Title: The tonal space of contrastive five level tones

Abstract

Multiple level-tone contrasts are typologically disfavoured because they violate the dispersion principles of maximizing perceptual distance and minimizing articulatory effort. This study investigates the tonal dispersion of a multiple level-tone system by exploring the cues used in producing and perceiving the five level tones of Black Miao. Both production and perception experiments show that non-modal phonations are very important cues for these tonal contrasts. Non-modal phonations significantly contribute to the dispersion of the five level tones in two ways: either by enhancing pitch contrasts or by providing an additional contrastive cue. Benefiting from more than one cue, the level tones T11, T33 and T55 are well distinguished in the tonal space; by contrast, the level tones T22 and T44, only contrasting in pitch, are the most confusable tones. The tonal registers model proposed in this paper sheds light on the different uses of non-modal phonations across languages.

1. Introduction

How many contrasting pitch levels can tonal languages have? Linguists [e.g., Chao, 1948; Maddieson, 1978] have observed that while people are able to produce many phonetically different levels of pitch in speech, no known language makes a phonological contrast of more than five pitch levels. Five-level phonetic transcriptions of tones, e.g., Chao's numbers and the IPA tone markers, have been found to be sufficient for known tonal languages. In fact, typologically the number of contrastive levels is even more restricted. According to Maddieson's [1978] cross-linguistic survey, five- and four-level tonal languages are extremely rare, compared to languages with fewer contrastive levels. A two-level contrast is the most frequently attested type among tonal languages.

Why is the number of possible contrastive levels so limited? Why do languages generally prefer fewer contrastive tonal levels? Dispersion Theory [Lindblom, 1986, 1990] and similar views [Martinet, 1952, 1955; Lindblom and Maddieson, 1988; Flemming, 2004] can shed light on these questions. The basic idea of Dispersion (including Dispersion theory and H&H theory) is that the structure of inventories is subject to two goals: maximize perceptual contrasts but minimize articulatory effort. So an optimal inventory space is the counterbalance between the constraints of speakers and listeners. In other words, the form of the tonal space should follow from its function of maximizing phoneme contrasts while requiring minimal articulatory effort.

Although it has been well accepted that speech communication should be effortless, previous attempts to implement "perceptual ease" and "articulatory ease" in dispersion theory have been challenged. On the one hand, if one strictly holds the principle of maximizing distinctions among the categories, languages may end up with exotic inventories (e.g., multiple non-peripheral high vowels; a consonant inventory consisting of one click, one implosive, one ejective, etc.) [Ohala, 1983]. To fix this problem, Lindblom [1986, 1990] revised the original proposal by replacing the notion of maximal

contrast with sufficient contrast, which took articulatory effort into account. However, on the other hand, how to define "articulatory effort" then becomes a challenge. It has been found impossible to reliably measure "how much" effort is made [c.f. Pouplier, 2003]. A more plausible way of explaining the counterbalance between perception and production is proposed by Lindblom and Maddieson [1988]: the phonetic space is partitioned into several dimensions based on "articulatory complexity" (from simple to complex: basic articulations, elaborated articulations and complex articulations). In order to make sufficient perceptual distinctions, languages first push the boundaries of the basic dimensions; if still not sufficient, then additional dimensions are recruited for additional contrasts, requiring greater effort but providing more contrasts [Lindblom and Maddieson, 1988, p. 71]. In this way physiological limitations of both production and perception can be taken into account: phonological contrasts should have reliable perceptual differences but lie within a comfortable production range.

Thus, if we consider the limitations of pitch production and perception, we can understand why multi-level tone systems are disfavoured: to maintain multiple level-pitch contrasts is extremely hard in speech communication. On the one hand, the pitch range used in normal speech is fairly small, said to be usually no more than 100 Hz [Baken and Orlikoff, 2000; Keating and Kuo, 2012, for isolated words and passages]. We can validate this pitch range estimate by a quick survey in the cross-linguistic corpus from the UCLA voice quality project. Multi-speaker recordings and a large range of voice measures [Keating et al. 2012] are available online [http://www.phonetics.ucla.edu/voiceproject/voice.html]. F0 has been calculated using the STRAIGHT algorithm [Kawahara et al., 1999]. The mean F0 value of each token produced by each male speaker from nine languages (both tonal and non-tonal) was retrieved. The boxplot in Figure 1 displays the overall distributions of these F0 values. As can be seen, across languages the overall pitch ranges (between the upper and lower whiskers) for male speakers are fairly similar, mostly around a 100 Hz range, lying between 80 Hz and 180 Hz;

medians are all around 140 Hz. This means that the physiological limits of pitch-range production are quite universal across languages. The physiological reason for a 100 Hz range that has been proposed is that this is about the pitch range for modal (i.e. with the least articulatory effort) vocal register [Hollien and Michel, 1968; Hollien, 1974; Titze, 1988; Baken and Orlikoff, 2000]. Although some tonal languages tend to have slightly larger ranges (e.g., Bo and Mazatec, compared to English), the differences are small, suggesting that the pitch range is rather physiologically restricted, at least at default physiological settings. (Of course, pitch ranges can be enlarged by extra vocal effort or stronger subglottal pressure, for example for exclamatory utterances.)

(Figure 1 about here)

On the other hand, the just-noticeable-difference (JND) for lexical tone appears to be not less than 9 Hz [indicated by Silverman 2003: Figure 17.3], and languages usually require a larger difference than the JND to maintain a phonological contrast; a 20-30 Hz difference (2-3 semitones in a comfortable pitch range for males from 150 Hz to 170 Hz, 2.1 st, or for females from 200 Hz to 230 Hz, 2.4 st) is just marginally sufficient [Hart, 1981]. Furthermore, Harris and Umeda [1987] found that JNDs are much greater in natural sentences. This is so even though the pitch JND for pure tones is fairly small, about 3 Hz for frequencies below 500 Hz [Kollmeier et al., 2008], i.e. speech pitch discrimination is much harder than non-speech pitch discrimination.

For example, Cantonese, a Yue dialect of Chinese, has four level tones (11, 22, 33, 55 in Chao's representation). A perception experiment [Mok and Wong, 2010] shows that tones 22 and 33 are the most confusable in perception, though they still have a pitch difference of 30 Hz. Their language survey shows that young speakers frequently merge these two tones. Therefore, three levels appear to be the maximum contrasts that listeners are able to perceive within a pitch range of 100 Hz. Four or

five levels of pitch contrasts violate both dispersion principles: so many levels cannot be produced differently enough within the modal range to be reliably distinctive.

How then is a five-level-tone system possible, since several such languages have been reported [Maddieson, 1978; Edmondson and Gregerson, 1992]? According to Dispersion Theory, to maintain the maximum contrasts, one possible consequence of increasing the size of inventories is that the overall acoustic space of the inventories is enlarged. This is true for vowel spaces. Cross-linguistic studies on the dispersion of vowels [Lindblom, 1986; Becker-Kristal, 2010; among many others] demonstrate that the size of the acoustic space is positively correlated with the size of vowel inventories. That is, languages with more vowels occupy a larger acoustic space than languages with fewer vowels.

Very few studies have addressed tonal dispersion, but Maddieson [1978, p. 339] observes a similar enlarging effect on tonal spaces: languages with more tonal levels tend to employ a relatively larger pitch range than languages with fewer tonal levels, as reproduced below (Table 1). For example, Yoruba, a three-level language, has an overall pitch range of 79 Hz, while Toura, a four-level language, has an overall pitch range of 90 Hz.

(Table 1 about here)

However, a recent quantitative cross-linguistic investigation [Alexander, 2010] claims that tonal space size is rather fixed across level-tone languages, and that size of inventory has little effect on size of pitch range. For example, the pitch range of Cantonese (four-level) is not significantly different from that of Yoruba (three-level) and Igbo (two-level) at mid-point and offset positions of the tones. However, as Alexander calculated these pitch ranges across level and contour tones combined, they

actually reflect the overall pitch ranges of the languages instead of the dispersion of just the level tones. Moreover, finding fixed overall pitch ranges is consistent with our survey in Figure 1. Therefore, it can be inferred from Alexander's study that the ability of speakers to expand the pitch space is rather limited.

The second way to optimize the dispersion of large inventories is to add contrastive dimensions. For example, Lindblom and Maddieson [1988] found that as the size of consonant inventories increases, more and more articulatory dimensions are utilized. In the case of tone, then, in order to incorporate more tonal contrasts than an enlarged pitch range can provide, the second possibility is to rely on cues other than pitch. One of these cues that can contribute to tonal contrasts is duration. Duration is a distinctive cue for tones (as well as for vowel lengths) in some languages, e.g., in Oujiang Wu [Rose, 2008]; and it is also an enhancement cue for the citation tone 35 vs. tone 213 contrast of Mandarin [Tseng, 1981; Blicher et al., 1990]. Pitch contours are the other most common cues for tonal contrasts (e.g., Chinese dialects, Vietnamese dialects, Thai dialects); even for level tones, slight phonetic falling is also commonly found, especially for low tones [Zhu, 2012]. Finally, another, less addressed, cue is phonation, which is commonly found in many tonal languages and has been found to be a salient cue in perception. Sometimes phonation functions as an allophonic cue, e.g., creaky voice on the low tone of Mandarin [Belotel-Grenié and Grenié, 1994; Yang, 2011] and Cantonese [Yu and Lam, 2011]; other times phonation functions as a phonemic dimension in addition to pitch: Green Mong [Andruski, 2006], White Hmong [Esposito, 2012; Garellek et al., 2012], Southern Yi [Kuang, 2011], and Northern Vietnamese [Brunelle, 2009], to cite just a few.

When all the dimensions that can contribute to tonal contrasts are considered, it becomes non-trivial to model the tonal space in which dispersion can be understood. Previous studies of tonal dispersion have instead used very simple spaces. In comparing the production effort for tones in normal vs. noisy

environments, Zhao and Jurafsky [2007, 2009] modeled tonal dispersion as variability of the overall pitch range. Adopting this method, Alexander [2010] also used only a one-dimensional cue, i.e. mean F0, to define tonal spaces. As she noted, this method was not adequate to capture the perceptual separability of tonal contrasts. Tonal space models that allow contour cues significantly improve the separability of the tonal space. For example, in a study comparing Cantonese tone productions by normal-hearing adults, normal-hearing children and cochlear-implanted children, Barry and Blamey [2004] defined the tonal space by F0 onset x F0 offset. This method enables the authors to capture the dynamic factors, such as direction and slope, of the tones. Yu [2011] demonstrated that mean F0 + pitch change, rather than mean F0 alone, can better model the distribution of tonal inventories. Taken together, these studies suggest that a tonal space should incorporate multi-dimensional cues to reflect the actual perceptibility of tonal contrasts, contours as well as levels. However, no tonal space models so far have incorporated cues like duration and phonation. Since previous tonal studies have been limited to pitch, we will test two competing hypothesized models: pitch (+ duration) cues only vs. pitch + phonation (+ duration) cues.

In sum, the question asked in this paper is, given normal hearing and speaking ability, how can native speakers produce and hear multiple contrasting level tones? We will try to address this question by exploring the tonal production and perception of a language with five level tones, the most contrasting levels to our knowledge [Edmondson and Gregerson, 1992]. We will argue that these tonal contrasts are much more than pitch contrasts. When pitch contrasts get crowded, other cues must be involved to enhance the contrasts (e.g., phonation cues), resulting in an expanded tonal space.

2. Black Miao

The five-level-tone language that will be discussed in this paper is a Black Miao dialect, called Qingjiang Miao (Ch'ing Chiang Miao), belonging to the Hmong-Mien or Miao-Yao family, which is unrelated to the Sino-Tibetan languages that dominate in China. This Miao dialect is spoken at Shidong Kou (Shih-Tung-K'ou), Taijiang (T'ai-Kung) county of Guizhou (Kweichow) province in China. Figure 2 is a map showing the location of this language. This particular Black Miao dialect was first documented by Fang-Kuei Li in the 1940s [Kwan, 1966], and since then has been the most famous five-level-tone language in tonal studies [e.g., Yip, 2002; among many others]. I went to the same village to conduct the experiments reported here.

(Figure 2 about here)

According to Li's transcription, there are eight tones (I-VIII) in this dialect; five of them are level tones, two rising and one falling (using Chao's tonal representation), as shown in Table 2. Tone VIII (11), tone IV (22), tone VI (33), tone I (44), and tone III (55) are the five levels. There is no “neutral” tone as found in some Sinitic languages.

(Table 2 about here)

This paper is organized as follows: We start by examining whether the five level tones are well dispersed in a pitch-based tonal space (Section 3). A follow-up perception experiment is then presented to reveal the dispersion of tonal categories in listeners' perceptual space (Section 4). Finally, the dispersion of the five level tones is further examined in a tonal space incorporating phonation cues (Section 5). Tonal spaces with different dimensionalities will be compared.

3. A pitch-based tonal space

3.1 Production recordings

A wordlist of minimal monosyllabic sets for the eight tones was created based on Li's transcriptions, which were partially reported in Chang [1947] and Kwan [1966]. A list of 23 minimal sets of words was compiled from these sources. These words were first elicited from a fifty-year-old male speaker, who had the best knowledge about Black Miao; ten complete minimal sets were confirmed with him. These words were then checked and rehearsed with every participant in the production experiment, and some additional words not in complete sets were identified. Each speaker confirmed 100 to 120 monosyllabic words.

The wordlist was then elicited in minimal sets by the experimenter with each speaker; speakers were instructed to skip the items that they could not recognize and to note any items that in their judgment had identical pronunciations. To avoid tone sandhi in continuous speech, the test monosyllabic words were spoken in isolation. Some monosyllables are morphemes that do not normally occur by themselves, but speakers can say them if instructed to. Elicitation was carried out in Southwestern Mandarin, as all the speakers were able to understand and speak this local Mandarin dialect of Guizhou Province. Simultaneous electroglottographic (EGG) and audio signals were recorded in a quiet room, directly to a computer via its sound card, in stereo using Audacity. The sampling rate per channel was 22050 Hz. The audio signal, from a Shure SM10A head-mounted microphone, was the first channel of the recordings. The EGG signal, from a two-channel Glottal Enterprises Electroglottograph, model EG2, was the second channel. Each token was produced as many times as needed until two good repetitions were obtained.

A total of 14 native speakers of Black Miao were recorded. Nine male speakers are native speakers of this particular dialect; five females were also recorded, but they had married into this village and were not native speakers of this dialect. For the purpose of consistency, we therefore only report the data of

the male speakers here, but the data of the female speakers are available upon request. One of the male speakers failed to produce the tonal distinctions in most tokens, and thus is also excluded from the current analysis. Therefore, production results from eight native male speakers will be presented here.

3.2 Measurements

Pitch values of nine time intervals were automatically obtained by VoiceSauce [Shue et al., 2011], using the STRAIGHT algorithm [Kawahara et al., 1999]. For all tokens, pitch was measured across the complete pitch-carrying portion of the rime, as segmented in Praat textgrid files. In cases where the pitch tracking failed, the circumstances were noted (such as glottalization or breathiness), and these tokens were manually double checked in Praat [Boersma and Weenink, 2012]. In general, STRAIGHT was successful in extracting correct values for most tokens, even for many vocal fry tokens. Mean F0 values were calculated over nine time sub-intervals, and three contour-related pitch measures were made: onset (first ninth), offset (last ninth) and ΔF0 (the range of pitch change within a syllable). In addition, rime duration was obtained from the Praat textgrids.

3.3 Results – pitch analysis

A series of pairwise mixed-effect models were used to decide which measures significantly distinguish one tone from another, with mean F0, ΔF0, F0 onset, F0 offset and duration as the dependent variables. Duration does not contribute to any tonal contrasts, so time-normalized F0 is appropriate. Mean F0 is significantly different between every pair of eight tones. ΔF0 can distinguish contour tones (i.e. T51, T13, T45) from level tones (i.e. T22, T33, T44, T55), and ΔF0 of T11 is also slightly different from T22. F0 onset and offset mostly help distinguish contours with different directions. Therefore, the only pitch cue needed for the level tones is the average pitch value. Figure 3

shows the average pitch trajectories of the five level tones for eight male speakers. F0 is presented in Hertz in order to show the actual physical pitch space of these tonal categories.

(Figure 3 about here)

Since test words read in isolation carry the sentence-level intonation, we can see slight pitch declination for all the level tones. The five tones under study are taken as level tones on the basis of auditory judgment which is consistent with Li's transcription in the 1940s. However, as shown in Figure 3, the five-level-tone space is very crowded, especially for the mid-range tones. Although the pitch difference between Tone 22 and 33 reaches statistical significance, their difference is less than 10 Hz, just about the JND. Likewise, the difference between 33 and 44 is only 20 Hz, which is also a very small difference. These small pitch differences make us wonder whether these tones are able to contrast in the tonal space. To see how these level tones are distributed in a physical tonal space, the distribution of the tonal categories is visualized by Multi-Dimensional Scaling (MDS) [Kruskal and Wish, 1977] in a low dimensional space. Typically, distances among contrasting categories (e.g., vowels) are calculated as Euclidean distances in a low-dimensional physical space (e.g., a 2-dimensional vowel-formant space). In contrast, with MDS the physical space can be based on a large number of phonetic measures. Each measure represents one dimension, and the values of the measure are the coordinates of the dimension. Here, each token is represented by five measures: mean F0, ΔF0, F0 onset, F0 offset, and duration. The MDS function is able to calculate the distances among tokens in a high-dimensional space (here, five-dimensional), and project the distances into a lower-dimensional and interpretable space.

The MDS solution was obtained by Kruskal's Non-metric Multidimensional Scaling algorithm (*isoMDS* function in R), and physical distances among the five level tones were calculated with the

five measures mean F0, ΔF0, F0 onset, F0 offset, and duration (all scaled) as the coordinates. A 2-D solution accounting for 84% of the variance (with stress value of 0.02) is presented as Figure 4.

(Figure 4 about here)

The intuitive interpretation of the plot is that the more distant the tokens are in the space, arguably the more contrastive the tonal categories are. Although we include all pitch and duration cues in the model, mean F0 is the primary cue that is responsible for the dispersion of the tonal categories, consistent with the results from mixed-effect models. As indicated in Figure 4, the tonal space seems to be structured by two pitch-height features (i.e. high and low), and tones cluster into three pitch ranges. Dimension 1 distinguishes low tone (T11) from non-low tones, while Dimension 2 separates high tone (T55) from non-high tones; mid tones (T22, T33 and T44) cluster together in a non-high and non-low range. T22 and T33 are not distinctive in the tonal space at all, and they are only marginally contrastive with T44. This is not surprising given the pitch differences in Figure 3: recall that the difference between T22 and T33 is less than 10 Hz, and T44 and T33 have only a 20 Hz difference. If this tonal space is accurate, native listeners should not be able to hear those contrasts reliably. In general, this space suggests that mean F0s can distinguish only three levels of tonal contrasts.

4. Perceptual space of tonal contrasts

The goal of the perception experiment is to determine whether native listeners are able to hear all the tonal contrasts in Black Miao, and to examine how these tones are distributed in a perceptual space.

4.1 Methods

4.1.1 Stimuli

The stimuli were a minimal set of eight real monosyllabic words with the syllable /pa/ (/p/ is an unaspirated voiceless bilabial stop): /pa44/ "send", /pa51/ "drop", /pa55/ "(water) full", /pa22/ "net",

/pa45/ "pig", /pa33/ "fail", /pa13/ "father", and /pa11/ "drive away (duck)". These words were chosen because they are frequently used in Black Miao people's daily life, and were found to be the set that was most easily recognized and produced by native speakers during the elicitation. This choice could partially overcome the possible influence of relative lexical frequency of the eight words (in the absence of an exhaustive source for an accurate estimation of lexical frequency for this language). Potential lexical frequency bias was further eliminated by a familiarity phase in the experiment (see next section).

A male native speaker produced all of these words in isolation; one token of each word was used in this experiment. This male speaker had a good education background and used to be a Black Miao language teacher. All participants in our study were personally familiar with him, which made his voice comfortable to them. He also recorded the experimental instructions in Black Miao. In the instructions, these monosyllabic targets were explained in Black Miao and used in appropriate contexts so that the subjects would unambiguously understand these words. For example, they would hear "/pa51/, as in 'I dropped my money'" (in Black Miao). The time-normalized F0 tracks of the stimuli are shown in Figure 5 (audio files of the 5 level tones are available as supplementary material at ...). Similar to Figure 3, the speaker's pitch range for the mid-level tones is only around 30 Hz, and T33 is barely distinctive from either T22 or T44. T11 is also very close to T22. Therefore, these particular tokens produced by the speaker are representative of the community's productions as seen in Figure 3.

(Figure 5 about here)

4.1.2 Procedures

The experiment was run by a Matlab script on a laptop in a quiet room, and audio stimuli were played through SONY MDR-NC60 noise-attenuating headphones. The experiment had two phases. The first

phase was a familiarization phase, during which subjects were asked to listen to an audio introduction in which they were presented with the 8 test words and told that they would hear two of them in each trial. This was to create the same expectation for all the words and thus overcome any prior bias about the test words. The instructions could be heard as many times as needed until a listener fully understood and memorized the words that would be presented in the following test. When they were ready, they were asked to produce the eight words by themselves first, repeating each word twice. This was to make sure these words were fully accessible for them.

The second phase was an AX discrimination task. In each trial, two audio stimuli were presented (The inter-stimulus interval was 500 ms, and the trial-initial silence was 600 ms), and two possible responses, "different" and "same" in Chinese, were displayed on the screen. Subjects were asked to judge whether the sounds they had just heard were same or different words by clicking on the screen (either by themselves directly or the author did the clicking based on their verbal answers). Subjects were asked to make their best guesses when they were not sure. The stimuli were all possible pairs among the eight stimuli in a random order (64 tokens in total). Thus in the “same” trials a single stimulus was played twice. The task was repeated three times.

4.1.3 Subjects

A total of 18 listeners, eight males (the same speakers from the production experiment) and ten females, with self-report of no hearing disabilities, were recruited from Shidong village. Listeners ranged in age from 25 to 55 with an average of 34 years. Most of them had been to local elementary school, so they were able to understand Chinese and were comfortable with reading Chinese characters on the computer screen. There were four females, not native speakers of this particular Black Miao dialect, who were excluded from the current analysis, leaving 14 listeners.

4.2 Results and discussion

Table 3 is the summary dissimilarity matrix of all eight tones from the discrimination task. Dissimilarity is calculated from the percentage of "different" responses to the tone pairs across all listeners. Responses for a stimulus pair in the two possible orders (e.g., T33 vs. T55, and T55 vs. T33) are averaged. In Table 3, the dissimilarity for the “different” pairs is close to 1, while for the “same” pairs it is close to 0. Thus in general, listeners showed high accuracy rates among all tonal pairs, including the “same” pairs, indicating that listeners are able to hear the tonal distinctions in the “different” pairs without any strong bias to answer “different” to all pairs. Among the level tones, the most confusable pair is T22 vs. T44, with 70% of the responses correct for that pair. Surprisingly, T33 is not confusable with either T44 (98% correct) or T22 (95% correct), even though Figure 3 shows that they barely differ in pitch. Similarly, the other two pairs of adjacent tones, i.e. T11 vs. T22 (also very close in Figure 5) and T44 vs. T55, are also nearly perfectly discriminated (93% and 92% respectively).

(Table 3 is about here)

Another Multidimensional-Scaling (MDS) solution was found to map the confusability of the five level tones, now in a "perceptual space", with distances calculated from the perceptual dissimilarity [Shepard, 1972; Kreiman et al., 1990]. The more distinctively the listeners perceive them, the more distant the tokens are in the space. A 2-D solution which accounts for 61.6% of the variance (with stress value of 0.009) is presented in Figure 6.

(Figure 6 about here)

As seen in Figure 6, the five level tones are well distinguished in native listeners' minds. Dimension 1 divides the tonal space into a mid-range (T22, T33 and T44) vs. extremes (T11 and T55). Dimension 2 mostly follows the low-to-high pitch scale, as seen for T22 vs. T44 and T11 vs. T55, with the exception of T33. It seems that the mid-range space is further divided into two: Tone 33 occupies its own space, which is very distinct from the space occupied by T22 and T44.

Ideally, this perceptibility of tonal categories should be reflected in the corresponding production space. However, comparing Figure 6 to Figure 4, we can see that the two spaces are very different. Only a very weak correlation ($r=0.17$, $P<0.05$) is found between the production distance-matrix and perceptual distance-matrix. Therefore, the pitch-and-duration-based model fails to reflect the actual distinctiveness of the well-separated five-level-tone contrasts.

In sum, the results from the perception experiment suggest that the tonal categories are well dispersed in a perceptual space; however, the relative perceptibility of the five level tones differs. Tones with extreme pitch values are well distinguished from the mid-range tones; the mid-range space is further divided into tone T33 vs. tone T22 and T44. Even though pitch-wise T33 is very close to both T22 and T44 (Figure 6), T33 is not confusable with them; and T22 and T44, the tonal pair with the larger pitch difference, are actually the most confusable. Since there is no significant contribution of pitch contours or duration to the production space, it is very mysterious why and how native listeners are able to hear these contrasts. Thus a more sophisticated tonal space model, which incorporates phonation cues, will be tested.

5. A pitch-phonation tonal space

Failing to reflect the perceptual distinctiveness for native listeners indicates that a tonal space modeled only on pitch (+ duration) cues is not sufficient to account for the Black Miao five-level-tone

contrasts; therefore, in this section, we will evaluate the competing hypothesis: both phonation and pitch contribute to tonal dispersion.

5.1 Measurements

Comprehensive acoustic measures reflecting different phonation properties were made using VoiceSauce [Shue et al. 2011]. (1) The amplitude difference between the first and second harmonics, H1*-H2* (with formant corrections by Iseli et al. [2007]), controversially reflecting open quotient of the vocal folds (Holmberg et al., 1995; Ní Chasaide & Gobl, 1999; for questions and issues see Gerratt and Kreiman, 2001; Kreiman et al., 2007; Kreiman et al., 2012). This measure has been found to successfully distinguish contrastive phonations across languages [Gordon and Ladefoged, 2001; Keating et al., 2011, 2012]. (2) The amplitude difference between H1 and the amplitudes of the harmonics nearest to F1, F2, and F3 (H1*-A1*, H1*-A2*, H1*-A3*), indicating the strength of higher frequencies in the spectrum, which might be related to closing velocity of the vocal folds [Stevens, 1977], and which have been found to reliably distinguish between breathy vs. non-breathy phonation types [Blankenship, 2002; Esposito, 2012; DiCanio, 2009]. (3) Individual harmonic amplitudes, H1*, H2* and H4*, which have been found to be important spectral landmarks of voice perception [Kreiman and Garellek, 2011]; H2*-H4*, significantly correlated with gender [Kuang, 2011; Bishop and Keating, 2012]. (4) Cepstral Peak Prominence (CPP) [Hillenbrand et al., 1994], reflecting the harmonics-to-noise ratio and periodicity, which has been found to be an indicator of contrastive breathy phonation [Blankenship, 2002; Garellek and Keating, 2011]. The idea here is to include all possible acoustic information without bias from any specific preconception of phonation and tone.

The EGG signals that were recorded simultaneously with the audio were processed by EggWorks [Tehrani, 2012] and three measures were extracted: Contact Quotient (CQ), which is defined as the proportion of the vocal fold contact during each single vibratory cycle [Rothenberg and Mahshie,

1988]; Peak Increase in Contact (PIC), defined as the amplitude of the positive peak on the DEGG wave, corresponding to the highest rate of increase of vocal fold contact [Michaud, 2004; Keating et al., 2011]; Speed Quotient (SQ), defined as the ratio between closing duration and opening duration, reflecting the skewness of the pulses [Marasek, 1996]. CQ was calculated using the "hybrid" method [Howard, 1990]: using the positive peak of dEGG to define closing events, and a 3/7 threshold to define opening events.

5.2 Phonation cues in five level tones

5.2.1 Acoustic phonation cues

A classification regression tree (*rpart* function in R) is first run to determine which measures are the most important to the tonal contrasts, with all pitch, duration, and voice measures as the predictors. The results show that the most important acoustic cues for classifying the five level tones are (mean) F0 ($p<0.001$), H1*-H2* ($p<0.001$), H1*-A1* ($p=0.002$) and CPP ($p<0.001$). These cues together can account for 64% of the data. Two spectral measures (H1*-H2* and H1*-A1*), which are related to open quotient of the vocal folds, and one measure (CPP), which reflects noise and periodicity, are included in this set of significant cues. A series of pairwise mixed-effect models are then used to decide the tonal effects on the three voice measures, with tonal categories as the fixed factor and speaker as the random factor. For H1*-H2* and H1*-A1*, significance is found for all tonal pairs, except for T22 and T44; as for CPP, there is no significant difference among T22, T44 and T55, but these three tones are all significantly different from T11 and T33. Figure 7 (a-c) shows the values of these three measures for the five tonal categories.

The patterns of H1*-H2* and H1*-A1* are very consistent. They both show that T33 is breathier than any other tones, as it has significantly greater H1*-H2* and H1*-A1* (Figure 7a, 7b). On the other hand, T11 and T55 are more constricted/laryngealized than any other tones, as they have significantly

smaller H1*-H2* and H1*-A1*. T22 and T44 have a similar voice quality, which is in between the breathier T33 and the more laryngealized tones (i.e. 11 and 55). These two figures suggest a three-way phonation distinction among the five level tones. Figure (7c) suggests some different information about these different phonations. CPP groups tones T11 and T33 together vs. the others. As discussed before, CPP reflects the periodicity and harmonic-to-noise ratio of phonation, so this means that T11 and T33 are less periodic and/or noisier than the other tones. It is likely that lower CPP for T33 is caused by higher breathy noise in the spectrum, and lower CPP for T11 is caused by irregularity of vibration. Therefore, although T55 and T11 are both laryngealized, T55 is much more periodic than T11, suggesting that they are actually not the same type of phonation, with T11 being more creaky.

(Figure 7 about here)

5.2.2 Physiological mechanisms

We now try to understand the physiological mechanisms involved in the tonal contrasts, especially the interaction between pitch and phonation. A Principle Component Analysis (PCA) biplot is employed to classify the five level tones, and evaluate the roles of phonation and pitch in tonal contrasts. CQ (Contact Quotient), SQ (Speed Quotient), PIC (Peak Increase in Contact) and mean F0 are the variables of the model. The inner interactions among these variables are also visualized by the biplot.

Figure 8 presents the first two principal components, which together account for 97.4% of the variance in the data. The biplot indicates the interactions among the variables. There are several kinds of information that can be read from this plot.

(Figure 8 about here)

First, in this kind of plot, the length of the lines approximates the variances of the variables (direction is indicated by arrow). The longer the line, the higher the variance (i.e. the more important is the cue). Reading this figure, F0 has the highest variance among the variables in the biplot, as it has the longest arrow line; while PIC has the lowest, as it has the shortest arrow line. Intuitively, this means that the F0 difference is the most important mechanism for tonal contrasts, which is not surprising. Among the EGG parameters, SQ, the skewness of the glottal pulse, and CQ, indicating the open quotient of the vocal folds, are the major phonation mechanisms.

Second, the plot also shows the roles of phonation and pitch in these tonal contrasts. Consistent with the results of the initial analysis of the acoustic phonation cues, T11 and T55 have the greatest CQ values, and smallest SQ and PIC, suggesting that these two tones are laryngealized such that the glottal pulses have small open quotients, skewed shape, and slow contact speed. By contrast, T33 has the smallest CQ, and greatest SQ and PIC, suggesting that this tone has a breathy phonation, the glottal pulses having greater open quotient, more symmetric shape and faster contact speed. T22 and T44 have the most similar voice quality, which is intermediate.

Third, the angle between the lines approximates the correlation between the variables they represent. The closer the angle is to 90, or 270 degrees (i.e. lines are perpendicular to each other), the smaller the correlation; whereas an angle of 0 or 180 degrees (i.e. lines are overlapping or in the exactly opposite direction) reflects a strong correlation of 1 or -1, respectively. Therefore, the biplot shows strong correlations among CQ, PIC and SQ, which almost fall on a single line. Meanwhile, F0 has only a weak positive correlation with the phonation parameters, and forms a largely independent dimension by itself. The weak correlation can be explained by the well-established mechanism that the longitudinal tension of the vocal folds increases as pitch increases [Ohala, 1978; Titze, 1988; Baken and Orlikoff, 2000, among many others], which leads to increased CQ values in EGG. Indeed, T22,

T44 and T55 follow this pattern. T55 is more laryngealized than the tones with lower pitches, suggesting that T55 is produced with a tense (or stiff) phonation. However, the locations of T11 and T33 cannot be explained by this mechanism. The high CQ of T11 is likely to be due to vocal fry, caused by the compression in the vocal folds, which naturally occurs with low pitches. Thus different phonatory mechanisms can explain the different acoustic properties of T11 and T55. Finally, the distinctive breathy phonation of T33 cannot be explained by the pitch production mechanism, but instead must have to be learned by native speakers.

5.2.3 Pitch-phonation tonal space

With this understanding of phonation cues for the five tones, we now turn to the main question: how do these phonation cues contribute to the dispersion of the tonal space? Incorporating acoustic phonation cues as well as pitch (+ duration) cues, we generate a new MDS tonal space (Figure 9), which accounts for 66.2% of the variance in this larger dataset (with stress value < 0.00001). We can see significant improvements from Figure 4: First of all, T33 now is well distinguished from T22 and T44, occupying its own quadrant; second, the scale of the space is much bigger than Figure 4, which indicates a better dispersion in general. The enhancement of distinctiveness is very important given that tonal contrasts are realized in a very limited pitch range. The new production space now matches better with the perceptual space. A much stronger correlation ($r=0.87$) of the perception distance-matrix and production distance-matrix is now found. This result indicates that non-modal phonations in Black Miao are very important in production, and by inference, also in perception.

As indicated in Figure 9, similar to Figure 6, the tonal production space is divided into a mid range (T22, T33 and T44) and an extreme range (T11 and T55). The extreme pitch ranges are also related to laryngealization (Figure 8). The mid-range space is further divided into two parts, breathy tone T33 vs. modal tones T22 and T44. Tones in the different quadrants benefit from both phonation and pitch

cues, whereas tones in the same quadrants are primarily distinguished by pitch cues. The laryngealization in T11 and T55 enhances the difference between tones with extreme pitch values and tones within the mid-range. On the other hand, breathy phonation creates an additional quadrant for the distinction between T33 and the other mid tones. In view of this production space, it is no longer mysterious how listeners are able to perceive the five-level-tone contrasts in Black Miao.

(Figure 9 about here)

6. Discussion – tonal space model

In this study, we conducted both production and perception experiments with Black Miao, to explore how native speakers produce and perceive the contrasting five level tones. We confirmed that pitch is not the only cue for tonal contrasts in this language, and non-modal phonations appear to be very important cues in both tonal production and perception. T55 and T11 can benefit from both pitch cues and phonation cues so that they have very good separability from the mid tones. For the mid-range tones that have very similar pitch cues, T33 is distinctive from T22 and T44 primarily by phonation cues. T22 vs. T44, the tonal contrast with only a pitch difference, is the hardest to produce and perceive distinctively.

The interaction between pitch and phonation leads to the well-dispersed tonal spaces shown in Figure 6 and Figure 9. The two tonal spaces share a similar dispersion pattern: Pitch range is first divided into three ranges, i.e. high, mid and low, where the high and low ranges are also cued by laryngealization. The mid-range space is further divided into breathy vs. modal parts, so that tone 33, the least dispersed tone in a pitch-based tonal space (Figure 4), is well distinguished in the new pitch-phonation tonal space. Figure 10 generalizes the organization of the tonal spaces from Figure 6 and Figure 9 (collapsing into a different view).

(Figure 10 about here)

In this schema, the five level tones are divided into different quadrants based on different phonations, as in Figure 9. The tones with the highest pitch and the lowest pitch form their own quadrants, and the tones with mid-range pitches can be further divided into two quadrants: T33 in the breathy quadrant, and T22 and T44 in the modal quadrant. With these dimensionalities, the burden of pure pitch contrasts reduces to T22 vs. T44 only.

Comparing Figure 9 with Figure 4, the non-modal phonations contribute to the improvement of tonal distinctiveness in two ways: On one hand, the phonation cues enhance the contrasts for T11 and T55, so that the general distinctiveness of the tone space is enlarged; on the other hand, the breathy phonation creates an independent dimension for T33, so that T33 is very distinct from the other mid-range tones, i.e. T22 and T44.

These two functions reflect the different possible relationships between pitch and non-modal phonations. The first kind of non-modal phonations are parts of the pitch scale, such as vocal fry, falsetto and tense voice. Vocal fry is correlated with the lowest pitch range, and falsetto or tense voice is usually associated with the highest pitch range. Referring to Figure 3, the mean F0 of the highest tone is around 220 Hz, which is a remarkably high pitch for male speakers, much higher than the average 175 Hz upper limit of the male speech range across languages [Baken and Orlikoff, 2000]. If not doing anything to reduce the longitudinal tension in the vocal folds (e.g. switch to falsetto), then these high pitches are produced with a tense voice [Kong, 2007]. This tension results in a greater CQ in EGG signals. Likewise, when pitch goes to the lowest end, e.g., below 75 Hz for males, speakers have to produce these pitches with creaky voice (e.g., vocal fry), which also leads to a greater CQ.

Unlike these pitch-driven non-modal phonations, the second type of non-modal phonation, such as breathy, is relatively independent from pitch[1]. The contrasts between breathy and modal (or essentially the degree of breathiness) are relative along the continuum of the glottal strictures from the most open (voiceless) to the most closed (glottal stop) [Ladefoged, 1971]. This type of non-modal phonation can create an independent dimension for tonal contrasts, so that tones with similar pitches (T33 vs. T22 and T44) but in different phonation registers are rarely confused.

Why do tonal languages need these two kinds of non-modal phonations? We can account for it from the view of optimizing the dispersion of tonal inventories. When level tone inventories are large, pitch cues are no longer sufficient, requiring too much perceptual or articulatory effort to maintain the crowded contrasts. As discussed earlier, there are two possible ways to optimize the tonal spaces with large size of inventories: expand the pitch space for tonal contrasts or add an additional contrastive dimension. Indeed, the pitch-driven phonations can help to produce extreme pitch targets, either super high or low, and thus enhance the physical and perceptual pitch distinctiveness for the highest and lowest tones. On the other hand, pitch-independent phonations create an independent dimension for tonal contrasts so that tones with similar pitches can be distinguished from each other. In sum, the well-distinguished five-level-tones of Black Miao can be attributed to both kinds of non-modal phonations. Lindblom and Maddieson [1988] proposed that small inventories can be distinguished on just the "basic" dimensions, while larger inventories have to expand to more complicated dimensions. This principle was proposed on the basis of a study of consonant inventories, and the present study shows that the same principle also holds for tones. Pitch cues are sufficient to distinguish small tonal inventories, but larger tonal inventories require more complicated dimensions.

[1] Breathy phonation tends to lower pitch across languages [Gordon and Ladefoged, 2001], but that usually happens throughout the pitch range. In addition, the perception of breathy tones is not affected by pitch (shown in Figure 6, also Garellek et al. [2012] for White Hmong, Andruski [2006] for Green Mong).

The tonal space model proposed in this paper can explain the typologically different use of non-modal phonations across languages. As pitch-driven non-modal phonations are related to realization of extreme pitch targets, they are usually found in extra-low tones or extra-high tones. Vocal fry in low tones is very common in languages, some famous cases being the Mandarin 213 tone and the Cantonese 11 tone [Belotel-Grenié and Grenié, 1994; Yu and Lam, 2011]. Perception experiments [Yu and Lam, 2011; Yang, 2011] have shown that this non-modal phonation can facilitate tonal recognition for these low tones. Non-modal phonation in super high tones is less documented, but a few languages that have multiple level tones, such as Yueyang Dialect [Peng and Zhu, 2010] and Pakphanang Thai [Rose, 1997], have been reported to have falsetto voice correlated with the highest tones. In all these cases, non-modal phonations are allophonic to tonal contrasts, as they are enhancement cues to pitch. By contrast, pitch-independent non-modal phonations usually are a phonemic dimension in languages. For example, tonal contrasts and phonation contrasts are crossed in the Chinese Wu dialect [Cao and Maddieson, 1992], Southern Yi [Kuang, 2011] and Mazatec [Garellek and Keating, 2011].

One last point: pitch contrast appears to be constrained by phonation types, that is, the possible pitch contrasts vary among phonation conditions. We see that pure modal phonation is able to host two levels, but non-modal phonations only host one level. This is consistent with the cross-linguistic facts. For example, in languages with multiple tones, such as White Hmong and Zapotec, most pitch contrasts occur with modal voice [Esposito, 2010, 2012]; fewer tones occur with non-modal phonations. Perceptually, Silverman [2003] showed that pitch discrimination by English speakers is less good during breathy phonation than during modal phonation. In addition, breathiness contrasts might more comfortably take place in mid-range tones, as the vocal folds are more adjustable within this range [Baken and Orlikoff, 2000].

The idea that phonation is part of the total phonetic space for tones is not new, and is seen especially in proposals about tonal registers [Yip, 2002; Duanmu, 1990; Bao, 1999]. Recently, Zhu [2012] has proposed a system of three registers, with different phonations each characterizing one or two of the registers. Each register also has four pitch levels drawn from a set of six total available pitch levels, such that phonations are associated with lower or higher parts of the pitch range. For example, falsetto phonation is associated with higher pitch and defines the High register, which encompasses pitch levels 3, 4, 5, 6, while creaky phonation can be used in both the Low register (pitch levels 1, 2, 3, 4) and the Mid register (pitch levels 2, 3, 4, 5). This system generates a large number of possible level tones, and level-tone inventories: any one language can have level tones in one, two, or all three registers, and up to four levels in each register. Like Zhu's proposal, the proposal in the present paper incorporates phonation into tonal descriptions, and refers to natural relations between certain phonations and pitch levels. Most notably, in both systems, pitch levels higher than 4 are not compatible with modal phonation.

Nonetheless, the two proposals differ in several important respects. The model presented here posits only the five traditional phonetic pitch levels, not six, and it limits pure-pitch contrasts (i.e. those with the same phonation) to two, or at most three, levels, certainly not four. This is because pitch contrasts are easiest to produce, and to perceive, in modal voice, i.e. within the 100-Hz pitch range of that voice quality, but tones are constrained by the auditory system to be at least 20-30 Hz apart. Any additional pitch level contrasts require associated phonation differences. Moreover, the model distinguishes pitch-driven phonations (creaky voice at pitch level 1, falsetto or tense voice at pitch level 5) from pitch-independent phonations (i.e. relative breathiness contrasts), based on the physiological interaction between pitch and phonation. In this way, this model not only accounts for the universal linguistic functions of non-modal phonations (e.g. enhance pitch targets and add contrastive dimensions), but also allows language-specific variability in which pitches combine with which

phonation types. For example, to distinguish tones with similar phonetic pitch, whether the contrast is between lax and modal (e.g. Southern Yi), or between breathy and modal (e.g. White Hmong), does not matter. All of these claims are supported in the present paper by substantial quantitative data on tone production and perception.

7. Conclusions

This study investigates the dispersion of multi-level tonal contrasts by exploring the cues used in producing and perceiving the five level tones of Black Miao. Both production and perception experiments show that five level tones are well dispersed when the language takes advantage of non-modal phonation cues. A new tonal space model based on the interaction between pitch and phonation is proposed, in which two different kinds of non-modal phonations - that either enhance pitch contrasts or provide an additional contrastive cue - divide tonal levels into several registers so as to optimize the distinctiveness of the tonal space.

The model makes several interesting predictions: Limited to a comfortable pitch range and reliable perceptual differences, pure pitch-level contrasts are only possible with two levels, perhaps at most three; moreover, such two-level-tone systems occupy the simplest tonal space, which only utilizes the most basic dimension (i.e. pitch levels) for contrasts. Additional tonal contrasts have to occupy more elaborated tonal spaces which recruit additional dimensions, such as duration, contours and phonation cues. These are strong predictions which seem to be contradicted by many known tone systems and thus will need to be extensively tested cross-linguistically in the future.

The findings of this paper also have great implications for future tonal studies. As tonal contrasts are more than pitch, future tonal studies should take multidimensional phonetic cues such as phonation and duration into account. Furthermore, the new method developed in this paper is a very powerful

tool for understanding the roles of multiple cues in production and perception; particularly, as it does not rely on speech synthesis or re-synthesis, this method is highly suited to fieldwork studies.

Acknowledgements

This work was supported by NSF grant BCS - 0720304 to Patricia Keating and a summer research award from the UCLA Department of Linguistics. I would like to thank Professor Keating for her valuable comments; Yen-Liang Shue for VoiceSauce and Henry Tehrani for EggWorks. I also would like to thank Weihan Zheng for his great assistance with contacting the Black Miao group in Guizhou province; Shilong Liu for his time and effort as my main language consultant, and for assisting with explaining the experiments to other participants; Professor Jiangping Kong for his assistance with equipment; Professor Defu Shi for sharing his knowledge of Miao languages. An earlier draft of this paper was presented at the Third International Symposium on Tonal Aspects of Languages (TAL2012) in Nanjing. Thanks are due to two anonymous reviewers, and to Wentao Gu as well as the Editor of Phonetica Klaus J. Kohler for helpful advice.

References

Alexander, J.A.: The theory of adaptive dispersion and acoustic-phonetic properties of cross-language lexical-tone systems. (Dissertation, Northwestern University, Evanston 2010).

Andruski, J.E.: Tone clarity in mixed pitch/phonation-type tones. Journal of Phonetics *34 (3)*: 388-404 (2006).

Baken, R.J.; Orlikoff, R.F.: Clinical measurement of speech and voice. (Singular Publishing Group, San Diego 2000).

Bao, Z.: The structure of tone. (Oxford University Press, New York 1999).

Barry, J.G.; Blamey, P. J.: The acoustic analysis of tone differentiation as a means for assessing tone production in speakers of Cantonese. Journal of the Acoustical Society of America *116*: 1739-1748 (2004).

Becker-Kristal, R.: Acoustic typology of vowel inventories and Dispersion Theory: Insights from a

large cross-linguistic corpus. (Dissertation, University of California, Los Angeles 2010).

Belotel-Grenié, A.; Grenié, M.: Phonation types analysis in standard Chinese. Proceedings of International Conference on Spoken Language Processing *3*: 343-346 (1994).

Bishop, J.; Keating, P.: Perception of pitch location within a speaker's range: fundamental frequency, voice quality and speaker sex. Journal of the Acoustical Society of America *132(2)*: 1100-1112 (2012).

Blankenship, B.: The time course of nonmodal phonation in vowels. Journal of Phonetics *30 (2)*: 163-191 (2002).

Blicher, D. L.; Diehl, R. L.; Cohen, L.B.: Effects of syllable duration on the perception of the Mandarin Tone 2/Tone 3 distinction: Evidence of auditory enhancement. Journal of Phonetics *18(1)*: 37-49 (1990).

Boersma, P.; Weenink, D.: Praat, http://www.fon.hum.uva.nl/praat, accessed on 10th November 2012.

Brunelle, M.: Tone perception in Northern and Southern Vietnamese. Journal of Phonetics *37 (1)*: 79-96 (2009).

Cao, J.; Maddieson, I.: An exploration of phonation types in Wu dialects of Chinese. Journal of Phonetics *20*: 77-92 (1992).

Chang, K.: Tones of the Miao and Yao languages (Chinese). Shiyusuo Jikan (Bulletin of the Institute of History and Philology, Academia Sinica) *16*: 93-110 (1947).

Chao, Y.R.: Mandarin primer. (Harvard University Press, Cambridge, MA 1948).

DiCanio, C.T.: The phonetics of register in Takhian Thong Chong. Journal of the International Phonetic Association *39 (2)*: 162-188 (2009).

Duanmu, S.: A formal study of syllable, tone, stress, and domain in Chinese languages. (Dissertation, Massachusetts Institute of Technology, Cambridge, MA 1990).

Edmondson, J. A.; Gregerson, K. J.: On five-level tone systems. Language in Context: Essays for Robert E. Longacre. Summer Institute of Linguistics and the University of Texas at Arlington Publications in Linguistics *107*: 555-576 (1992).

Esposito, C.: Variation in contrastive phonation in Santa Ana del Valle Zapotec. Journal of the International Phonetic Association *40 (2)*: 181-198 (2010).

Esposito, C.: An acoustic and electroglottographic study of White Hmong tone and phonation. Journal of Phonetics *40*: 466-476 (2012).

Flemming, E.: Contrast and perceptual distinctiveness. In Hayes, B.; Kirchner R.; Steriade, D. (eds.): The Phonetic Bases of Markedness, pp. 232-276, (Cambridge University Press, Cambridge, UK 2004).

Garellek, M.; Keating, P.: The acoustic consequences of phonation and tone interactions in Mazatec. Journal of the International Phonetic Association *41(2)*: 185-205 (2011).

Garellek, M.; Esposito, C.; Keating, P.; Kreiman, J.: Perception of spectral slopes and White Hmong tone identification. UCLA Working Papers in Phonetics *110*: 24-45 (2012).

Gerratt, B. R.; Kreiman, J.: Toward a taxonomy of nonmodal phonation. Journal of Phonetics *29 (4)*: 365-381 (2001).

Gordon, M.; Ladefoged, P.: Phonation types: a cross-linguistic overview. Journal of Phonetics *29(4)*: 383-406 (2001).

Guo, Q. J.; Strauss, H.; Liu, C. Q.; Zhao, Y. L.; Pi, D. H.; Fu, P. Q.; Zhu, L. J.; Yang, R. D.: Carbon and Oxygen Isotopic Composition of Lower to Middle Cambrian Sediments at Taijiang, Guizhou Province, China. Geological Magazine *142(6)*: 723-733 (2005).

Harris, W. J.; Umeda, N.: Difference limens for fundamental frequency contours in sentences. Journal of the Acoustical Society of America *81*: 1139-1145 (1987).

Hart, J.: Differential sensitivity to pitch distance, particularly in speech. Journal of the Acoustical Society of America *69*: 811–821 (1981).

Hillenbrand, J.M.; Cleveland, R.A.; Erickson, R.L.: Acoustic correlates of breathy vocal quality. Journal of Speech and Hearing Research *37*: 769-778 (1994).

Hollien, H.; Michel, J.F.: Vocal fry as a phonational register. Journal of Speech and Hearing Research *11(3)*: 600 (1968).

Hollien, H.: On Vocal registers. Journal of Phonetics *2*: 125-143 (1974).

Holmberg, E.B.; Hillman, R. E.; Perkell, J. S.; Guiod, P.C.; Goldman, S.L.: Comparisons among aerodynamic, electroglottographic, and acoustic spectral measures of female voice. Journal of Speech and Hearing Research *38*: 1212-1223 (1995).

Howard D. M.; Lindsey, G. A.; Allen, B.: Toward the quantification of vocal efficiency. Journal of Voice *4*: 205-212 (1990).

Iseli, M.; Shue, Y.L.; Alwan, A.: Age, Sex, and Vowel Dependencies of Acoustic Measures Related to the Voice Source. Journal of the Acoustical Society of America *121*: 2283 (2007).

Kawahara, H.; Masuda-Katsuse, I.; de Cheveigne, A.: Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency based F0 extraction. Speech Communication *27*:187-207 (1999).

Keating, P.; Esposito, C.; Garellek, M.; Khan, S.; Kuang, J.: Phonation contrasts across languages. Proceedings of ICPhS *XVII*: 1046-1049 (2011).

Keating, P.; Kuang, J.; Esposito, C.; Garellek, M.; Khan, S.: Multi-dimensional phonetic space for phonation contrasts. LabPhon13 in Stuttgart, Germany (2012).

Keating, P.; Kuo, G.: Comparison of speaking fundamental frequency in English and Mandarin. Journal of the Acoustical Society of America *132(2)*: 1050-1060 (2012).

Kollmeier, B.; Brand, T.; Meyer, B.: Perception of speech and sound. In Benesty, J.; Sondhi, M.; Huang, Y. (eds.): Springer handbook of speech processing, pp. 61-82 (Springer, Berlin 2008).

Kong, J.: Laryngeal dynamics and physiological model, pp. 66-95 (Publishing House of Peking University, Beijing 2007).

Kreiman, J.; Gerratt, B.R.; Precoda, K.: Listener experience and perception of voice quality. Journal of Speech and Hearing Research *33*: 103-115 (1990).

Kreiman, J.; Gerratt, B.R.; Antoñanzas-Barroso, N.: Measures of the glottal source spectrum. Journal of Speech, Language and Hearing Research *50*: 595-610 (2007).

Kreiman, J.; Garellek, M.: Perceptual importance of the voice source spectrum from H2 to 2 kHz. Journal of the Acoustical Society of America *130(4)*: 2570 (2011).

Kreiman, J.; Shue, Y-L.; Chen, G.; Iseli, M.; Gerratt, B.R.; Neubauer, J.; Alwan, A.: Relationships among voice quality, harmonic amplitudes, open quotient, and glottal area waveform shape in sustained phonation. Journal of the Acoustical Society of America *132 (4)*: 2625-2632 (2012).

Kruskal, J. B.; Wish. M.: Multidimensional Scaling. (Sage Publications, Beverly Hills 1977).

Kuang, J.: Production and perception of phonation contrasts in Yi. (MA thesis, University of California, Los Angeles 2011).

Kwan, J.C.: A phonology of a Black Miao dialect. (MA thesis, University of Washington, Seattle 1966).

Ladefoged, P.: Preliminaries to linguistic phonetics. (University of Chicago, Chicago 1971).

Lindblom, B.: Phonetic universals in vowel systems. In Ohala, J.J.; Jaeger, J.J. (eds.): Experimental phonology, pp. 13-42 (Academic Press, Orlando 1986).

Lindblom, B.: Explaining phonetic variation: A sketch of the H&H theory. In Hardcastle, W.; Marchal, A. (eds.): Speech production and speech modeling, pp. 403-439 (Kluwer, Dordrecht 1990).

Lindblom, B.; Maddieson, I.: Phonetic universals in consonant systems. In Hyman, L.M.; Li, C.N. (eds.): Language, speech and mind, Studies in honor of Victoria A. Fromkin, pp. 62-78 (Routledge, London 1988).

Maddieson, I.: Universals of tone. In Greenberg, J.H.; Ferguson, C.; Moravcsik, E.A. (eds.): Universals of human language, pp. 335-365 (Stanford University Press, Palo Alto 1978).

Marasek, K.: Glottal correlates of the word stress and the tense/lax opposition in German. Proceedings of ICSLP *96*: 1573-1576 (1996).

Martinet, A: Function, structure, and sound change. Word *8(1)*: 1-32 (1952).

Martinet, A: Economie des Changements Phonétiques. (Francke, Berne 1955).

Michaud, A.: A measurement from electroglottography: DECPA, and its application in prosody. Proceedings of Speech Prosody, 633-636 (2004).

Mok, P.; Wong, P.: Perception of the merging tones in Hong Kong Cantonese: Preliminary data on monosyllables. Proceedings of Speech Prosody, 1-4 (2010).

Ní Chasaide, A.; Gobl, C.: Voice source variation. In Hardcastle, W.J.; Laver, J. (eds.): The handbook of phonetic sciences, pp. 427-461 (Blackwell, Oxford, UK 1997).

Ohala, J. J.: The production of tone. In Fromkin, V. A. (eds.): Tone: A linguistic survey, pp. 5-39 (Academic Press, New York 1978).

Peng, J.G.; Zhu, X.N.: Falsetto in Yueyang dialect (Chinese). Journal of Contemporary Linguistics *001*: 24-32 (2010).

Pouplier, M.: Units of phonological encoding: Empirical evidence. (Dissertation, Yale University, Newhaven 2003).

Rose, P.: A seven-tone dialect in Southern Thai with super-High: Pakphanang tonal acoustics and physiological inferences. In Abramson, A.S. (eds.): Southeast Asian linguistic studies in honour of Vichin Panupong, pp. 191-208 (Chulalongkorn University Press, Bangkok, Thailand 1997).

Rose, P.: Oujiang Wu tones and acoustic reconstruction. In Bowern, C.; Evans, B.; Miceli, L. (eds.): Morphology and Language History, pp. 235-250 (John Benjamins, Amsterdam 2008).

Rothenberg, M.; Mahshie, J. J.: Monitoring vocal fold abduction through vocal fold contact area. Journal of Speech and Hearing Research *31*: 338-351 (1988).

Shepard, R. N.: Psychological representation of speech sounds. In David, E. E.; Denes, P. B. (eds.): Human communication: A unified view, pp. 67-113 (McGraw-Hill, New York 1972).

Shue, Y.L.; Keating, P.; Vicenik, C.; Yu, K.M.: VoiceSauce: A program for voice analysis. Proceedings of ICPhS *XVII*: 1846-1849 (2011).

Silverman, D.: Pitch discrimination during breathy versus modal phonation. In Local, J.; Ogden, R.; Temple, R. (eds.): Papers in laboratory phonology VI, pp. 293-304 (Cambridge University Press, Cambridge, UK 2003).

Stevens, K. N.: Physics of laryngeal behavior and larynx modes. Phonetica *34*: 264-279 (1977).

Stevens, K. N.: Acoustic phonetics. (The MIT Press, Cambridge, MA 1998).

Tehrani, H.: EGGworks, http://www.linguistics.ucla.edu/faciliti/sales/software.htm, accessed on 10th November 2012.

Tseng, C.Y.: An acoustic phonetic study on tones in Mandarin Chinese. (Dissertation, Brown University, Providence 1981).

Titze, I.R.: A framework for the study of vocal registers. Journal of Voice *2 (3)*: 183-194 (1988).

Yu, K.M.: Learning tones from the speech signal. (Dissertation, University of California, Los Angeles 2011).

Yu, K.M.; Lam, H.W.: The role of creaky voice in Cantonese tonal perception. Proceedings of ICPhS *XVII*: 2240-2243 (2011).

Yang, R.X.: The Phonation factor in the categorical perception of Mandarin tones. Proceedings of ICPhS *XVII*: 2204-2207 (2011).

Yip, M.J.W.: Tone. (Cambridge University Press, Cambridge, UK 2002).

Zhang, S.Y.; Kreiman, J.; Gerratt, B.R.; Garellek, M.: Acoustic and perceptual effects of changes in body layer stiffness in symmetric and asymmetric vocal fold models. Journal of the Acoustical Society of America *133*: 453-462 (2013).

Zhao, Y.; Jurafksy, D.: The effect of lexical frequency on tone production. Proceedings of ICPhS *XVI*: 477-479 (2007).

Zhao, Y.; Jurafksy, D.: The effect of lexical frequency and Lombard reflex on tone hyper-articulation. Journal of Phonetics *37 (2)*: 231-247 (2009).

Zhu, X.N.: Multi registers and four levels: A new tonal model. Journal of Chinese Linguistics *40 (1)*: 1-17 (2012).

Table 1. Pitch intervals between tones in different languages [Maddieson, 1978, p. 339], combining various sources, averaged across gender.

	Two levels		**Three Levels**			**Four levels**
	Siswati	**Kiowa**	**Yoruba**	**Thai**	**Taiwanese**	**Toura**
						+50
			+52	+28	+32	+30
	+18	+22	+27	+16	+18	+10
Lowest tone	+0	+0	+0	+0	+0	+0

Table 2. Black Miao tonal system.

I	II	III	IV	V	VI	VII	VIII
44	51	55	22	45	33	13	11

Table 3. Dissimilarity matrix for all listeners. (1.00= perfectly discriminated; 0.00=not at all; therefore, we expect 0 for the same pairs, and 1 for the different pairs)

	T11	T13	T22	T33	T44	T45	T51	T55
T11	0.05							
T13	0.94	0.00						
T22	**0.93**	0.88	0.03					
T33	**0.97**	0.78	**0.95**	0.05				
T44	**0.98**	1.00	**0.70**	**0.98**	0.03			
T45	0.94	1.00	1.00	1.00	1.00	0.00		
T51	0.94	1.00	1.00	1.00	1.00	1.00	0.00	
T55	**0.95**	1.00	**1.00**	**1.00**	**0.92**	0.88	0.88	0.00

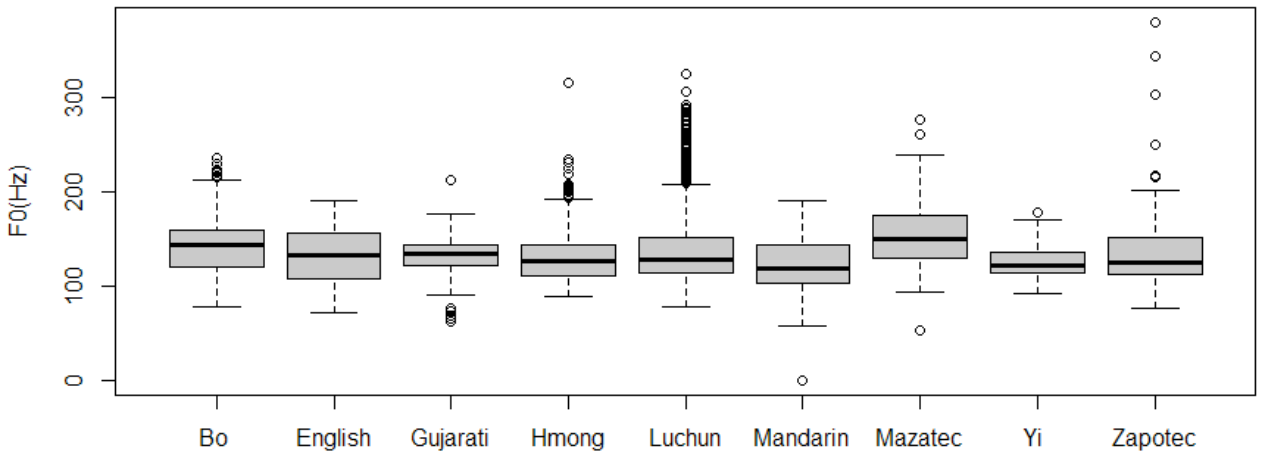

Fig.1. Speech pitch range of male speakers across languages. The measure "strF0_mean" for all tokens in the corpus is plotted here. For each language, the plot indicates: the median (the horizontal line in the box), the highest 25% of the datapoints (the upper whisker), the lowest 25% of the datapoints (the lower whisker), and 50% of the datapoints (within the box between the upper and lower quartiles); outlier datapoints are shown as circles.

Fig. 2. Map showing the location of Taijiang county of Guizhou province, reproduced from the geological study of Guo et al. [2005]. The triangle indicates the location of Taijiang.

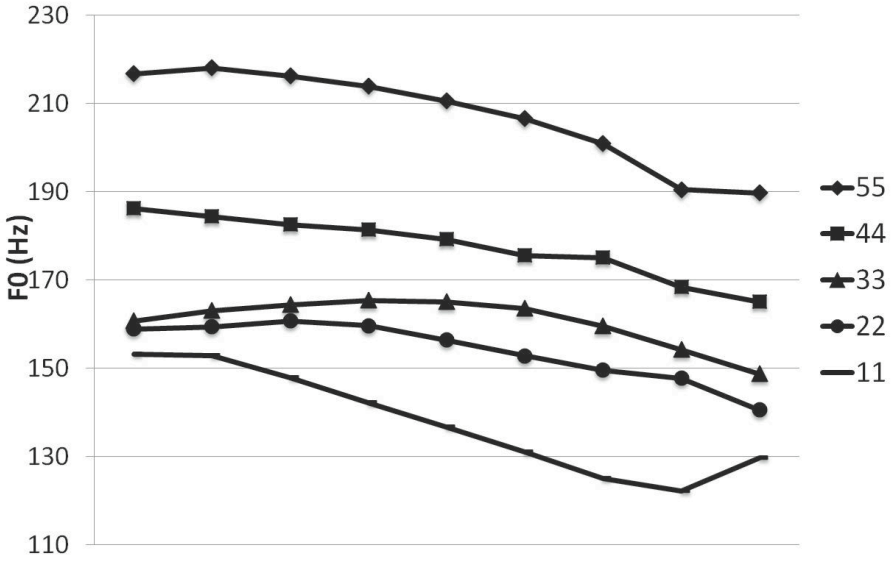

Fig. 3. Average F0 trajectories of five level tones for eight male speakers (time normalized).

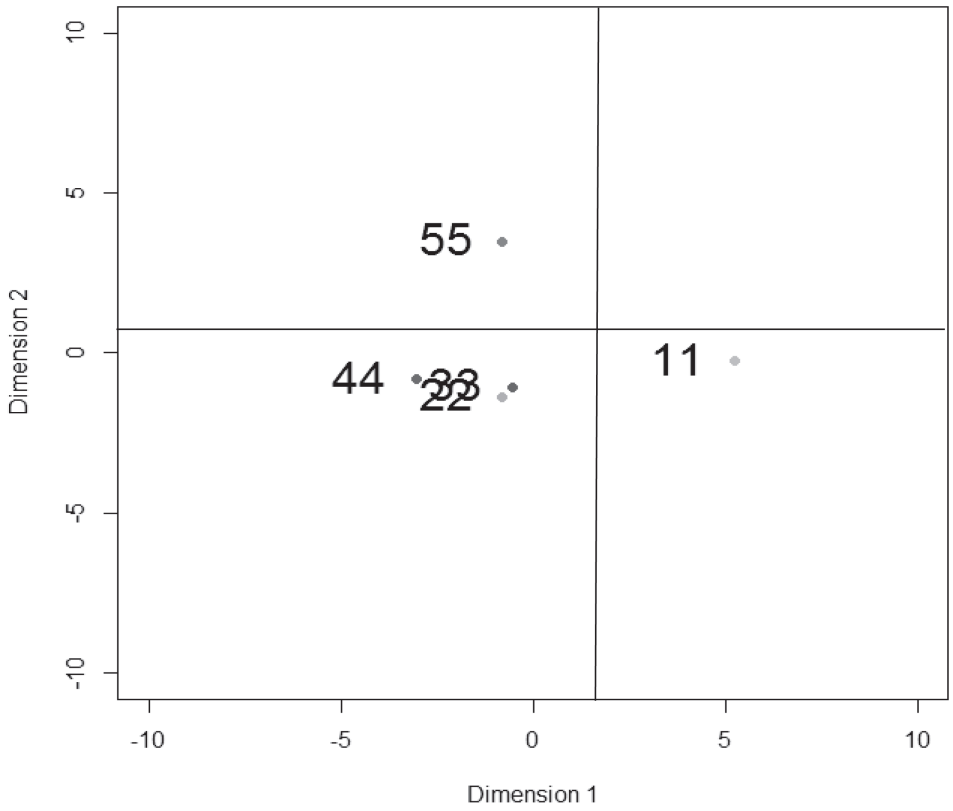

Fig. 4. Tonal space derived by MDS from pitch and duration measures, level tones only. This is a physical space showing acoustic differences. The solid lines that divide the space are added for visual convenience and are not part of the MDS solution

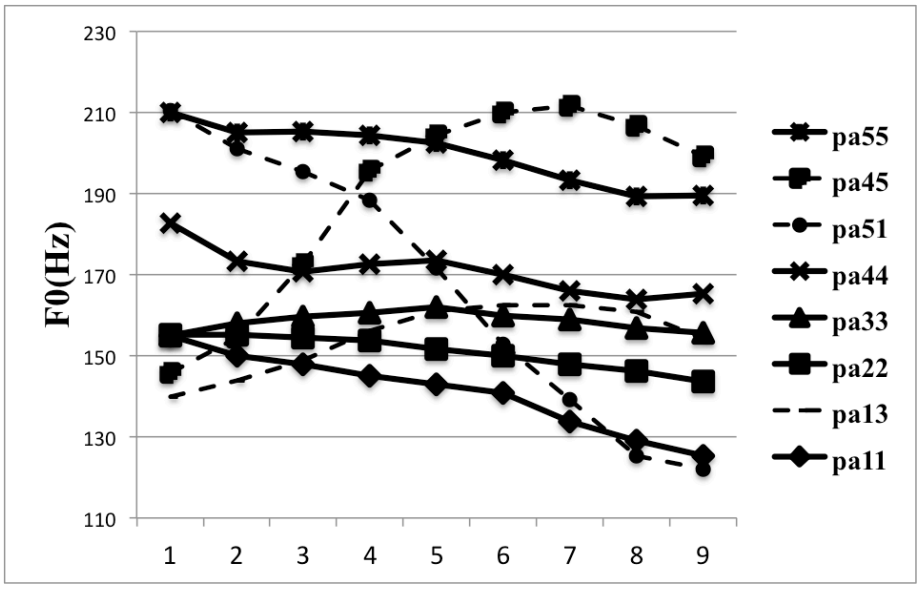

Fig. 5. F0 trajectories of all the 8 lexical tones of Black Miao on the syllable [pa], produced by a male speaker for the perception experiment (audio files of the 5 level tones are available as supplementary material at ...)

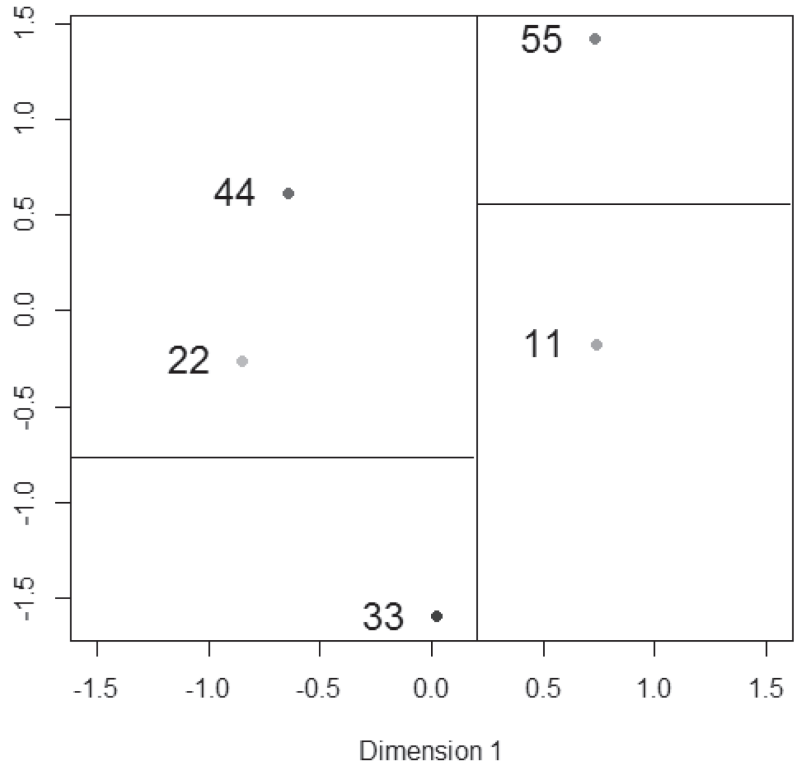

Fig. 6. Perceptual space of Black Miao five level tones derived by MDS from discrimination responses. The solid lines that divide the space are added for visual convenience and are not part of the MDS solution.

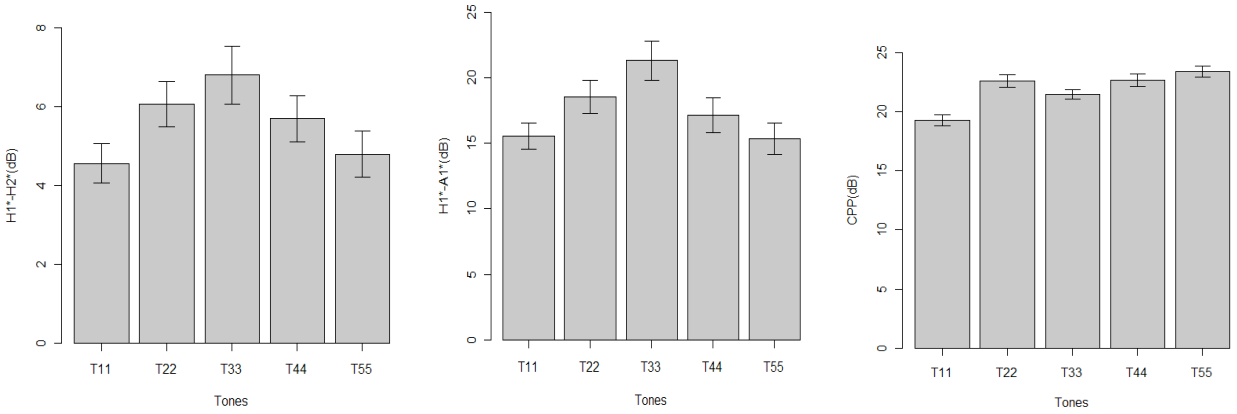

Fig. 7. Acoustic measures related to phonation contrast, by tone. (7a: H1*-H2*; 7b: H1*-A1*; 7c: CPP)

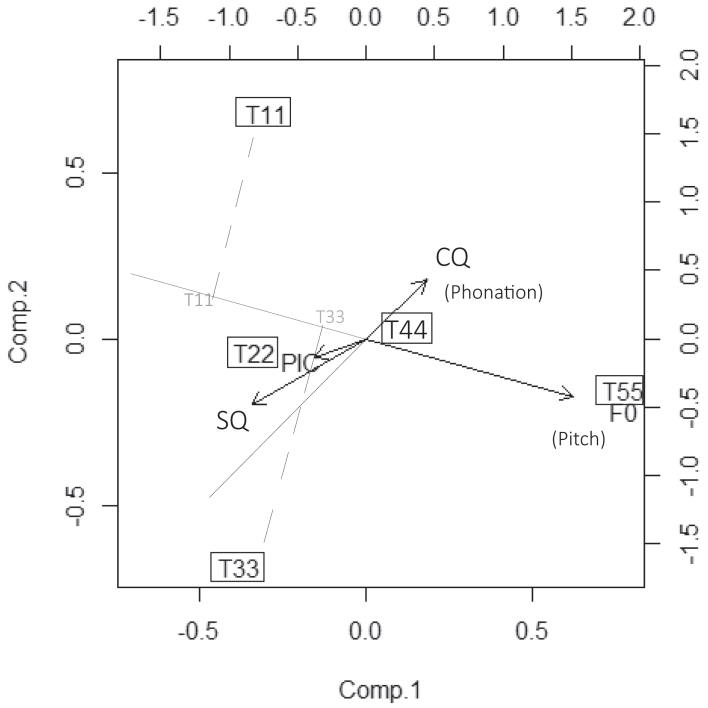

Fig. 8. PCA biplot from factor analysis of the interaction between pitch and laryngeal parameters. Length of lines=strength of this parameter, arrow=direction, angle between the lines=correlation. Tonal categories' positions are determined by the interaction of the parameters. E.g., T33 pitch-wise is located between 22 and 44 (ref. the projection of T33 on the pitch dimension), and phonation-wise is the breathiest among the tonal categories.

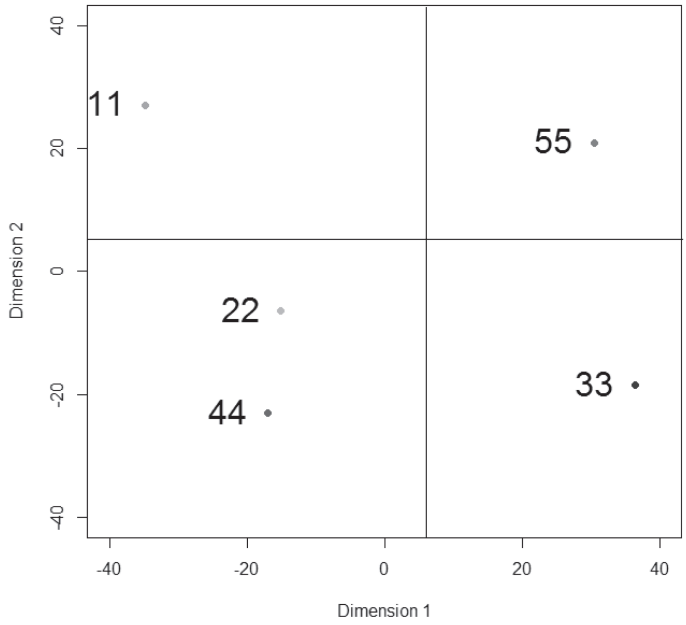

Fig. 9. MDS tonal space with pitch, duration and phonation measures, level tones only. Note the scale of Figure 9 is much larger than that of Figure 4. The solid lines that divide the space are added for visual convenience and are not part of the MDS solution.

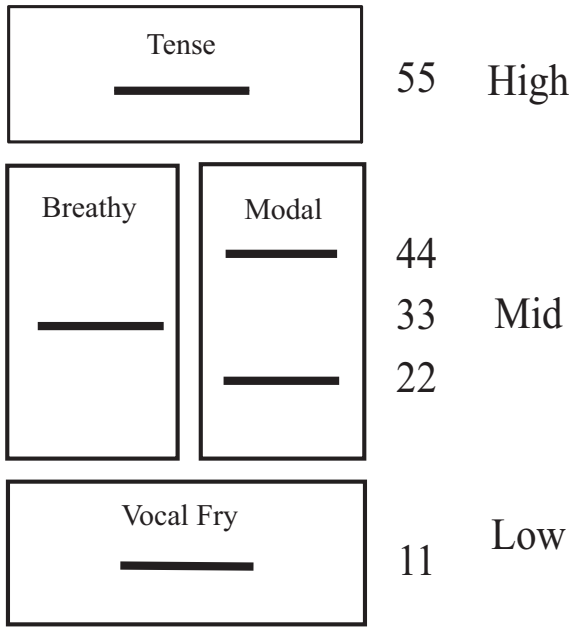

Fig. 10. Phonation registers of the five contrasting levels: a model of Black Miao tones