span over multiple voiced segments, a window wider than the normal time window used for extracting the short-time spectral features should be used for tonal feature extraction and recognition. In the approach of embedded tone modeling, F0 is part of the feature vector and the analysis window for tone extraction is exactly the same as the time window used for other short-time acoustic features. The constraints of short-time feature vector based HMM are also imposed onto tone modeling. For example, the 1st-order Markov property of state transitions, the explicit state-dependent output probability of the observations, and the independent assumption of the neighboring observations [11]. Additionally, it is convenient but apparently wrong to assume that the suprasegmental F0 features and the segmental spectral features are in synchrony with each other. In the approach of explicit tone recognition, the tone model is built separately and independently. It can have a tailor-made design that explicitly takes the supra-segmental characteristics of tones into account. For examples, long-term dynamics can be modeled in the time-frequency domain and the transition trajectories can be captured across segmental units. However, how to integrate the tone information into the search process is a puzzling problem, especially for large-vocabulary continuous speech recognition (LVCSR) tasks. Post-processing of the output from the phonetic recognition appears to be an easy way from implementation point of view.

In this paper, we first investigate potential impact of tone information on the improvement of speech recognition performance. Based on the performance analysis of our previously proposed tone recognition method [3], we propose a post-processing technique for utilizing the tone information in an N-best hypothesis re-scoring/re-ordering process. In the rescoring process an utterance-level confidence measure is adopted.

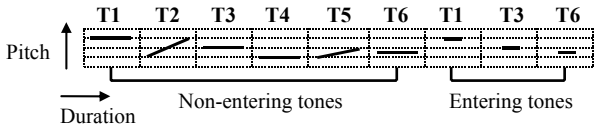

Figure 1: *Tones in Cantonese (schematic description).*

2. Tone Information for Cantonese Speech Recognition

The output of a Cantonese LVCSR system is in the form of a sequence of Chinese characters. The generation of such a sequence inevitably involves a process of conversion from syllable-level units to characters since each character is pronounced as a syllable. An Initial-Final combination is usually called a base syllable (BS) while a tonal syllable refers to a base syllable carrying a specific tone. Cantonese has about 600 legitimate base syllables and 1,700 tonal syllables. Given that the total number of Chinese characters is over 10,000, a syllable typically corresponds to a number of different characters (homophones). In other words, the average number of homophonic characters per BS and per TS is 17 and 6 respectively. It should be obvious that correct tone information can reduce the number of confusing multiple mappings between a syllable and its corresponding homophonic Chinese characters by almost three times on the average. This reduction is purely from a BS unit to a TS unit without any extra help from contexts or language models.

Let us consider the following mathematical formulation of the speech recognition process,

$$W^* = \arg\max_w p(w \mid o)$$
$$= \arg\max_w p(w)\, p(tr \mid w)\, p(o \mid tr)$$

where o is the acoustic observation, tr is the syllabic transcription of a hypothesized word w. $P(o \mid tr)$, $P(tr \mid w)$ and $P(w)$ are the acoustic model, the pronunciation dictionary and the language model respectively.

Conceptually, the recognition process can also be seen to involve two passes of conversion,

$$W^* = \arg\max_w p(w \mid o)$$
$$= \arg\max_w p(w \mid tr)\, p(tr \mid o)$$

in which $P(tr \mid o)$ governs the conversion from acoustic observations to phonetic (syllable) symbols and $P(w \mid tr)$ governs the conversion from phonetic representation to text (Chinese characters). From this perspective, we have the following thoughts on the possibility of making tone information contributive to Cantonese speech recognition.

Acoustic features to phonetic

Typically, the acoustic models used for Cantonese speech recognition are context-dependent Initial/Final models. But in small-vocabulary task, whole-syllable models may be used. For example, the recognition of digit strings can be done with 10 digit models: 0(ling4), 1(jat1), 2(ji6), 3(saam1), 4(sei3), 5(ng5), 6(luk6) ,7(cat1), 8(baat3) and 9(gau2). In this case, every word (digit) is a unique BS or TS and there is no homophonic confusion among different words. In the search process of digit recognition, only legitimate combinations of a BS and its corresponding tone are allowed to be hypothesized and conceptually we believe that this should give a better recognition results for its reduced search space.

However, for a large-vocabulary task, a syllable, depending upon its lexical context, can have more than one unique tone and confusions may result. For example, the syllable /si/ can have up to 6 different tones. Accordingly, the effect of constraining search space by using the F0 feature is no longer that obvious as in digit recognition. Nevertheless, as shown in Table 1, nearly 30% of the Cantonese base syllables are allowed to carry only one specific tone. More than 50% of the syllables carry at most two different tones. This indicates that tone information is still useful in large-vocabulary tasks.

Table 1: *Statistical breakdown of syllable distributions according to the number of tones per syllable.*

Number of tones / syllable	Percentage distribution of syllables
6	2.67%
5	5.65%
4	18.05%
3	19.31%
2	24.49%
1	29.83%

Phonetic-to-text conversion

The statistics of BS and TS transcriptions in the 6,400-word pronunciation dictionary, in which most words are either di-syllabic (54.61%) or monosyllabic (41.94%), are shown as in Table 2. The total number of BS entries and TS entries are both larger than the number of words due to polyphone transcriptions, or multiple pronunciations of a character. Table 2 also indicates that tone information can largely increase the number of one-to-one mapping entries (which means that BS or TS entry corresponds to only one word) from 3,747 to 4,568 and decrease the average number of homophonic words from 5.4 to 3.5. This conjecture, that the tone information is useful for identifying a word more distinctively, has been confirmed in our experiment on a large-vocabulary speech corpus, CUSENT [12]: using a backward Viterbi beam search with 6,400-word bi-gram language model, incorporating perfect tone information improves the accuracy of phonetic-to-text conversion rate from 90.70% to 95.81%.

Table 2: *The statistics of BS and TS transcriptions in a 6400- word pronunciation dictionary.*

	Based on BS transcriptions	Based on TS transcriptions
Total no. of entries	7168	7456
No. of one-to-one mapping entries	3747	4568
No. of transcriptions that correspond to more than one word	636	837
Average no. of homophonic words	5.4	3.5

3. Analysis of the Results of Automatic Tone Recognition

We have proposed a novel approach to tone recognition in continuous Cantonese speech with overlapped di-tone Gaussian mixture models (ODGMM) [3]. This method was designed to take into account the fact that Cantonese tone identification relies more on the relative pitch level than on the pitch contour shape. Experimental results show the ODGMM approach significantly outperforms other methods of tone recognition in continuous Cantonese speech. However, compared with the accuracy of syllable recognition, the accuracy of tone recognition is still low. As discussed in Section 2, tone information with phonological constraints on the combination of syllable and tone can improve the accuracy of syllable recognition. On the contrary, syllable information with the same constraints also can contribute to the accuracy

of tone recognition. As tone recognition relies on syllable initial and final segmentation, using syllable information for tone recognition is a more straightforward way.

In order to investigate the real confusion cases for tone recognition, we try first to isolate the problem by eliminating mis-aligned syllable boundaries and using correctly aligned data. Confusion matrix of tone recognition is shown as in Table 3. The overall accuracy is 79.95%. The best accuracy, 96.40%, 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 58.94%. The most confused tone pairs are T3-T6 and T4-T6. We think the reason causes confusion between T3 and T6 results from that the second entering tone which is regarded as abbreviated counterparts of T3 is a middle register tone. The height of it is also close to the low register tone T6.

Table 3: *Confusion matrix of tone recognition.*

	T1	T2	T3	T4	T5	T6
T1	2061	49	96	14	7	54
T2	6	1008	64	43	117	69
T3	35	66	1192	57	21	232
T4	4	33	17	1399	44	101
T5	3	47	56	51	399	79
T6	29	76	215	227	89	1552
Accuracy (%)	96.40	78.81	72.68	78.11	58.94	74.37
	79.95					

Our method of tone recognition employs Viterbi based search algorithm with phonological constraints to search for the best path of tone sequence given a sequence of di-tone feature vectors. We output the results of di-tone recognition, which skips the step of Viterbi search, to study the tone confusion. The overall accuracy of recognizing di-tone units is 68.26%. The most confused di-tone pairs are X3-X6, 3X-6X, X4-X6 and 4X-6X (X indicates the same tone in an di-tone pair). This result is consistent with the result shown in Table 3.

At this moment we still can't explain why T5 has much lower accuracy than the others and what might have caused the confusions between T4 and T6. However, the above analysis suggests that different level of reliability in tone recognition, which varies from one tone to the other, might be exploited to further enhance the performance of our Cantonese LVCSR.

4. Integrating Tone Information into Cantonese LVCSR

4.1. The Baseline Cantonese LVCSR System

The baseline Cantonese LVCSR system (CURec) was developed at the Digital Signal Processing Laboratory, the Chinese University of Hong Kong [13,14]. Details are given below.

The acoustic models
The acoustic models are the context-dependent Initial/Final models. The acoustic feature vector is composed of 12 Mel-Frequency Cepstral Coefficients (MFCC), Energy, and their first- and second-order time derivatives. Each Initial model is an HMM with three emitting states, while a Final model can be with either three or five emitting states, depending on its phonetic composition. Each emitting state is described by 16 Gaussian mixture components. Decision-tree based state clustering technique is used to facilitate the parameter sharing among models.

The language model
The language model is a word tri-gram for a 6400-word lexicon. It was trained on a text corpus of 98 million characters, compiled from five Hong Kong newspapers.

The decoder
The decoder uses the tree-trellis algorithm to generate N-best results. The forward search is based on a time-synchronous Viterbi based trellis search and the search space is arranged in a tree-structured lexicon. The backward search is A* stack-decoder based tree search.

The speech 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. This corpus is used for acoustic models training in baseline LVCSR. The baseline results of Cantonese LVCSR are given in Table 4.

Table 4: *The baseline results of Cantonese LVCSR.*

	BS	Character	
		Top1	Top 10
Accuracy(%)	82.92	80.46	87.23

4.2. Tone Information for Cantonese LVCSR Post-processing

The CUSENT corpus was used also for training and test of the tone models. The accuracy of tone recognition is 74.68%. If we combined the recognition results of tone and BS directly, the TS accuracy was lower than each individual performance since unreliable TS affected PTT conversion negatively. Alternatively, the recognized tone can be used to re-score the N-best word recognition hypotheses. Using the correct tone information in an oracle experiment to re-score the N-best hypotheses, an increase of 4.1% was obtained in terms of the accuracy of Chinese character conversion. While using the recognized tones with 74.68% accuracy did not improve the performance. On the contrary, a decrease of 0.5% on character conversion accuracy was observed.

4.2.1. Confidence Measures for Speech Recognition

In order to solve the above problem, we propose a post-processing scheme to apply weighted tone information to re-scoring the N-best list. It uses confidence measures to estimate the reliability of a hypothesized output. Confidence scores can be defined at various levels: frame, word and utterance. Examples of scores used in conventional decoding are frame acoustic likelihood, phone and word duration penalties, word language probabilities and prosodic confidence score [15]. In this study, we employ the utterance level score to re-order the N-best utterance hypotheses. The utterance confidence score $c(u)$ is computed by

$$c(u) = \frac{1}{h}\sum_{j=1}^{h} c(c_j)$$

$$c(c_j) = \begin{cases} w(t_n) & \text{recognized tone} \approx \text{tone of character} \\ 0 & \text{otherwise} \end{cases}$$

where $c(c_j)$ is character confidence score and h is the number of character. t_n is the tone identity. The symbol "$\approx$"

means that the most confused tone pairs such as T3 and T6 are counted to be equal except for the real equal pairs. We assign different value of $w(t_n)$ according to the following aspects:

- The accuracy of tone recognition shown in Table 3
- The degree of confusion with other tones
- The frequency of occurrences in the N-best hypotheses
- The number of N-best hypotheses

$w(t_n)$ can be seen as the weighted tone information. Its value is generated on-the-fly according to the rules embedded into a post-processor of the system.

4.2.2. Experimental Results

With the weighted tone information, the accuracy of the LVCSR system was improved from 80.46% to 82.02%, i.e. a 8% relative reduction of character error rate. Table 5 gives detailed recognition performance for each test speaker in CUSENT. Integrating weighted tone information into LVCSR improves the performance for all but one of the speakers.

For better understanding of the usefulness of tone information, we transcribed the recognized characters back to syllables and found that syllable accuracy is slightly higher than the one without tone information, from 82.92% to 83.31%. That is, the weighted tone information gives less improvement of acoustic-to-phonetic than phonetic-to-text conversions.

Table 5: *the LVCSR post processor performance for individual speakers.*

Speaker	Character Accuracy (%)	Tone Accuracy (%)	Character Accuracy Improvement(%)
F4F	78.76	73.47	0
F5F	80.14	73.25	1.87
F6F	74.66	69.93	2.04
F7F	82.54	70.39	1.21
F8F	84.27	75.38	2.86
F9F	83.22	79.09	0.44
FAM	81.63	76.91	2.76
FBM	84.9	78.67	1.8
FCM	76.89	71.61	1.58
FDM	81.1	74.94	0.82
FEM	78.65	76.73	1.04
FFM	79.19	74.65	2.23
Average	80.46	74.68	1.56

5. Discussion and Conclusions

Tone information is critical to human perception of natural speech. How to use it as an additional knowledge source for automatic speech recognition of tonal language is a challenging problem. Under the current framework of LVCSR, we study possible ways to incorporate tone information into speech recognition. In our previous work, we have reported the tone recognition improvement of a new di-tone based statistical model. In this paper, we analyze the results of tone recognition and integrate the tone information into our Cantonese LVCSR system as a confidence measure based N-best rescoring process. The experimental results show that the weighted tone information used as the confidence measure indeed helps to improve the overall performance of LVCSR.

6. Acknowledgements

This research is partially supported by a Research Grant from the Hong Kong Research Grant Council (Ref: CUHK4206/01E).

7. References

[1] Linguistic Society of Hong Kong (LSHK),1997. *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).

[2] Clark, J.; Yallop, C., 1990. *An Introduction to Phonetic and Phonology.* Cambridge, MA: Blackwell.

[3] Qian, Y.; Lee, T.; Li, Y.J., 2003. Overlapped di-tone modeling for tone recognition in continuous Cantonese speech. In *Proceedings of the 8th Eurospeech,*1845-1848.

[4] Lee, T.; Lau, W.; Wong, Y.W.; Ching, P.C., 2002. Using tone information in Cantonese continuous speech recognition. *ACM Transactions on Asian Language Information Processing,* Vol.1, No.1, 83-102.

[5] Chen, C.J.; Gopinath, R.A.; Monkowski, M.D.; Picheny, M.A.; Shen, K., 1997. New methods in continuous Mandarin speech recognition. In *Proceedings of the 5th Eurospeech,* 1543-1546.

[6] Huang, H.; Seide, F., 2000. Pitch tracking and tone features for Mandarin speech recognition. In *Proceedings of ICASSP,* 1523-1526.

[7] Wong, Y.W.; Chang, E., 2001. The effect of pitch and tone on different Mandarin speech recognition tasks. In *Proceedings of the 7th Eurospeech*, 1517-1521.

[8] Lin, C.H.; Wu, C.H.; Ting, P.Y.; Wang, H.M., 1996. Frameworks for recognition of Mandarin syllables with tone using sub-syllabic units. *Speech Communication,* Vol.18, No.2 , 175-190.

[9] Seide, F.; Wang, N.J.C., 2000. Two-stream modeling of Mandarin tones. In *Proceedings of the 6th ICSLP,* 495-518.

[10] Cao, Y.; Deng, Y.; Zhang, H.; Huang, T.; Xu, B., 2000. Decision-tree based Mandarin tone model and its application to speech recognition. In *Proceedings of ICASSP,* 1759-1762.

[11] Huang, X.; Acero, A.; Hon, H., 2001. *Spoken Language Processing: A Guide to Theory, Algorithms and System Development*. Prentice Hall PTR.

[12] Lee, T.; Lo, W.K.; Ching, P.C.; Meng, H., 2002. Spoken language resources for Cantonese speech processing. *Speech Communication,* Vol.36, No.3-4, 327-342.

[13] Wong, Y.W.; Chow, K.F.; Lau, W.; Lo, W.K.; Lee, T.; Ching, P.C., 1999. Acoustic modeling and language modeling for Cantonese LVCSR. In *Proceedings of the 6th Eurospeech,* 1091-1094.

[14] Choi, W.N.; Wong, Y.W.; Lee, T; Ching, P.C., 2000. Lexical tree decoding with a class-based language model for Chinese speech recognition. In *Proceedings of the 6th ICSLP,* 174-177.

[15] Koo, M.W.; Lee C.H.; Juang, B.H., 2001. Speech recognition and utterance verification based on a generalized confidence score. *IEEE Transaction on Speech and Audio Processing*, Vol. 9, No.8. 288-298.