Tone Ratios Combined with F0 Register in Cantonese as Speaker-dependent Characteristic

Yujia Li

Department of Electronic Engineering
The Chinese University of Hong Kong, Hong Kong
yjli@ee.cuhk.edu.hk

Abstract

F0 is considered to provide speaker-specific information in some extent. Based on the widely agreement that extrinsic F0 is helpful for speaker identity, this paper investigates the possibility of making use of both extrinsic and intrinsic features of Cantonese tone system as speaker-dependent characteristic. Considering the special characteristic of Cantonese tone system, relative tone ratios and F0 register are proposed to model the tone systems generated by different speakers. The investigation is carried out over both recognition and analysis. The results primarily show the potential of implementing such features on speaker characterization.

1. Introduction

As an acoustical attribute of prosody, F0 does not only deliver perceived naturalness of continuous speech; it also has great potential to provide speaker-specific information. In some research F0 has been implemented as a feature in speaker recognition to improve the system performance [1] [2]. F0 register/range is considered most as a cue to speaker identity [3] [4]. Obviously male and female speakers are distinguished mainly by different F0 ranges and F0 registers [5]. Most research on speaker F0 normalization also indicates that F0 register/range helps to reduce speaker effects.

In tonal language like Mandarin or Cantonese, there is well defined tone system. However, there also exists significant difference among different tone systems. Mandarin is more like a CONTOUR (or GLIDING-PITCH) system which uses distinctive tone shapes to contrast with each other. While Cantonese is close to a REGISTER system, which uses distinctive pitch levels to distinguish tones [6]. In continuous speech, these tones shift from time to time in acoustical realization, and are recognized mainly by contrasting with the neighboring tones, in a relative sense, referred to the speaker's F0 range [4].

As F0 range/register is widely considered as speaker-dependent characteristic, further, based on the results of perceptual test, Moore discussed the possibility that intrinsic F0 (tone realization) may also enable listeners to establish a representation of speaker's identity [4]. However, how to describe such "intrinsic F0" in Cantonese? Is it possible that how to realize the relativity between different tones is a meaningful kind of intrinsic F0 for speaker identity? In previous studies [7-9], to facilitate the acoustical analysis of Cantonese tone contours, a F0 normalization method is proposed to alleviate F0 variations resulted from different speakers, utterance to utterance and intra-utterance intonation at once. In which relative tone ratios are proposed to build up the height relationship between different tones, so as to capture the stable characteristic of tone realization. Even though the analysis is based on a single speaker's data, the relative tone ratios combined with a scaling factor reflecting speaker's F0 register are considered to contribute most for removing speaker's variation. It is also discussed that this kind of parameters actually describe how the speakers produce their own tone system.

This paper presents a primary investigation based on the recognition and analysis experiments, by using a feature set—relative tone ratios and F0 register, over different speakers. The target is to find if such feature is speaker related. Positive result will benefit technologies such as speaker recognition, speaker-independent speech recognition or speaker-dependent text-to-speech synthesis.

2. Tones in Cantonese

2.1. Tone system

Spoken Cantonese is made up of a sequence of monosyllabic sounds. Each Chinese character is pronounced as a tonal syllable that carries a base syllable plus a specific tone. Meanwhile each base syllable can be associated with different tones to represent different characters, thus to deliver different lexical meanings. So tones actually work as a part of lexicon.

Cantonese is said to have nine citation tones that are characterized by different stylized pitch patterns as illustrated in Figure 1. The so-called "entering" tones occur exclusively with "checked" syllables, i.e. syllables ending in an occlusive coda /p/, /t/ or /k/. They are contrastively shorter in duration but coincide with a non-entering counterpart in terms of pitch contour. In many transcription schemes, only six distinctive tone categories, labeled as Tone 1 to Tone 6, are defined [10]. Among the six non-entering tones, four have flat pitch patterns, differing only by distinct pitch heights.

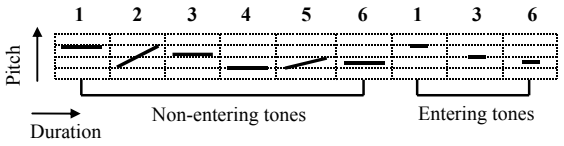

Figure 1: *Tones in Cantonese: schematic description*

Although the tone system is well defined, all the tones are defined in a relative sense by contrasting with each other. How they internally contrast is referred as intrinsic feature; while where to put the whole system is considered as extrinsic feature. Such relative sense is reflected in both tone production and tone perception.

2.2. Acoustical realization

Acoustically, tone is realized by the F0 movement across the voiced portion of a syllable. In isolation case, the acoustical realization of Cantonese tones can reflect the schematic

patterns very well in terms of surface contour and internal contrast [9]. However, because of the relative sense in definition, in continuous speech, there is much variation in acoustical realization, resulted from factors such as intra-utterance, inter-utterance and inter-speaker. Figure 2 gives an example of acoustical realization of tones in continuous speech. The utterance contains a Cantonese sentence. The curve in the upper part is the extracted F0 contour and the lower part is the concatenation of the respective schematic tone patterns. The characters and their respective tone identities are also shown in alignment with the F0 contour. Clearly, in continuous speech, tones are not realized as their canonical patterns. Even the same tone can be realized quite differently, especially in terms of tone height, like the five occurrences of Tone 3 and the three occurrences of Tone 1 (marked with circles) in the example.

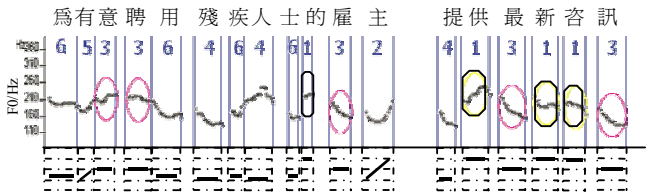

Figure 2: *F0 contour of a continuous speech utterance*

Crossing over different speakers, tones can be realized in very different ways. Figure 3 depicts the F0 contours obtained from a male speaker and two female speakers, who utter the same sentence "我從小就很愛魔術。" ("I like magic very much since I was a child.") in Cantonese. The obvious difference is observed from the three realized F0 contours. Compared with the two female speakers, the male speaker completes it within a very limited range. Between the two female speakers, the second one appears to have a relatively larger excursion of F0, which results in a wider F0 range.

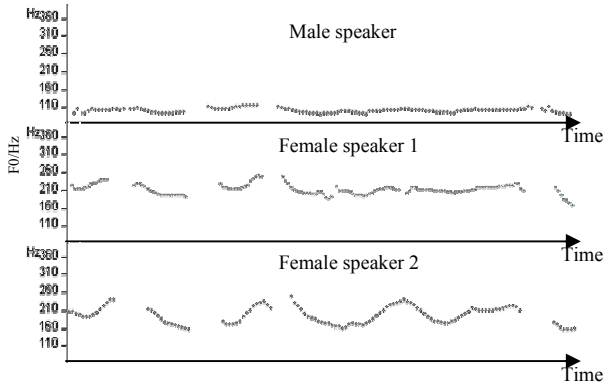

Figure 3: *F0 contours of one sentence from three speakers*

The question is can we make use of such difference to develop speaker-dependent characteristic? And how can we effectively model such difference to represent the speaker-dependent characteristic?

3. Relative Tone Ratios and F0 Register

Relative tone ratios aim to capture the internally stable feature of tone realization. The creation of such tone ratios is based on the following assumption. For the same speaker, although the absolute F0 level of a particular tone may vary greatly, its relative height with each other tone remains largely invariant. Such invariance is preserved locally, i.e. between neighboring syllables, for the requirement of communication accuracy and the continuous muscle movement of the vocal cords.

Given a pair of neighboring tones (i, j), where i and j denote the preceding and the succeeding tones respectively, the height ratio of this tone pair is computed as

$$R(i,j) = \frac{Height\ of\ Tone\ i \quad i = 1,2,...,6}{Height\ of\ Tone\ j \quad j = 1,2,...6} \qquad (1)$$

Here, the height of a tone is defined as the mean value of the respective F0 contour. Table 1 gives a six-by-six matrix of the tone ratios, computed based on 1,200 utterances of a single speaker in *CUProsody* [7], where R_{ij}, denoting the relative height ratio of Tone i over Tone j, is the average of all occurrences of $R(i, j)$.

Table 1: *Relative tone ratios derived from CUProsody*

		j					
	R_{ij}	1	2	3	4	5	6
i	1	0.97	1.39	1.28	1.60	1.39	1.35
	2	0.71	0.99	0.92	1.11	0.95	0.97
	3	**0.80**	**1.07**	**1.02**	**1.32**	**1.13**	**1.13**
	4	0.65	0.91	0.83	1.08	1.00	0.94
	5	0.71	0.99	0.93	1.16	1.02	1.01
	6	0.73	1.01	0.95	1.22	1.07	1.05

It is observed: all the diagonal elements, which are the ratios between the same tones, are around 1 with a slight deviation; $R_{ij} \approx R_{ji}^{-1}$ means the occurrence order of the tones does not affect their relative ratio of heights; $R_{ij} / R_{kj} \approx R_{ik}$ $(j = 1, 2, \cdots, 6)$ indicates internal consistence. The above observations prove that the estimation of the relative tone ratios is reliable.

Relative tone ratios reflect how the speaker internally produces his/her tone system. For a speaker, given his/her relative tone ratios and the height of a reference tone, the heights of all the other tone classes can be regained, meanwhile, the knowledge about speaker's F0 register and F0 range is also learned. According to the left-to-right control pattern of Cantonese [9], for the ratio R_{ij}, i should refer to the reference tone while j should be the transmitted tone. Being located in the middle of the tone space and with relatively small contextual variation, Tone 3 is selected as the reference tone. Thus the height of Tone 3, referred as H_3, is used to represent the general F0 register of a speaker. Ratios R_{3k} $(k = 1,2,...,6)$ are selected to describe the contrast relation among the six tones.

We use a feature set $[H_3\ R_{31}\ R_{32}\ R_{33}\ R_{34}\ R_{35}\ R_{36}]$ to model the realization of a speaker's tone system. According to the feature set, the general F0 register is decided by H_3; how to extend the range of the system is controlled by R_{31} and R_{34}. The other four tone ratios decide how to put other tones in the range. Figure 4 partially gives visual evidence how the different feature sets produce different tone systems. Retaining the surface tone contour of different tone classes in continuous speech, only changing H_3, will result in much different tone systems. Both the register and range of the whole system are changed. Now consider if we change the tone ratios and H_3

together, what we will get? Can we generate different stylized speech as from different speakers? Reversally, if the perceived speech identified from different speakers, is based on the perception of such features?

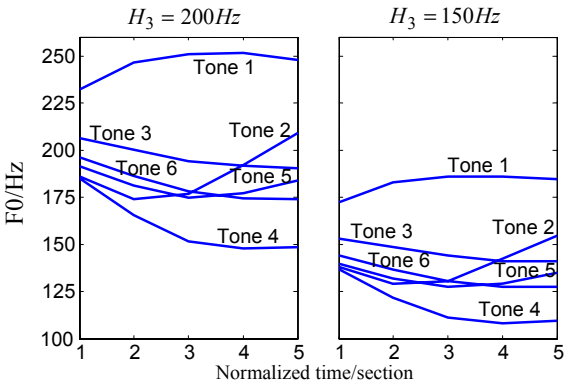

Figure 4: *Based on the retained surface tone contours, recovered tone contours by different F0 register*

4. Experiments and Discussion

Towards the target, two experiments are proposed. One is in terms of recognition; another one is based on acoustical analysis and perceptual analysis.

4.1. Corpus

The database we used in the experiments is *CUSENT*. It is a multi-speaker Cantonese reading speech corpus. Totally 68 speakers, half is male and half is female. Control material is 5100 sentences from local newspaper. They are repeated 4 times and give 21,000 utterances. Each speaker's speech data is 300 unrepeated utterances among the whole database.

The experiments focus on the tone system, so only F0 is used. It is post-processed feature out of speech data. F0 is extracted automatically by ESPS software [11]. F0 contour is obtained by HMM forced alignment.

4.2. Experiment 1: template matching

This experiment investigates the possibility of using defined feature set in speaker recognition. As we do not know exactly how effective the feature is, this single offline experiment is proposed to be the first stage.

In this experiment, each speaker's data is partition into training and testing sets. Each set builds a template for the speaker. The template matching is conducted between two template pools. For a given testing template, in training template pool the one giving the smallest distance is the matched one. If the feature is good, the distance between templates from the same speaker should be very small, meanwhile the distance between templates from different speakers should be very large. The template is the 7-dimension feature set, in which each feature is modeled by the mean of all occurrences of a feature. The feature set is post-processed by normalizing the 7 features into similar level and weighting H_3. The distance measure method is Euclidean distance as shown in the formula:

$$d(p,q) = \left[\sum_{k=1}^{7}(p_k - q_k)^2\right]^{\frac{1}{2}} \quad \begin{matrix} p,q : template \\ k : feature\ index \end{matrix} \qquad (2)$$

In order to obtain reliable results, each speaker's data, 300 utterances, is randomly partition into 6 sets. Alternate testing is conducted among these sets. Each time any 4 sets 200 utterances are used as training data and the left 1 or 2 sets are as testing data, until any combination of sets are used in both training and testing. This is a text-independent experiment because the speech data, among different speakers or among the training and testing data of the same speaker, are generated from different texts.

Table 2 gives the results. The matching accuracy is much higher than 50%, indicating the potential of the feature set in speaker recognition. When testing data size is 100 utterances, the averaged accuracy is about 70%; but if the testing data size is reduced to half, the accuracy is about 11 percent lower, and the accuracy variation also grows. The results depend on the testing data size, which implies the variation of each feature is also important. Gaussian model is expected.

Table 2: *Results of template matching*

Testing data size	Testing times	Averaged accuracy	Std. Variation of accuracy
100	15	69.51%	3.5%
50	30	58.48%	5.26%

4.3. Experiment 2: Clustering

This experiment is a kind of analysis. The target is to find if the speakers are in different clusters, what is the difference of their tone systems? Would their speech be perceived very differently?

Clustering is conducted over the speakers' templates. Each speaker's template is built up based on all the 300 utterances. Experienced with experiment 1, each feature in the 7-dimension feature set is modeled by a Gaussian, which is defined by mean and variance of all occurrences of a feature. The distance measure among templates is Kullback distance (KLD), which measures the distance between multi-dimensional Gaussians. It is shown in the following formula:

$$d(p,q) = \frac{1}{2}tr[\left(\sum_p^{-1} + \sum_q^{-1}\right)(\mu_p - \mu_q)(\mu_p - \mu_q)^T + \sum_p \sum_q^{-1} + \sum_q \sum_p^{-1} - 2I] \qquad (3)$$

where p,q are multi-dimensional Gaussian template. In clustering, single linkage hierarchical clustering is adopted, with the consideration that we do not know the actual distribution of templates and we do not know how many clusters are approximate. Hierarchical clustering provides a good solution for our consideration. This kind of clustering produces a set of nested clusters organized as a hierarchical tree called dendrogram. By observing the tree, any number of clusters can be easily obtained by cutting the link from the highest distance to the lowest distance.

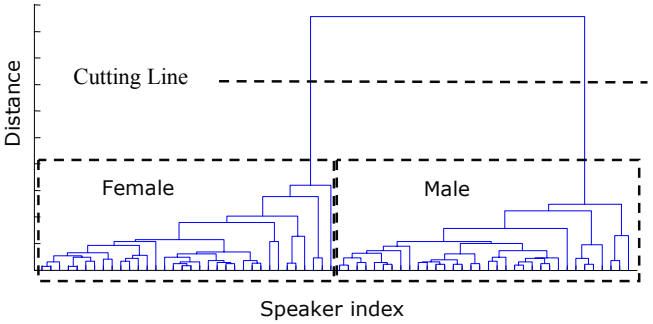

Figure 5: *Dendrogram of hierarchical clustering*

Figure 5 is the dendrogram of templates clustering. If two clusters are wanted, just split the link with the highest distance. Then two clusters with similar density appear. Checking the members in each cluster, we got a very clean result. One cluster totally consists of female speakers' templates and another one contains only male's. This result primarily proves that we control the features and models in a right way.

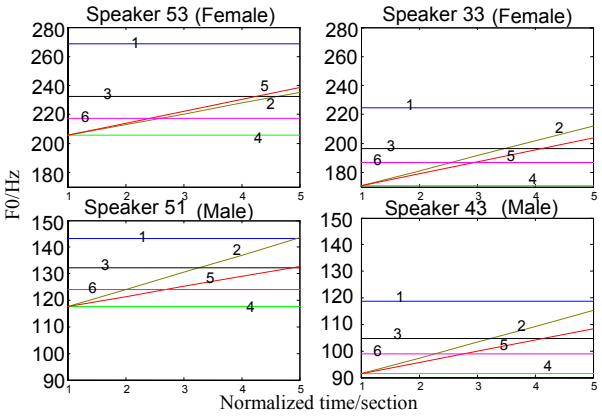

Figure 6: *Visualization of feature sets from the templates in close clusters*

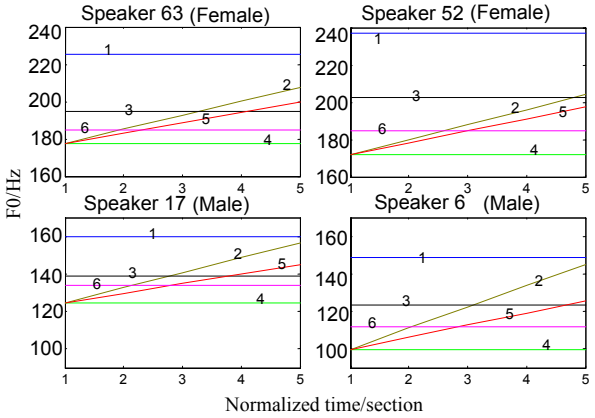

Figure 7: *Visualization of feature sets from the templates in far away clusters*

Concentrating on intra-gender difference, for each gender the templates in two clusters are compared. According to the mean parameter of each feature in the template, Figure 6 gives a visual comparison of the feature sets, by approximating the tone system. The height of Tone k is obtained from H_3 / R_{3k}. Level tones are plotted by extending their heights along the five time sections. Rising tones are expressed as a line decided by the height of Tone 4 at the first section and the height of their own at the third section. Both the two female and male speakers are from close clusters. It is found that if the templates are from two close clusters, their difference would rest with the difference of F0 register. While if the templates are from two clusters with large distance, as shown in Figure 7, the great difference both in range and relative position of each tone is observed. It indicates the ratios in two templates are very different. An informal perceptual test is carried out to investigate the speech data of the selected speakers. In terms of tone system perception, if the speeches of the two speakers are from close clusters, perceived difference is not so obvious. While if the speeches of the two speakers are from far away clusters, great difference will be perceived. The result proves the proposed acoustical measurement is consistent with the perceptual measurement. F0 register is not the most important feature to decide perceived difference, while tone ratios are more important towards this kind of perception.

5. Conclusion

This paper describes a preliminary study on the role of Cantonese tone system in speaker recognition, focusing on a feature set which consists of F0 register and relative tone ratios. The investigation is conducted over both simple recognition test—template matching and acoustical and perceptual analysis—clustering. Results primarily proves the importance of Cantonese tone system in speaker recognition. Not only the extrinsic F0 feature but the intrinsic realization of the whole tone system is important. Informal perceptual test confirms the proposed feature set is effective for modeling the Cantonese tone system in speaker characterization.

Further improvement is much possible. Because at present stage, the F0 information in speech is not fully used, only a part of them is used in feature extraction. Towards the tone system, the feature set only tries to model level tones, but not to fully model rising tones. The last point, better model for each feature can be proposed.

6. Acknowledgement

This research is partially supported by an Earmarked Research Grant (Ref: CUHK 4227/04E) from the Hong Kong Research Grants Council.

7. References

[1] Sonmez, K., Shriberg, E., Heck, L. and Weintraub, M., 1998. Modeling dynamic prosodic variation for speaker verification, *Proceedings of ICSLP 1998*, vol. 7, 3189-3192.

[2] Atal, BS, 1976. Automatic recognition of speakers from their voices, *Proceedings of the IEEE 64* , 460 - 475.

[3] Wong, Patrick C.M. and Diehl, Randy L., 2003. Perceptual normalization for inter- and intratalker variation in Cantonese level tones, *in the Journal of Speech, Language, and Hearing Research*, vol. 46, 413-421.

[4] Moore, C.B. and Jongman A., 1997. Speaker normalization in the perception of Mandarin Chinese tones, in the *Journal of Acoustical Society of America*, vol. 102(3), 1864-1877.

[5] Syrdal, A.K, 1996. Acoustic variability in spontaneous conversational speech of American English talkers, *Proceedings of ICSLP 1996*, 438-441.

[6] John, C. and Yallop, C., 1990. *An Introduction to Phonetic and Phonology.* Cambridge, MA: Basil Blackwell.

[7] Li, Y.J., 2003. *Prosody Analysis and Modeling for Cantonese Text-to-Speech.* MPhil. Thesis, Department of Electronic Engineering, the Chinese University of Hong Kong.

[8] Li, Y.J., Tan Lee and Qian, Y., 2004. F0 analysis and modeling for Cantonese text-to-speech, in *proceedings of the International Conference on Speech Prosody 2004*.

[9] Li, Y.J., Tan Lee and Qian, Y., 2004. Analysis and modeling of F0 contours for Cantonese text-to-speech, in *the Journal of ACM Transactions on Asian Language Information Processing*, volumn 3, issue 3, 169-180, September, 2004.

[10] 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).

[11] Talkin, D. and Lin, Derek, "ESPS/waves online documentation", Entropic Research Laboratory.