

## Voice Quality Variation with Tone and Focus in Mandarin

Xiaoju Zheng

Linguistics Program  
Purdue University, West Lafayette  
zheng2@purdue.edu

### Abstract

Mandarin Chinese is one of the well studied tone languages, and related contextual tonal variations as a function of prosody have attracted many researchers' interest. Change of voice quality, which is not phonemic in Mandarin, is also subject to prosody. The focus of this study will be the effects of prosodic structure, focus in particular, on voice quality in Mandarin Chinese, and the results show that there are consistent effects of prominence and T3 on voice quality change. There are also great individual variations in terms of rates and kinds of non-modal voice quality utilized.

### 1. Introduction

Using the 5-level scale proposed by [2], the four Mandarin tones are High tone (55), Rising tone (35), Low tone (21), and Falling (51). The underlying pitch targets are only partially reflected in the surface F0 contours due to articulatory constraints, and non-final focus would largely expand tonal contours and the longer duration of a syllable under focus will give articulators enough time to fully realize the underlying pitch targets [17, 18]. Prosodic locations, such as accented position, domain-initial and domain-final position, also have an effect on the variation in voice quality, resulting in domain-initial glottalization or sentence-final creak [3, 4, 5].

The purpose of this study is to evaluate the interactive effect of focus and tone on voice quality in Mandarin Chinese.

1. Is there an effect of tones on voice quality?
2. Is there an effect of focus on voice quality?
3. Is there an effect of the combination and interaction of tones and focus on voice quality?

### 2. Voice Quality and Prosody

#### 2.1 Previous Studies on Voice Quality Change

Recent studies have shown that the prosodic structure of spoken English is manifested in the speech signal by fine-grained phonetic details, and physical realizations of sounds can never be characterized adequately without the influence of prosodic structure on them being taken into account. For example, segmental variation in speech is partially attributable to the prosodic location of the segments [10, 11, 13]. Important prosodic positions for voice quality change are identified as domain initial, accented, and domain final position: glottalization rate is higher for word-initial vowel at the beginnings of intonational phrases [10, 11, 12, 13].

Pitch accent also has a significant effect on glottalization. For example, prominent words have increased levels of allophonic word-initial vowel glottalization [3, 11]; prominent words have greater spectral intensity and more high frequency spectral energy [1, 4].

Study of voice quality variation as an effect of prosody

in tonal language like Mandarin, which doesn't use voice quality contrastively, would extend previous studies on voice quality variation in English, and further evaluate the reliability and extent of voice quality change as a function of segmental and prosodic context.

#### 2.2 On Voice Quality

Voice quality is a term that subsumes a wide range of possible meanings, covering supralaryngeal and laryngeal aspects[9], and voice quality are described from perspectives of perception, acoustic signal, and physiological process, which are inter-related but not necessarily in a one-to-one relationship.

From the physiological perspective, voice has been defined with reference to activity at the level of the larynx, aspects of vocal tract excitation associated with the control, vibration of the vocal folds, and associated laryngeal structures. Modal phonation has been defined as phonation in which full contact occurs between the vocal folds during the closed phase of a phonatory cycle [15], and the vocal folds are closed during half of each glottal cycle and open during the other half (approximately), which results in about 50% open quotient [8]. Glottalization is defined by [15] as "transient sounds resulting from the relatively forceful adduction or abduction" of the vocal folds, and creaky voice is caused by holding vocal folds loosely together with air bubbling up through them, which results in longer closed phase, hence smaller open quotient than that of modal voice, both of which refer to a tendency to constriction at the laryngeal level. Creaky phonation, a different term from creak, is characterized by irregularly spaced pitch periods and decreased acoustic intensity relative to modal phonation [5]. Whereas, breathy voice is characterized as hyper-abduction of the vocal folds resulting in less vocal folds contact or greater open quotient than the modal and creaky phonation [8], and the acoustic manifestation is substantial turbulent energy which makes it difficult to discern individual pitch pulses. [5] suggested a conceptual continuum of phonation types, "defined in terms of the arytenoids cartilages, ranging from voiceless, through breathy voiced, to regular, modal voicing, and then on through creaky voice to glottal closure", which correspond to gradient degrees of openness between arytenoids cartilages.

Acoustically, vocal registers are characterized as falsetto, modal, and vocal fry, creak, or pulse register, respectively corresponding to highest, normal, lowest range of the human vocal frequencies [7]. According to [7], vocal fry register, ranging from 20-70 Hz with a mean of approximately 50 Hz, which is approximately one octave below the frequencies noted for the normal male modal register. Vocal fry results from a train of discrete laryngeal excitations, or "pulses", manifested as vertical striation in the spectrogram, and there is nearly complete damping of the vocal tract between successive excitations [16]. The effect of continual, separate taps in rapid sequence is an essential

part of the characteristics auditory quality of vocal fry.

### 3. Method

#### 3.1 Speech Material

A base corpus and a test corpus were designed to factor out the interactions between personal/segmental voice quality differences and voice quality differences induced by focus, and the factors manipulated in the corpora were under two large categories: linguistic factor and personal voice quality. Segmental features of vowel in CV syllable, lexical tones, and narrow focus (sentential stress) are linguistic factors under control.

Disyllabic words of CVCV structure are used as test words, and the first CV syllable is the test syllable controlled for vowel feature, tonal feature, and narrow focus, and thus receives careful inspection for voice quality change. The phonetic transcription of the disyllabic test words is [timo], [tamo], and [tumo], with four different tones assigned on the first CV and only Tone 2 on the second CV. Therefore, there are four test sentences, each containing a word with the same tone but different vowel segments ( $384 \times 2 = 768$  measurements).

The test corpus consists of test words inserted in test sentences as responses to corresponding prompt questions. Data from test sentences will help identify if a change in voice quality is due to focus, tones, or the interaction of the two. The base corpus consists of test words inserted in base sentences with focus on words other than test words and corresponding prompt questions, which ask for the specific information bearing focus in the base sentences. The words receiving focus in base sentences are highlighted with boldface type and bigger font.

The following examples are from test corpus (1) and base corpus (2), including prompt question asking for the specific information receiving focus in the test sentence/base sentence.

- (1) *Prompt question:* 刚才你说的是地膜这个词么?  
*Did you say [ti4 mo2] just now?*  
*Test sentence:* 我说的是[底膜 这个词]  
*What I said is [ti3 mo2] this word.*
- (2) *Prompt question:* 刚才[底膜 是谁说的?  
*Who said [ti3 mo2] just now?*  
*Base sentence:* 刚才[底膜 是]我说的。  
*Just now [ti3 mo2] is said by me.*

The word in brackets is test word, inserted in test sentence and base sentence, respectively

#### 3.2 Subjects

Six native speakers of Mandarin, with normal hearing capability, were recruited and analyzed for this study, three male and three female, all of whom are between 20-30 years of age with no hearing and voice disorders.

#### 3.3 Recording procedures

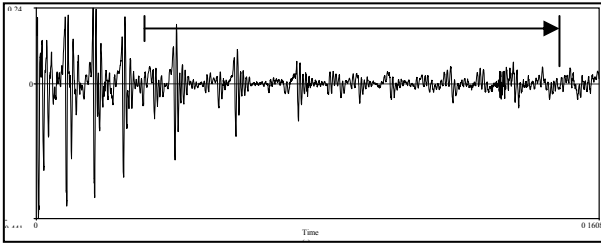
Subjects were recorded in a sound booth, and the mouth-to-microphone distance is held constant about 10cm. Signals were sampled at 48 kHz in PCM format. Speakers were asked to read both the prompt questions and base sentences at a comfortable loudness level and at normal speech rate. Each sentence was repeated six times with three-to-five-seconds intervals, and the middle four repetitions were analyzed.

#### 3.4 Classification of non-modal voice quality

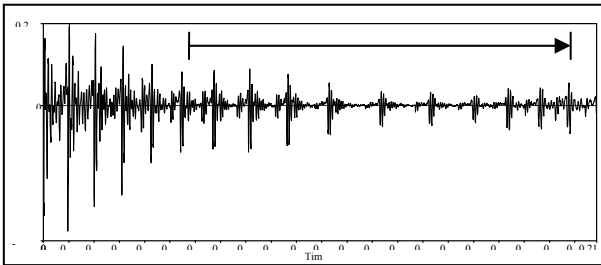
Visual inspection of the speech waveform and spectrogram is used as the major criterion of inspection for non-modal voice quality. Conceptually, periodicity is referred to as the property of modal voice quality and aperiodicity is the major characteristic of creaky phonation, but perturbation from neurological, biomechanical, aerodynamic, and acoustic sources make the perfect periodicity physically impossible [15]. Therefore, this study only discusses the obvious non-modal pattern deviating significantly from periodic pattern, instead of closely examining the small random jitter or shimmer.

Four acoustic categories of non-modal quality were proposed, based on the voice data collected from these eight subjects. The four categories were: (1) *aperiodicity*: irregularities in duration of glottal pulses from cycle to cycle (Fig.1); (2) *creak*: prolonged low fundamental frequency or almost total damping of glottal pulses of low fundamental frequency (Fig. 2); (3) *breathy*: characterized by supraglottal turbulence or noise caused by complete lack of vocal fold vibration and glottal pulses (Fig. 3); (4) *phonatory break*: the vibration of the vocal folds stop complete for a duration of time, which is visually identified from the acoustic waveform for each trial of the sustained vowel samples (Fig. 4). Here, I distinguish aperiodicity from creak, instead of using creaky voice as a whole, in that most male subjects produce rather periodic glottal pulses with very low F0, and damping of glottal pulses is exhibited in their sound samples.

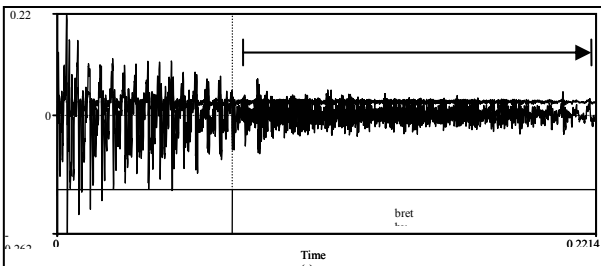
Both perceptual and acoustic evidence are sought concomitantly for each token. For aperiodicity, visual displays (e.g., spectrograms, phase portraits, or next-cycle parameter contours) are most useful for understanding the physical characteristics of the oscillating system [15]. When transglottal pressure, vocal-fold tension, and vocal-fold adduction are in particular ranges, the vibrations might become irregular, and the acoustic manifestation is the unevenly spaced glottal pulses in the waveform. Therefore, the major criterion of inspection for aperiodicity is to examine waveform and find the unevenly spaced glottal pulses. As long as the glottal pulses are irregularly spaced, they are classified as "Aperiodicity". Breathiness is the result of excessive air leakage at the glottis when the vocal folds do not fully approximate during phonation [6]. For moderate voicing, voicing may continue but the noise component in the mid or higher frequency range is stronger than the modal periodic component [9]. However, in this study, only the extreme case of breathiness is recognized: extremely large glottal opening producing noise (aspiration) without voicing. There is also an extreme case of non-modal voice quality, phonatory break, due to vocal folds' failure to vibrate. Since the periodic vibration involves sufficient transglottal pressure, moderate vocal fold tension and resistance. The vibration of the vocal folds is the result of the interaction of both transglottal pressure and the vocal fold resistance. Extremely low transglottal pressure or too much tension in the vocal folds might limit the vibration of the vocal folds or even prevent vocal folds from vibrating. In this case, phonatory break occurs, which results from the vocal folds' failure to vibrate.



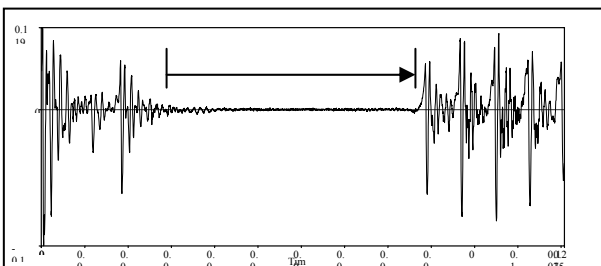
**Fig. 1** Items of non-modal voice quality exhibiting aperiodicity. The part of waveform in between two added vertical lines demonstrates the quality of aperiodicity.



**Fig. 2** Items of non-modal voice quality exhibiting crack. The part of waveform in between two added vertical lines demonstrates the quality of crack.



**Fig. 3** An item of non-modal voice quality exhibiting noise-like breathy voice quality. The part of waveform in between two added vertical lines demonstrates the quality of noise-like breathy voice quality.

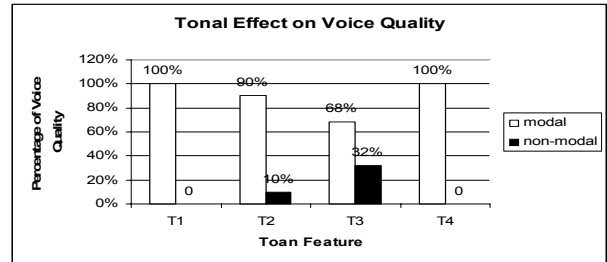


**Fig. 4** An item of non-modal voice quality exhibiting phonatory break. The part of waveform in between two added vertical lines demonstrates the quality of phonatory break.

## 4. Results

### 4.1 The effect of tones

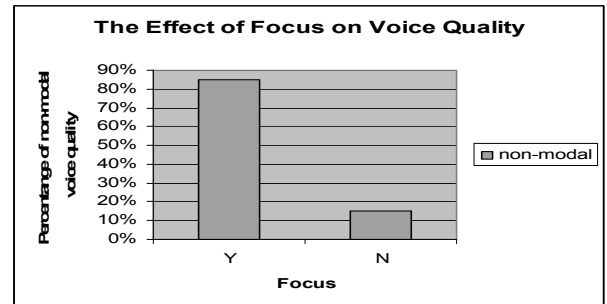
It is obvious from Fig. 5 that only T2 and T3 would induce voice quality change, and T1 and T4 seem to lack the potential for the implementation of non-modal voice. T3 has much higher rates of non-modal voice quality than T2. The higher rate of non-modal voice quality of T3 may be contributed by the presence of focus, vowel feature, or personal factors, which will be discussed in the following section, respectively.



**Fig. 5.** Percentage of non-modal voice quality induced by Tone 1, Tone 2, Tone 3, and Tone 4 across 8 speakers, base corpus, and test corpus.

### 4.2 The effect of focus

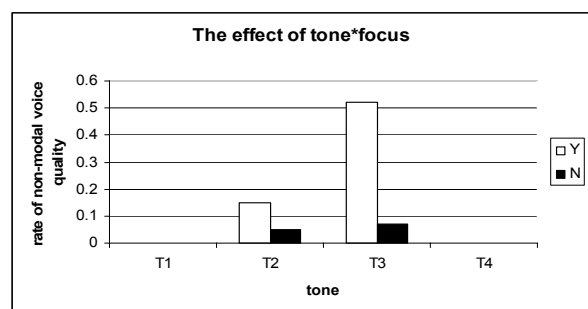
Fig. 6 shows that focus has a significant effect on the realization of non-modal voice quality, across individual, tones, and vowels.



**Fig. 6.** Percentage of tokens of non-modal voice quality across individual, tonal feature, and vowel feature. In the graph, Y means the presence of focus, and N means the absence of focus.

### 4.3 The effect of the interaction between tone and focus

It is obvious in Fig. 7 that when focus is assigned to T3, it produces significantly higher rate of non-modal voice quality than the other three tones, without considering personal factors, which is further illustrated in Fig. 7 demonstrating individual variation in the use of non-modal voice quality to produce tones under focus.



**Fig. 7.** The rates of non-modal voice quality as an effect of the interaction between tone and focus, across 8 speakers and vowel features

There is a consistent across-individual effect of the interaction of Tone 3 and focus on the occurrence of non-modal voice quality, shown in Fig. 8. T2 under focus also could produce non-modal voice quality for some speakers more significantly than T1 and T4. A one-sided *t* test indicates that T3 under focus has a much more

significant effect on non-modal voice quality than T2 ( $P < 0.001$ ).

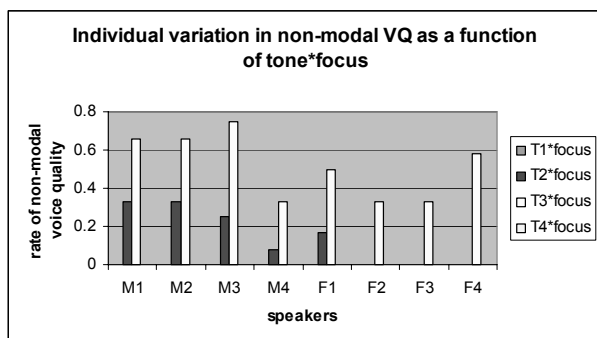


Fig. 8. Individual variation in the rate of non-modal voice quality occurrence when narrow focus is allocated.

#### 4.4 Individual variation in preferred acoustic characteristics

It is clear in Fig. 9 that speakers utilized various characteristics at different rates, with some speakers apparently preferring certain characteristics over others. Aperiodicity and creak are two most preferred acoustic characteristics, while breathy and phonatory break are specifically produced by one subject, respectively.

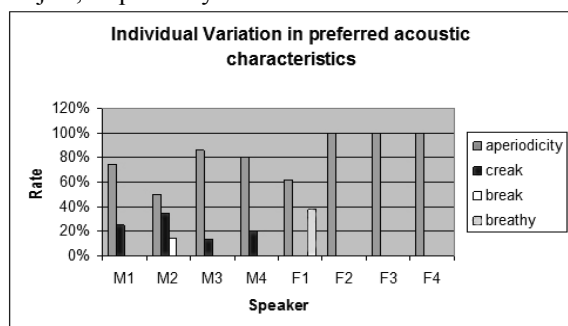


Fig. 9. Percentage of non-modal tokens exhibiting aperiodicity, creak, break, and breathy quality for eight subjects (four male and four female).

#### 5. Conclusions

Results from this study confirm the hypothesis that normal speakers exhibit non-modal voice quality in association with the focus in a sentence in spoken Mandarin utterances, especially when focus is allocated on Tone 3. At the same time, speech samples from both male and female showed a wide range of non-modal voice quality rates and significant differences in preferred acoustic characteristics.

One explanation may be sought from articulatory perspective. T2 and T3 have similar tonal contour, but T3 dips into the lower frequency region while T2 doesn't. When under focus, the tonal contour largely expanded, the dipping action of T3 is fully realized or even emphasized with extra effort of articulation, and the starting part of the T2 is lowered to make the rising action more salient. However, some people cannot produce with this extremely low frequency with modal voice quality, and substitute with various kinds of non-modal voice quality, as indicated in this study. Since voice quality is not

phonemic in Mandarin, great individual variation exists in terms of rates of non-modal voice quality and preferred acoustic characteristics.

[14] found that listeners perceived sound would be lower in pitch as the amount of modulation (amplitude and frequency modulation) increases, which is actually equivalent to aperiodicity in the present study. In this study, aperiodicity, creak, breathy and phonatory break are all used to fulfill the lower frequency range of tonal contour (if the hypothesis is valid), and it didn't create problems for the identification of tonal identity. In this case, the fact that speakers utilize various non-modal voice qualities to achieve the extremely low frequency is probably due to the fact that this low F0 is an important acoustic cue for T3 perception, and people intentionally amplify the relevant acoustic characteristic when it is established as a perceptual cue. At this point, extremely low F0 is the significant factor inducing non-modal voice quality, and, on the other hand, utilization of various non-modal voice qualities is a way of enhancing F0 lowering. Another significance of this low F0 effect is that it points to the possibility that utterance-final creak might be an indirect way to enhance F0 lowering, which is probably more directly related to boundary marking. However, further evidences are to be found in more detailed and well-designed studies.

#### Acknowledgements

I would like to thank Xu Yi for his comments on this work.

#### References

- [1] Campbell, W. N. (1995). Loudness, spectral tilt, and perceived prominence in dialogues. *Proceedings of the XIIIth International Congress of Phonetic Sciences, Stockholm 3*, 676-679
- [2] Chao, Y. R., (1930) A system of "tone letters". *Le Maitre Phonétique* 45, 24-27.
- [3] Dilley, L., S. Shattuck-Hufnagel and M. Ostendorf (1996) Glottalization of word-initial vowels as a function of prosodic structure. *Journal of Phonetics* 24, 423-444.
- [4] Gobl, E. (1988) voice source dynamics in connected speech. *STL-QPSR* 1, 123-159.
- [5] Gordon, Matthew & Ladefoged, Peter (2001) Phonation types: a cross-linguistic overview, *Journal of phonetics*, 29, 383-406 Shattuck-Hufnagel S. and Dilley L. (1996) Glottalization of word-initial vowels as a function of prosodic structure. *Journal of Phonetics* 24, 423-444.
- [6] Hanson, M. Helen, Stevens, N. Kenneth, Kuo, Jeff, Hong-Kwang (2001) Towards models of phonation. *Journal of Phonetics*, 29, 451-480.
- [7] Hollien, H. (1974). On vocal register. *Journal of Phonetics*, 2, 25-43
- [8] Johnson, K. (1997) *Acoustic and auditory phonetics*. Blackwell
- [9] Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, 87, 820-857.
- [10] Pierrehumbert J. & Talkin (1992) lenition of /h/ and glottal stop. In *Papers in laboratory phonology II: gesture segment prosody* (g. Doherty and D. R. Ladd, editors), 90-117. Cambridge: Cambridge University Press.

- [11] Pierrehumbert, J. (1995) Prosodic effects on glottal allophones. In *Vocal fold physiology: voice quality control* (O. Fujimura and M. Irano, editors), 39-60. San Diego: Singular Publishing Group.
- [12] Redi, L. and Shattuck-Hufnagel (2001) Variation in the realization of glottalization in normal speakers. *Journal of phonetics* 29, 407-429.
- [13] Shattuck-Hufnagel S. and Dilley L. (1996) Glottalization of word-initial vowels as a function of prosodic structure. *Journal of Phonetics* 24, 423-444.
- [14] Sun, X. and Xu, Y. (2002). Perceived pitch of synthesized voice with alternate cycles. *Journal of Voice*, 16: 443-459
- [15] Titze R. Ingo (2000) *Principles of Voice Production*. National center for voice and speech, Iowa City, IA.
- [16] Wendahl, R., Moore, P., and Hollien, H. (1963). Comments on vocal fry. *Flia Phoniatic*. 15:251-255
- [17] Xu, Yi (1999) Effects of tone and focus on the formation and alignment of F0 contour. *Journal of Phonetics* 27, 55-105
- [18] Xu, Yi & Wang Q. Emily (2001) Pitch targets and their realization: Evidence from Mandarin Chinese, *Speech Communication* 33, pp. 319-337