A memory-based approach to Cantonese tone recognition

Michael Emonts

Spoken Language Processing
Sony Electronics
mike@slp.sel.sony.com

Deryle Lonsdale

Department of Linguistics
Brigham Young University
lonz@byu.edu

Abstract

This paper introduces memory-based learning as a viable approach for Cantonese tone recognition. The memory-based learning algorithm employed here outperforms other documented current approaches for this problem, which is based on neural networks. Various numbers of tones and features are modeled to find the best method for feature selection and extraction. To further optimize this approach, experiments are performed to isolate the best feature weighting method, the best class voting weights method, and the best number of k-values to implement. Results and possible future work are discussed.

1. Introduction

This paper introduces memory-based learning as a viable approach for Cantonese tone recognition. Speech recognition has only recently been applied to Cantonese[1], though considerable effort has been spent in recognizing Mandarin[2].

Various numbers of tones and features are modeled to find the best method for feature selection and extraction. To further optimize this approach, experiments isolated the best feature weighting method, the best class voting weight method, and the best number of k-values to implement. The recognition accuracy achieved was 90.9% in a speaker-dependent system and 81.8% in a speaker-independent system.

Cantonese syllable structure is simple: (C)V(C), with optional elements parenthesized. The optional onset may involve any valid consonant, the nucleus may involve a diphthong, and the optional coda may only be a nasal consonant or a voiceless unaspirated stop (which is unreleased). Cantonese is a highly tonal language, and tone is an intonational property that extends over the entire syllable. A tone reflects the movement of pitch (or fundamental frequency) over time.

Cantonese has a complicated tone system with tone-based minimal pairs. The number of Cantonese tones has been hotly debated: linguists have argued for 6, 7, 9, 10, or even 12 tones. It is well documented, however, that there are 9 citation tones in Cantonese, observable when a native speaker of Cantonese is prompted with a Chinese character in isolation. We assume a basic 9-tone scenario and also experiment with a 6-tone one; tonal notation follows the LSHK system [1].

Though Chinese dialects are tonal in nature, many current Chinese recognizers do not use tone information. Still, the inclusion of tone information has been shown to greatly enhance recognition accuracy. Even errorful tone recognition has been shown to improve recognition rates for Mandarin and Cantonese [2].

Very little tone recognition research has been performed to date for Cantonese. Monosyllabic tone recognition would seem to be an appropriate target since 23% of words in Chinese are monosyllables (62% are disyllables, 6% are trisyllables, and 9% are more than three syllables) [3]. This paper reports on isolated monosyllabic tone recognition excluding segmental (i.e. phonemic) identification; other work on disyllabic recognition [4], polysyllabic recognition [5] and continuous recognition of Mandarin [6, 7] and Cantonese [8, 9, 2] or both [10] has been described elsewhere.

Prior to the present work, the only published work on monosyllabic Cantonese tone recognition used neural networks [11]. Recognition accuracy of 89.0% and 87.6% was achieved for single-speaker and multi-speaker recognition respectively. Neural networks have also been applied to Mandarin tone recognition.

Hidden Markov models (HMMs) are currently the best-performing speech recognition approach, both for word spotting and for continuous speech recognition. Several researchers have applied hidden Markov models to tone recognition of syllables. Other classification methods successfully used in tone recognition include fuzzy logic, decision trees, and vector quantization. Hybrid approaches are also increasingly popular.

Since experiments differ greatly in complexity, scope, corpora used, subjects, and even language, comparison between approaches on basis of net accuracy alone is not without difficulties. Given this caveat, a summary of relevant research is presented in Figure 1. Note that most systems have used neural networks, HMMs, vector quantization, or a hybrid of these three methods.

[1] a Chinese dialect with over 40 million speakers worldwide
[2] the standard Chinese dialect with over 700 million speakers

Method	Ref.	Spk-Dep	Spk-Ind
Neural Network	[12]	97.5%	92.5%
Neural Network (w/ backprop)	[13]	—	93.8%
HMM + differential coding	[14]	96.0%	—
HMM (w/o pitch features)	[15]	94.9%	—
HMM (60% lower complexity)	[16]	89.7%	87.9%
Vector Quantization (VQ)	[17]	—	98.8%
Fuzzy Sets	[18]	99.5%	—
Decision Tree	[19]	—	90.2%
VQ/HMM	[20]	98.3%	96.5%
VQ/HMM	[3]	—	97.9%
VQ/Multilayer perceptron	[21]	99.8%	—
HMM (Thai)	[22]	90.0%	—
Neural Network (Cantonese)	[11]	89.0%	—

Figure 1: Summary of Monosyllabic Tone Recognition Research (all for Mandarin unless noted)

2. Methodology

Our tone recognition process includes three general steps: (i) extracting speech sample features from the Chinese University Syllable Corpus[3] (CUSYL, version 1.0) [23]; (ii) building input files for the classification engine; and (iii) using the memory-based classifier for identifying tones.

2.1. Feature extraction and vectorization

Fundamental frequency (F_0) values were extracted via autocorrelation and two slight variations of the cepstral algorithm (cepA and cepB) [24] using the Speech Filing System[4]. Since only voiced speech contains tone information, zero values (pre-, post-, and even mid-syllable ones) from the sample were all removed.

Our work assumed equal-length input feature vectors. Others have tried vector lengths such as 10, 16, or even 80 [14, 11, 4]; we chose lengths of 16 and 8. When necessary, vectors were left-padded with values of zero ("zero-addition") or 100 ("100-addition"). Overly long vectors were reduced by selecting the first and last features, as well as maximally equidistant remaining ones; this helped preserve tonal contours.

For speaker-independent experiments, we normalized feature vectors using standard methods in order to directly compare data from different speakers.

Cantonese syllable onsets are unvoiced and hence carry no tone information, so tone recognition for Cantonese is limited to rime data: the nucleus (vowel) and the optional nasal coda. Data from the peripheries, especially from nasal codas, tend to be less predictable. Various techniques have been tried to prune potentially harmful features from the periphery of syllable data [7, 25] with no emerging concensus. We will show that featyre trimming in vertain situations can be indeed helpful.

[3]The data includes speech from 2 male and 2 female speakers, each speaking the entire set of about 1,800 Cantonese syllables.

[4]Freely available at ftp://ftp.phon.ucl.ac.uk/pub/sfs.

2.2. Tone classification

Tone classification was carried out using memory-based learning (MBL), an exemplar-based machine-learning approach which employs various similarity metrics. The basic assumption underlying memory-based learning is that the use of stored experience outperforms the application of knowledge (such as rules or decision trees) abstracted from experience. MBL has attained adequate to excellent generalization accuracies on a wide variety of complex NLP tasks.

The memory-based learning techniques presented in this paper were implemented using the widely-employed TiMBL system [26]. TiMBL's learning component loads labeled feature vectors into memory without using any abstraction, structuring principles, or explicit rules. Its performance component then calculates the similarity between test feature vectors and each exemplar stored in memory, in order to ascertain the test vector's best output classification. Five basic parameters were crucial in obtaining optimal results, all of which are fundamental to TiMBL's operation; thousands of parameter permutations were attempted to find the optimal combinations.

TiMBL provides three *similarity metrics* for computing the likeness between two features: Overlap, Modified Value Difference (MVDM), and Numeric. As our features involved numerical F_0 values, the Numeric metric consistently outperformed the other two.

Five available *feature weighting* methods were tried with variable results: none, Information Gain, Gain Ratio, Chi-squared, and Shared Variance.

The number of relevant nearest neighbors (i.e. k-values) affects the classification outcome. Increasing the k-value increases generalization of the system, but decreases efficiency. We used odd-numbered k-values from 1 to 15; even k-values were avoided to limit the number of ties.

Class voting weights allow for differential weighting of the k-nearest neighbors. In this work four functions were tried: Normal, Inverse Linear, Inverse Distance, and Exponential Decay.

Four *classification algorithms* specify an overall approach for classifying each given instance: IB1, IGTREE, TRIBL, and IB2. The most accurate (and hence default) algorithm, IB1, was chosen for this work although its use entails increased computational cost. Test cases confirmed its outperformance of the other algorithms.

Training data for speaker-dependent experiments consisted of 1600 CUSYL files, with 200 being held back for test data. For speaker-independent experiments, all data files from three speakers were used for training with the remaining speaker (all 1800 files) being used to test. In this case, all possible permutations were performed and final results were averaged across each of the four speakers.

3. Results

90.9% accuracy was attained for the speaker-dependent 6-tone system, and 87.0% for the 9-tone system. For speaker-independent systems, accuracies of 81.8% and 77.5% were attained for the 6-tone and 9-tone systems respectively. The 90.9% accuracy in speaker-dependent experiments compares favorably with previous work using other techniques.

Choice of feature extraction method has a considerable effect on recognition accuracy. Autocorrelation was found to be the best method when using only 8 features. However, the cepstrum methods (both cepA and cepB) work best for 16-feature settings, particularly when using the zero-addition or 100-addition methods.

Normalizing fundamental frequencies was found to be essential for speaker-independent experiments. Accuracy due to normalization increased from 51.4% to 81.0% (6-tone experiments) and 46.8% to 72.5% (9-tone experiments).

8-feature systems worked best without feature trimming (contra popular practice for Cantonese and Mandarin tone recognition). Indeed, a direct correlation was observed: the more features were dropped, the lower the accuracy. In the speaker-dependent 16-feature system, it was found that ignoring the first two features attained best results.

Choice of a particular feature weighting method has almost no effect on accuracy. It was clear, however, that in 9-tone/8-feature systems, disabling feature weighting was detrimental. Generally, using the Gain Ratio method achieved highest recognition accuracy albeit by very small margins.

In calculating feature distances, the number of nearest neighbors (k-values) plays an important role. This work confirms the growing concensus that accuracy increases as the value of k does. This was especially true in speaker-independent systems. It is possible that using values greater than our maximum (15) could result in even higher accuracy.

Intertwined with k-values is the class voting weights method. Each method implemented performed similarly, with best results from Inverse Linear (for 6-tone systems) and Inverse Distance (for 9-tone systems). Speaker-independent systems tended to benefit most from the former.

The optimal number of F_0 features remains an open question. This work used commonly-adopted 16-feature vectors, and then compared them with smaller, computationally friendlier 8-feature vectors. The average recognition accuracy was 84.3% for 16-feature systems, and 81.4% for 8-feature systems. The difference was much greater for 9-tone systems than for 6-tone systems. Using 8 features instead of 16 resulted in a loss of only 0.4% in the speaker-dependent 6-tone system and 0.9% in the speaker-independent 6-tone system. In 9-tone systems, however, losses of 6.1% and 4.1% (respectively) are observed. If computational restrictions do not exist, then clearly using 16 features is better. Using 8 features is a viable option to lessen computational burden with only a slight loss of accuracy, especially in 6-tone recognition systems.

In addressing the issue of how many tones to model, we compared 6-tone systems and 9-tone systems; the former achieved 86.0% on average, whereas the latter only achieved 79.7%. This is due to the fact that the additional three tones are extremely similar to existing tones in the 6-tone system, and the extra tones introduce more opportunities for misclassification. Thus, unless the specific need exists for a 9-tone system, the use of a 6-tone system is strongly recommended.

Figure 2 summarizes which parameters attained best results for each of the 8 systems evaluated.

Analysis of confusion matrices helped isolate which tones were particularly easy or difficult to recognize. In all experiments, the most common classification error was due to the recognizer's difficulty in distinguishing among level tones. The addition of three more level tones in 9-tone systems added to the confusion, especially between tones that differed primarily by duration rather than contour or pitch.

Presumably our approach would achieve even higher results when applied to the Mandarin tone recognition task. Not only are there only four tones in Mandarin, but also the contours of each tone are more distinct from each other. Limiting the tone recognition task to the four Cantonese tones most similar to the four Mandarin tones produced hypothetical results for a Mandarin tone recognizer. Hypothetical results predict 98.1% (speaker-dependent) and 92.7% (speaker-independent).

4. Future work

The system presented here can be improved upon in further computationally-based investigations, for example: (i) finding and using an optimal number of features in each feature vector; (ii) increasing the amount of training data; (iii) using higher k-values; and (iv) using a more accurate feature extraction method.

This paper has only addressed Cantonese tone recognition (not addressing segmental or phonemic recognition), and shows that our tone recognition approach is viable. Since, as noted above, Cantonese recognizers typically do not use tonal information, it would be interesting to integrate our approach with a strictly segmental recognizer to study the expected increase in Cantonese speech recognition accuracy.

Work in the future might also apply this approach to more complicated tasks such as continuous or large-vocabulary speech and to other tonal languages such as Mandarin, Thai, Vietnamese, or other dialects of Chinese.

System	FeatExtrMeth	FeatsTrim	FeatWeighting	k	ClassVoteWt	%
Spkr-Dep 6-Tone/8-Feat	autocorr	—	gain ratio	13	InvDist	90.5
Spkr-Dep 9-Tone/8-Feat	cepB	—	gain ratio	11/13	InvLinear	80.9
Spkr-Dep 6-Tone/16-Feat	cepB(0)	first two	shared-var/χ^2	5	InvDist	90.9
Spkr-Dep 9-Tone/16-Feat	cepA(100)	first two	gain ratio	11	InvLinear	87.0
Spkr-Indep 6-Tone/8-Feat	cepA	—	shared-var	15	InvDist	80.9
Spkr-Indep 9-Tone/8-Feat	autocorr	—	gain ratio	15	InvLinear	73.4
Spkr-Indep 6-Tone/16-Feat	cepA(spread)	—	shared-var/χ^2	13	InvDist	81.8
Spkr-Indep 9-Tone/16-Feat	cepA(0)	first two	gain ratio	15	InvDist	77.5

Figure 2: Parameter settings producing best results for all systems

5. References

[1] W.K. Lo, Tan Lee, and P.C. Ching. Development of Cantonese spoken language corpora for speech applications. In *Proceedings of 1998 International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 102–107, 1998.

[2] Wai Lau, Tan Lee, Y.W. Wong, and P.C. Ching. Incorporating tone information into Cantonese large-vocabulary continuous speech recognition. In *Proceedings of the 2000 International Conference on Spoken Language Processing (ICSLP)*, volume 2, pages 883–886, 2000.

[3] W. Yang Liu, H. Wang, and Y. Chang. Tone recognition of polysyllabic words in Mandarin speech. *Computer Speech and Language*, 3:253–264, 1989.

[4] Jin-song Zhang and Keikichi Hirose. Anchoring hypothesis and its application to tone recognition of Chinese continuous speech. In *Proceedings of the 2000 International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1419–1422, Istanbul, 2000.

[5] Yadong Wu, Kazou Hemmi, and Kazuo Inoue. A tone recognition of polysyllabic Chinese words using an approximation model of four tone pitch patterns. In *Proceedings of the 1991 International Conference on Industrial Electronics, Control and Instrumentation*, pages 2115–2117, 1991.

[6] Sin-Horng Chen and Yih-Ru Wang. Tone recognition of continuous Mandarin speech based on neural networks. *IEEE Transactions on Speech and Audio Processing*, 3:146–150, 1995.

[7] Hengjie Ma. The four tones recognition of continuous Chinese speech. In *Proceedings of the 1987 International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 65–68, 1987.

[8] K.F. Chow, Tan Lee, and P.C. Ching. Sub-syllable acoustic modeling for Cantonese speech recognition. In *Proceedings of 1998 International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 75–79, 1998.

[9] Y.W. Wong, K.F. Chow, W. Lau, W.K. Lo, Tan Lee, and P.C. Ching. Acoustic modeling and language modeling for Cantonese LVCSR. In *Proceedings of the 6th European Conference on Speech Communication and Technology*, volume 3, pages 1091–1094, 1999.

[10] Sheng Gao, Tan Lee, Y.W. Wong, Bo Xu, P.C. Ching, and Taiyi Huang. Acoustic modeling for Chinese speech recognition: A comparative study of Mandarin and Cantonese. In *Proceedings of the 2000 International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1261–1264, Istanbul, Turkey, 2000.

[11] Tan Lee, P.C. Ching, L.W. Chan, and Brian Mak. An NN based tone classifier for Cantonese. In *Proceedings of 1993 International Joint Conference on Neural Networks (IJCNN)*, volume 1, pages 287–290, Nagoya, 1993.

[12] Jhing-Fa Wang, Chung-Hsien Wu, Shih-Hung Chang, and Jau-Yien Lee. A hierarchical neural network model based on a C/V segmentation algorithm for isolated Mandarin speech recognition. *IEEE Transactions on Signal Processing*, 39:2141–2146, 1991.

[13] Pao-Chung Chang, San-Wei Sun, and Sin-Horng Chen. Mandarin tone recognition by multi-layer perceptron. In *Proceedings of the 1990 International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 517–520, 1990.

[14] Xi-Xian Chen, Chang-Nian Cai, Peng Guo, and Ying Sun. A hidden Markov model applied to Chinese four-tone recognition. In *Proceedings of the 1987 International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 797–800, 1987.

[15] Chih-Heng Lin, Lin-Shan Lee, and Pei-Yih Ting. A new framework for recognition of Mandarin syllables with tones using subsyllabic units. In *Proceedings of the 1993 International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 227–229, 1993.

[16] Xia Wang and Juha Iso-Sipilä. Low complexity Mandarin speaker-independent isolated word recognition. In *Proceedings of the 2002 International Conference on Spoken Language Processing (ICSLP)*, pages 1589–1592, Taipei, 2002.

[17] Cuntai Guan and Chen Yongbin. Speaker-independent tone recognition for Chinese speech. *Acta Acustica*, 18:380–385, 1993.

[18] Shilin Xu and Samuel C. Lee. A fast real time Chinese tone recognition system using fuzzy sets. *International Journal of the Chinese and Oriental Languages Information Processing Society*, 2:1–13, 1992.

[19] Kai-Cheng Chang and C.C. Yang. A real-time pitch extraction and four-tone recognition system for Mandarin speech. *Journal of the Chinese Institute of Engineers*, pages 37–49, 1986.

[20] Wu-Ji Yang, Jyh-Chyang Lee, Yueh-Chin Chang, and Hsiao-Chuan Wang. Hidden Markov model for Mandarin lexical tone recognition. *IEEE Transactions on ASSP*, pages 988–992, 1988.

[21] Liang Zhou and Satoshi Imai. Chinese all syllable recognition using combination of multiple classifiers. In *Proceedings of the 1996 International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3494–3497, 1996.

[22] A. Tungthangthum. Tone recognition for Thai. In *Proceedings of the 1998 IEEE Asia-Pacific Conference on Circuits and System*, pages 157–160, 1998.

[23] Tan Lee, W.K. Lo, and P.C. Ching. Spoken language resources for Cantonese speech processing. *Speech Communication*, 36:327–342, 2002.

[24] Michael Emonts. Memory-based tone recognition of Cantonese syllables. Master's thesis, Brigham Young University, 2002.

[25] J. Chen, H. Li, L. Shen, and G. Fu. Recognize tone languages using pitch information on the main vowel of each syllable. In *Proceedings of the 2001 International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2001.

[26] Walter Daelemans, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch. TiMBL: Tilburg Memory Based Learner, version 4.1, reference guide. Technical Report 01-04, ILK, 2001. http://ilk.kub.nl/downloads/pub/papers/ilk0104.ps.gz.