Variation in Spectral Slope and Interharmonic Noise in Cantonese Tones

Phil Rose

Australian National University Emeritus Faculty, Australia
philjohn.rose@gmail.com, https://philjohnrose.net

Abstract

To provide reference data for studies of voice quality variation in lexical tone, an experiment is described to investigate the nature of intrinsic variation in spectral slope and interharmonic noise for Cantonese citation tones. 23 spectral slope and interharmonic noise measures are extracted with *VoiceSauce* from the tones on /o/ Rhymes of five male and five female speakers of conservative Cantonese. Significant correlation between F0 and both spectral slope and interharmonic noise is demonstrated. It is shown with probabilistic bivariate discriminant analysis that even tones with no extrinsic voice quality differences can be identified at rates considerably above chance from a combination of their spectral slope and interharmonic noise. Male tones, with a minimal error rate of 5.7%, are identified twice as well as female, with a minimal error rate of 14.5%. Combinations with uncorrected spectral slopes perform better than corrected. The best combinations for both sexes involve slope parameters *H2H4* (difference between the 4th and 2nd harmonic amplitudes); and *H42K* (difference between the 4th harmonic and nearest harmonic to 2 kHz), irrespective of noise parameters. The worst combinations involve *CPP (*cepstral peak prominence) as a noise parameter.

Index Terms: Cantonese citation tones, voice quality, spectral slope, HNR-ratio, phonation type.

1. Introduction

Although pitch has long been assumed the main perceptual feature for tonal contrasts [1, 2, 3], it is well-known that other features, both segmental and suprasegmental, can interact with tone. One such feature is phonation type: many so-called *mixed* or *tonatory* tone systems, especially in SE Asia, include tones primarily characterized by extrinsic – i.e. deliberately produced – phonation types like creak, whispery voice, glottal-stop or growl [4, 5, 6, 7]. (The term *extrinsic* in the sense of *deliberately produced* was introduced in [8] as part of a classification scheme for allophones.)

Extrinsic phonation types result from deliberate adjustment to laryngeal structures which cause the vocal folds, and often other laryngeal structures like the ventricular or aryepiglottic folds, to vibrate in different ways [9]. Creaky voice, for example, is produced by constricting the epilaryngeal tube at the level of the ventricular folds [9]. Aerodynamically, different phonation types will be manifested in the volume velocity at the glottis and epilarynx, which will in turn be reflected in the distribution of energy in the source spectrum [10].

It is of phonetic interest, and indeed it is common nowadays, to seek the acoustic correlates of phonation types, for example, tense voice in Yi [11], breathy and slack voice in Chinese Wu and Gan dialects [12, 13, 14], or creak in Hmong [15]. This is done not only when phonation type functions as a short-term phonological property, but also in its longer-term para-linguistic and non-linguistic functions, where it is treated as an aspect of voice quality [9, 16, 17]. Given that phonation type is manifested in aspects of the source spectrum, usual acoustic measures are spectral, falling broadly into the two classes of spectral slope (also called spectral tilt) and interharmonic noise.

Now, although extrinsic tonation is undoubtedly more fascinating for the light it sheds on the degrees of freedom of the human larynx as articulator, this paper is concerned with the more mundane aspect of variation in tonal production in the *absence* of extrinsic phonation type. For the way the vocal-folds vibrate can differ in the course of producing different tonal F0 itself. Tonal F0 is produced primarily by controlling vocal-fold tension with crico-thyroid, thyro-arytenoid and strap muscles [18], and this muscle activity in turn can affect the way the vocal folds vibrate. For example, increased thyro-arytenoid activity "bulges the vocal-fold body and creates a thicker and deeper vibrating structure" [19]. The aim of this paper is to investigate the extent to which different tones are associated with different intrinsic voice quality features of spectral slope and interharmonic noise (abbreviated as *slope* and *noise* below).

There are two main reasons why it is important to know about this kind of variation. Firstly, such changes can be important for the perception of tonal pitch from F0. It has been shown in [20] for example that the same F0 will be perceived with a higher pitch if it is produced with a less steep spectral slope, and in [21] that the contrastivity of tones with similar pitch values can be enhanced by different phonation types - a nice demonstration that pitch perception is mediated by factors other than F0 [22]. Secondly, if slope and noise can vary intrinsically with phonation, then using them to determine the presence of *extrinsic* phonation type, as is often done, is problematic in the absence of additional confirmatory auditory evidence. Before one can say from the acoustics alone whether a particular slope or noise value indicates extrinsic phonation type, one logically needs to know how likely one is to get that value if no extrinsic phonation type is present. This paper thus investigates the kinds of differences in noise/slope values one finds in the absence of extrinsic tonation.

Cantonese is a good variety for investigating this. On syllables ending in a sonorant, conservative Hong Kong Cantonese contrasts six tones [23]. Its three level-pitched tones are located at the top, in the middle and just below the middle of the speaker's pitch range. These will be referred to as *high level*, *mid level* and *lower-mid level* respectively. Its two rising tones start low in the pitch range, with one rising to high (*low to high rise*) and one to mid (*low to mid rise*). Its falling tone (*low fall*) starts low and falls still lower, such that its phonation type usually becomes creaky or breathy as it falls below the speaker's modal pitch range. More importantly,

Cantonese is not known for extrinsic tonation: its tone system is based on pitch contrast alone [24]. To be sure, its low falling tone often becomes breathy or creaky, but that occurs intrinsically as the result of the low falling pitch target.

2. Procedure

2.1. Speakers and corpus

Data for analysis were taken from recordings of five female and five male speakers of conservative Hong Kong Cantonese. The speakers are referred to below as CM(ale) / CF(emale) 1-5. The recordings were made in the late nineties, when the informants were all young linguistics students at the Australian National University. They reported no speech pathology. The data-set was originally used for a linguistic-phonetic analysis of Cantonese tones [25] and in the development of bionic-ear implants for Cantonese [26]. Mean F0 and duration measurements from the data-set may be downloaded from the author's website.

The corpus itself consisted of a list of Chinese characters, with dummy characters inserted to minimize listing effects, which each informant read out at least four times. Details are in [25]. This paper uses only morphemes from the list with the Rhyme /o/, which has the allophone [ɔː]. A mid vowel monophthong, rather than a high or low vowel, was chosen to avoid problems with estimating spectral slope in segments with a low frequency pole, where source-filter theory suggests its transfer function might obscure the relationship of the amplitudes of the lower harmonics of the source spectrum. A low frequency pole will of course occur as the first formant in high vowels, and therefore high vowel Rhymes were not used; but a low frequency pole will also occur with low vowels as a result of coupling to the nasal cavity where the velic port can be pulled open to a certain extent from the low tongue position. Thus low vowels were also best avoided. Finally, the monophthongal quality of /o/ minimizes any change in voice quality parameters that might arise as a function of filter change. There were no usable low to mid tone morphemes in the data set, so the corpus comprised a subminimal tonal quintuplet with /o/ and a voiceless Onset stop. This is given in table 1. It can be seen the morphemes all have a voiceless stop, which is aspirated in the case of the low falling tone, but otherwise unaspirated (Cantonese phonotactics do not allow a voiceless unaspirated stop in the low falling tone).

Table 1: /o/ *tokens analysed. M = morpheme, R = Yale romanisation, P = phonetic representation.*

Tone	M		P	R
High level	歌	*song*	kɔ˥	gō
Mid level	個	*classifier*	kɔ˧	go
Lower-mid level	惰	*lazy*	tɔ˨	doh
Low fall	婆	*woman*	pʰɔ˩	pòh
Low to high rise	躲	*hide*	tɔ˨˧	dó

Cantonese has been undergoing considerable phonological change, including tonal mergers [24]. Although all speakers contrasted six tones, it was clear from listening that, even in the late nineties, some did not distinguish certain morphemes, especially on /aa/ and /ei/. A native-speaking Cantonese phonetician was therefore recruited to check that all /o/ tokens were produced with the appropriate tones.

2.2. Processing

Recordings were digitised at 16K, the /o/ tokens in them identified in *Praat* and their phonation onset and offset labeled in a textgrid. The *VoiceSauce* package [27] was then used to extract a suite of acoustic parameters intended for voice quality quantification.

VoiceSauce provides for extraction of the conventional acoustic parameters of F0, and formant centre frequencies and bandwidths, estimated with several different algorithms. The manual has a default extraction with *snack*, but F0 and F-pattern (formants 1 – 3) were extracted with both *Praat* and *snack* algorithms with, following forensic procedure, optimum settings chosen separately for each speaker. Visual checking showed that *snack* usually extracted F0 better, but since neither procedure was clearly superior in F-pattern extraction, the manual was followed and *snack* used. *VoiceSauce* estimates harmonic-to-noise ratios over four spectral regions from dc to an increasing upper bound of 0.5 kHz, 1.5 kHz, 2.5 kHz and 3.5 kHz. V*oiceSauce* names these *HNR05*, *HNR15*, *HNR25* and *HNR35* respectively. All these regions were used. *VoiceSauce* employs several different ways of estimating spectral slope, mostly from differences between harmonic amplitudes (where H1 = F0). Two measures select harmonics without reference to formant centre frequencies: *H1H2*, *H2H4*. Three measures compare H1 with the harmonic nearest to one of the first three formants' centre frequency: *H1A1*, *H1A2*, *H1A3*. Four measures select harmonics closest to a given frequency: *H2K H5K H42K H2KH5K*. All these harmonic measures are made both corrected and uncorrected for an all-pole LPC transfer function. One *VoiceSauce* measure is based on the cepstrum: *CPP*, or cepstral peak prominence. This is the amplitude difference between a peak in the cepstral power spectrum and the value of a trend line at the same quefrequency. Originally an automatic measure of dysphonia, it is now commonly used as a general voice quality measure. In all, (9 uncorrected slope + 9 corrected slope + 5 noise =) 23 measures in *VoiceSauce* were used. Parameters were estimated every millisecond and imported into *R* for further processing.

3. Results

3.1. Illustrative time-courses

To give a general idea of the results, figure 1 illustrates, with the five repeats of CM1's five tones, typical time-courses for noise and slope during the /o/. I have chosen for this just two of the many slope and noise measures available: H1H2c and HNR05. The bottom two panels show the slope and noise trajectories, and, for orientation, F0 and F-pattern configurations are also given for the same data in the top two panels. The tones are colour-coded to facilitate comparison. In each panel thick lines indicate the arithmetical mean values over his five repeats, which are plotted with thin dotted lines.

The speakers' F0 shapes in the top panel have a typical configuration. In particular, the mid and lower-mid level tones lie fairly close in the middle of the F0 range with the high level tone at the top. Typical consonantally-induced falling onset perturbations of different magnitudes are seen over about the first 10 centiseconds. Offset perturbations are also evident over about the last five centiseconds, especially in the low falling tone. (These tend to be speaker-specific, indicating the way in which a speaker ends phonation.)

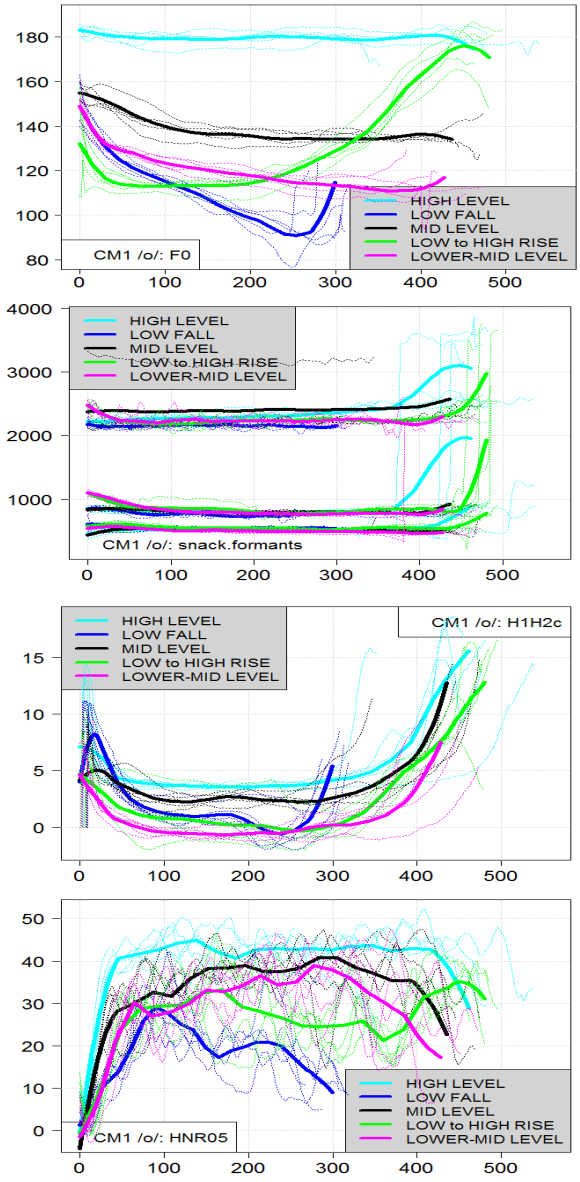

Figure 1: *Time-aligned F0, F-pattern, spectral slope and interharmonic noise trajectories for five repeats of five of CM1's sonorant-final tones on /o/ Rhymes. X-axis = duration (msec), thick lines = means, thin dotted lines = individual tokens*

The monophthongality of the /o/ can be seen in the F-pattern panel, with all formants showing little change through most of the Rhyme. Typical F2 onset perturbations associated with different place of articulation in the Onset are seen. It is also clear that F-pattern extraction was not always successful, especially at the end of the Rhyme in tones with high F0 (high level, low to high rise). F3 in one mid-level token can also be seen estimated with a frequency typical of F4, thus pulling the mean F3 value up. Taking these aberrations into account, F-pattern does not appear to vary with tone, apart from a possibly lower F3 in the low falling tone.

Turning now to the noise and slope data in the bottom two panels, the most important thing to note is that, just as with F0, CM1's different tones have different mean values for noise and slope. In particular it can be seen that for his three level tones both slope and noise appear to correlate with F0. The lower the F0, the greater the amount of relative noise and the smaller the difference between H1 and H2 amplitudes.

It is also evident from the noise and slope trajectories that, again like the F0, they all display similar onset and offset perturbations. In particular it can be seen that the slope parameter increases towards the end of the Rhyme. In the high level tone, for example, the slope parameter starts to increase at about 60% - 70% of duration. Since figure 2 shows that this tone has effectively level F0 and a static F-pattern, the slope changes are presumably due to changes in vocal fold vibratory behavior. There are also typical abrupt rising-falling perturbations in the first few centiseconds after phonation onset which tend to correlate with the above-mentioned F0 perturbations due to Onset properties. Trajectory changes in noise parameters are less marked but appear to also correlate with F0 contour.

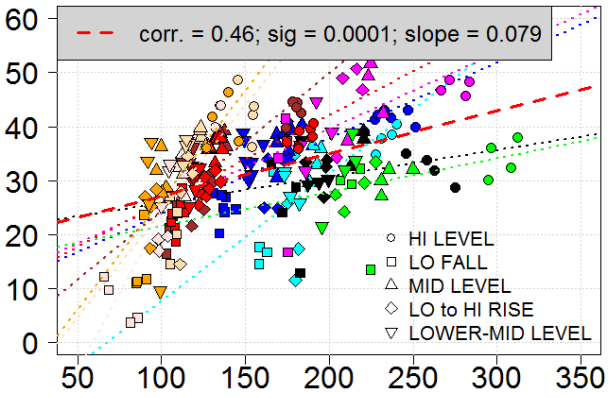

Figure 2: *Relationship between mean HNR05 and F0 values for all tokens of all speakers. Red dashed line = least squares fit with perturbations removed. X-axis = F0 (Hz), y-axis = HNR05.*

If intrinsic variation exists between tonal F0 and noise/slope parameters, then one would expect to see it in correlations between F0 and slope, and F0 and noise. In order to estimate the correlation between F0 and the slope/noise parameters, the first 5 and last 8 centiseconds of the Rhyme data in each token were removed to control for onset and offset perturbations, and mean values then estimated for the remaining Rhyme values. Illustrative results are shown in figure 2 for HNR05 against F0 for the five tones of all 10 speakers. Speakers are colour-coded; tones are plotted with different symbols. Least squares lines are fitted for each speaker (dotted) and for the data overall (red dashed). It can be seen that F0 and HNR05 are positively correlated both within- and also across speakers, where the correlation of 0.46 is very highly significant. Many other *VoiceSauce* parameters also showed very highly significant correlation with F0.

3.2. Prediction of tonal categories from slope and noise measures.

Rather than just assessing their degree of correlation, a more insightful way of interpreting the relationship between tonal F0 and these noise and slope values is to determine how well the three level tones can be statistically identified from them.

In order to do this, a portion of each token was selected from 20% to 70% of Rhyme duration to exclude perturbatory effects, and mean slope and noise values calculated over this

portion. These data – mean values for all speakers and all noise/slope combinations – can be downloaded from the author's web-site. Because of the small number of tokens for most speakers, which will tend to give overly optimistic identification, 100 slope and noise values per speaker per tone were then bootstrapped from the mean and standard deviation of each speaker's sample. These bootstrapped values were then analyzed with probabilistic Fisher discriminant analysis (PFDA), with mean slope and noise measures as bivariate predictor variables and tonal category (*high level*, *mid level*, *lower-mid level*) as predicted. (PFDA relaxes the homoscedasticity constraint of linear discriminant analysis and is reported to give equal or better results [28, 29].) Tone identification was tested with all (18 slope * 5 noise =) 90 noise/slope combinations. Figure 3 is an example of PFDA on CM2's tones with slope and noise parameters of H1H2c and HNR05. The top panel shows the distribution, in the H1H2/HNR05 plane, of the three tones' original tokens and the values bootstrapped from them. It can be seen that the mid level tone and the lower-mid level tone overlap in both dimensions, and the high level tone and mid level tone overlap in noise dimension. The percent identification error rate of each speaker was determined separately, for each combination - for the combination in figure 3 it is 6.3% - and an overall performance for each noise/slope combination obtained from the mean of the 10 speakers' individual error rates.

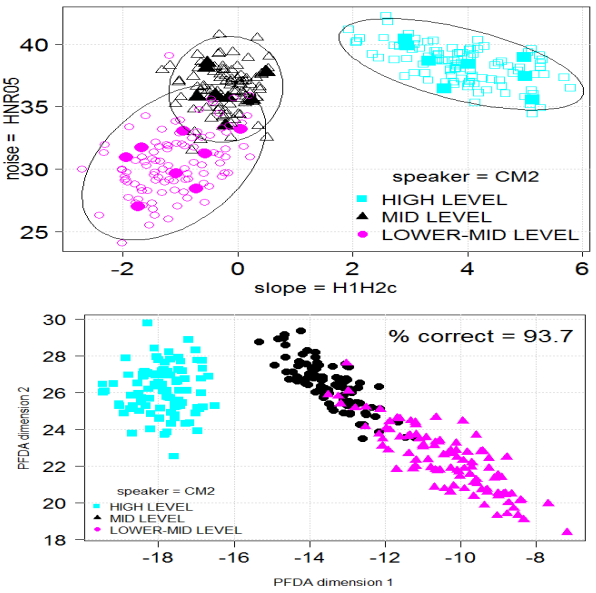

Figure 3: *Demonstration of Probabilistic Fisher Discriminant Analysis with CM2's level tones. Top = raw data (solid symbols = raw values, empty symbols = bootstrapped data). Bottom = PFDA resolution*

These overall mean error rates were then analysed with ANOVA using Tukey HSD as post-hocs. Results showed that male tones could be substantially better identified than female tones (df=1, F=10.38, p = 0.001). The best male error rate was 5.7%, achieved with a slope/noise combination of H2H4u and HNR35, compared to the best female error rate of 14.5% with a slope/noise combination of H2H4 and HNR05. The mean error rates over all slope/noise combinations were 12.8% for males and 26.7% for females. Surprisingly, perhaps, combinations with uncorrected slopes performed significantly better than corrected.

Figure 4 shows 3d bar charts of males' and females' mean error rates with uncorrected slope combinations. The difference between male and female tones is readily visible (the female vertical axis showing error rate is twice the height of the male). But there are also clear similarities between the sexes. It can be seen that, of the noise parameters, combinations with CPP (along the back) generally perform the worst for both sexes, and good identification rates are obtained with slope parameters of H2H4 and H42K (4th and 8th columns) irrespective of which noise parameter is used.

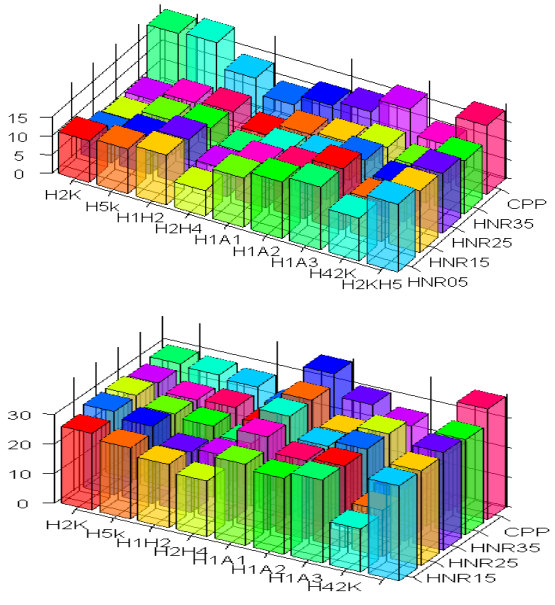

Figure 4: *Mean percent error rates for PFDA identification of the 3 Cantonese level tones from noise and slope parameters. Top= 5 males, bottom = 5 females. Vertical axis = error rate, x- axis = 9 slope parameters, y-axis = 5 noise parameters.*

4. Summary

This paper has investigated intrinsic voice quality in Cantonese tones and shown that spectral slope and interharmonic noise vary, even in tones that are not associated with any extrinsic phonation type. The variation is sufficiently systematic, and the parameters sufficiently independent, to enable the three level tones to be identified from combinations of slope and noise. Future work must extend the analysis to different Rhymes and other dialects. The most important consequence of these findings is this. Given that the *VoiceSauce* parameters have been shown to vary with tonal F0 in the absence of extrinsic voice quality, it is logically not possible to claim the existence of extrinsic voice quality simply on the basis of observed differences between tones in *VoiceSauce* parameters. The ideal solution to the problem would be to be able to estimate a likelihood ratio for an observed difference [30]. The present data at least permit the estimation of one half of such a likelihood ratio: the probability of getting an observed *VoiceSauce* parameter difference assuming no extrinsic voice quality is present. Perhaps a more practical solution – rather perversely, given this paper's focus! – is to choose those *VoiceSauce* parameters which show the *least* correlation with tonal F0. Figure 4 suggests CPP may be a good candidate.

5. References

[1] K.L. Pike, *Tone Languages*. Ann Arbor: The University of Michigan Press, 1948.

[2] V.A. Fromkin, *Tone - A linguistic Survey*. New York: Academic Press, 1978.

[3] L. Hyman, "Word prosodic typology", *Phonology*, vol. 23, pp. 225 – 257, 2006.

[4] E.J.A. Henderson, "The topography of certain phonetic and morphological characters of South East Asian languages," *Lingua*, vol.15, pp. 400 – 434, 1965.

[5] P. Rose, **"**Tonation in Three Chinese Wu Dialects," in *IChPS 2015 18th International Congress of Phonetic Sciences*, Glasgow, UK, Proceedings, 2015, no page nos.

[6] J. Kirby and M. Brunelle, "Southeast Asian tone in areal perspective," in R. Hickey (ed.) *The Cambridge Handbook of Areal Linguistics*, Cambridge Handbooks in Language and Linguistics, Cambridge: CUP, pp. 703–731, 2017.

[7] P. Rose, **"**Tonatory Patterns in Taizhou Wu Tones," in *IChPS 2019 19th International Congress of Phonetic Sciences*, Melbourne, Australia, Proceedings, 2019.

[8] M.A.A. Tatham, "Classifying Allophones," *Language and Speech*, vol. 14, pp. 140–145, 1971.

[9] J.H. Esling, S.R. Moisik, A. Benner and L. Crevier-Buchman, *Voice Quality – The Laryngeal Articulator Model.* Cambridge Studies in Linguistics 162, Cambridge: CUP, 2019.

[10] C. Gobl and A. Ní Chaside, "Voice Source Variation and its Communicative Functions," in W.J. Hardcastle, J. Laver and F.E. Gibbon (eds.), *The Handbook of Phonetic Sciences* 2nd ed., pp. 378–423, Chichester: Wiley-Blackwell, 2013.

[11] I. Maddieson, S. Hess, "The effect on F0 of the Linguistic Use of Phonation type", *UCLA Working Papers in Phonetics: Studies of Phonation types,* vol. 67, pp. 112–118 , 1987.

[12] J.F. Cao and I. Maddieson, "An exploration of phonation types in Wu dialects of Chinese," *J. Phonetics*, vol. 20, pp. 77–92, 1992.

[13] J-Y. Gao, P. Hallé, K. Honda, S. Maeda and M. Toda, "Shanghai Slack Voice: Acoustic and EPGG data," in *IChPS 2090 - 17th International Congress of Phonetic Sciences*, *Hong Kong, China, Proceedings*, 2009, pp. 719–722.

[14] Y.Y. Zhou, *"The Tonal Typology and Evolution of Northern Gan – A Case Study of Xiushui,"* unpublished Ph.D. thesis, Hong Kong University of Science and Technology, 2020.

[15] J. E. Andruski and M. Ratliff, "Phonation types in production of phonological tone: the case of Green Mong," *Journal of the International Phonetic Association,* vol. 30, pp. 37–61, 2000.

[16] P. Keating, M. Garellek and J. Kreiman, "Acoustic properties of different kinds of creaky voice," in *IChPS 2015 - 18th International Congress of Phonetic Sciences*, *Glasgow,* UK, Proceedings, 2015, no page nos.

[17] L. C. Rusilo, Z.A. de Camargo and S. Madureira, "*The validity of some acoustic measures to predict voice quality settings: trends between acoustic and perceptual correlates of voice quality,"* in A. Botinis (ed.) ExLing 2011 - 4th International Speech Communication Association Tutorial and Research Workshop on Experimental Linguistics, May 25-27, Paris, France, Proceedings, 2011, pp. 115–118.

[18] D. M. Erickson, "*A Physiological Analysis of the Tones of Thai*," unpublished Ph.D. thesis, University of Connecticut, 1976.

[19] I.R.Titze, *Principles of Voice Production*, New Jersey: Prentice Hall, 1994.

[20] J.J. Kuang and M. Libermann, "Integrating Voice Quality Cues in the Pitch Percpetion of Speech and Non-speech Utterances," *Frontiers in Psychology* vol. 9, Article 2147, 2018.

[21] J.J. Kuang, *Phonation in Tonal Contrasts*. Unpublished Ph.D. thesis, UCLA, 2013.

[22] P. Rose, "On the non-equivalence of fundamental frequency and pitch in tonal description," in D. Bradley, E. Henderson and M. Mazaudon, (eds.) *Prosodic Analysis and Asian Linguistics: to Honour R.K. Sprigg.* Pacific Linguistics, pp. 55–82, 1989.

[23] S. Matthews, V. Yip, Cantonese – A Comprehensive Grammar. London: Routledge, 1994.

[24] P.P.K. Mok, H.S.H. Fung and V.G. Li, "Assessing the Link Between Perception and Production in Cantonese Tone Acquisition," *Journal of Speech, Language, and Hearing Research*, vol. 62, pp. 1243–1257, 2019.

[25] P. Rose, "Hong Kong Cantonese Citation Tone Acoustics: A Linguistic-Tonetic Study," in M. Barlow (ed.) SST 2000 - *8th Australian International Speech Science and Technology Conference*, Proceedings, Canberra, Australia, 2000, pp. 198–203.

[26] J. G. Barry, "*Speech Development in Profoundly Hearing Impaired Cantonese-Speaking Children Using a Cochlear Implant*," unpublished Ph.D. Thesis, Department of Otolaryngology, University of Melbourne, 2002.

[27] Y.L. Shue, P. Keating, C. Vicenik and K. Yu, "VoiceSauce: A Program for voice analysis," in ICPhS 2009 - *17th International Congress of Phonetic Sciences*, *Hong Kong, China, Proceedings*, 2009, pp.1846–1849.

[28] C. Bouveyron and C. Brunet, "Probabilistic Fisher discriminant analysis: A robust and flexible alternative to Fisher discriminant analysis," *Neurocomputing* vol. 90, pp. 12–22, 2012.

[29] S. Ioffe, *"Probabilistic Linear Discriminant Analysis,"* in A. Leonardis, H. Bischof and A. Pinz (eds.), *ECCV 2006 - 9th European Conference on Computer Vision*, pp. 531–542, 2006.

[30] P. Rose, "The Acoustics and Probabilistic Phonology of Short-stopped Syllable Tones in Hong Kong Cantonese". In S. Cassidy (ed.) SST 2004 - *10th Australian International Speech Science and Technology Conference*, Proceedings, pp. 445–450, 2004.