THE ROLE OF VOICE QUALITY IN SHANGHAI TONE PERCEPTION

Jiayin GAO[1], Pierre HALLÉ[2]

[1]LPP (Paris 3–CNRS); [2]LMC (Paris 5–INSERM)
jiayin.gao@univ-paris3.fr; pierre.halle@univ-paris3.fr

ABSTRACT

This study investigates the perceptual aspect of voice quality in Shanghai Chinese. Previous studies show that breathy voice is a redundant feature of low tones, but tends to disappear in young speakers' productions. Does this mean that voice quality is not playing any role in tone perception? To find this out, we conducted forced-choice identification tests with young listeners, using synthesized and natural, breathy and modal syllables as stimuli. Our results show that breathy voice still is an important cue for the perception of low tone syllables, with the exception of nasal onset syllables.

Keywords: Shanghai Chinese, perception, breathy voice, tone

1. INTRODUCTION

In Shanghai Chinese, low tone or "yang" syllables are accompanied with some slightly breathy voice, whereas high tone or "yin" syllables are not. This is supported by impressionistic descriptions [7, 12] and phonetic evidence based on acoustic [6, 9], and physiological [13, 14] measurements. A recent study [11] showed that young speakers produce less phonation difference between high and low tones, suggesting an evolution towards the loss of breathy voice in the production of yang syllables. Yet, is breathiness still *perceived* as a cue to yang syllables?

In this study, we test whether breathy voice helps perceiving low tone, yang syllables. Previous studies on Shanghai tone perception are scarce [5,14,10]. To our knowledge, the role of breathy voice in Shanghai tone perception has only been examined in Ren's dissertation [14]. His data suggest that breathy voice can be used as a perceptual cue to a syllable's yang identity. We therefore summarize his study. In Shanghai Chinese disyllabic words, the tonal contour realized on the second syllable is determined by the tone of the first syllable by virtue of a tone sandhi rule. That is, the tonal component of the original yin or yang identity of the second syllable is, in principle, lost. Ren asked whether voice quality and/or F0 onset height differences might nevertheless help to recover this lost identity. He used the potentially ambiguous sequence [ɦa.ta], which, he claimed, may correspond to either 鞋带 'shoelace' or 鞋大 'the shoe is big'; [ta] in the former and latter [ɦa.ta] originally bears the yin tone T2 (34 in Chao's notation [8]) and the yang tone T3 (23), respectively, but the sandhi imposes a 22-33 contour on both [ɦa.ta] sequences, determined by the tone T3 of [ɦa]. Ren synthesized [ɦa.ta] stimuli, varying [ta] along two dimensions: F0 onset height (three contours from 110, 120, or 130 Hz to 135 Hz) and voice quality (from breathy to modal continua). Listeners were tested with these stimuli on a forced-choice identification task with 鞋带 'shoelace' and 鞋大 'the shoe is big' as the two possible responses. Both manipulated dimensions influenced listeners' responses. Higher F0 onsets induced more yin, high tone /ta/ responses. For each F0 onset, most listeners perceived the voice quality continua in a categorical way, from yin to yang, based on breathiness: breathy voice stimuli induced a large majority of yang, low tone /da/ responses, and modal voice stimuli a large majority of yin, high tone /ta/ responses. These findings thus suggest that listeners perceived breathy voice as a cue to the yang identity of a syllable at the time of the study. But Shanghai Chinese changed in the meantime, with a trend toward disuse of the breathy voice quality in yang syllables. This is the first reason why we reexamine the issue of the role of breathy voice in perception. The second reason is that Ren's study has some shortcomings. As a main concern, while 鞋带 'shoelace' is a frequent word, indeed pronounced [ɦa22.ta33] after sandhi, 鞋大 'the shoe is big' is an artificially constructed verb phrase (鞋 would be more acceptable if suffixed with 子[tsɻ]), in which the application of the sandhi rule is problematic: if it applies, the yang syllable 大 /da/ should surface as [da], not [ta], with little or no breathiness (see [6]); if not, the two syllables 鞋大 should be pronounced as if they were word-initial, that is, [ɦa23.ta23]. In both cases, the critical ambiguity in Ren's material, segmental and tonal at the same time, is quite artificial. To substantiate this point, we informally asked nine native speakers of Shanghai Chinese to read aloud the questionable sequence 鞋大. Five of them produced it as [ɦa.tu], with the second syllable in its original tone T3; four of them produced it as [ɦa.du] and applied the tone sandhi rule for lexicalized words. (Note that 大 can

be read with the rime /a/ or, more frequently, /u/.) All nine speakers judged the sequence unnatural anyway. The listeners in Ren's study, we believe, might have focused on which of *isolated* 大 /da/ or 带 /ta/ the synthesized [ta] matched better.

In the present study, instead of breathy voice continua, we used tone contour continua from yin T2 to yang T3 (34 to 23) with two patterns of voice quality (breathy and modal). We reasoned that breathy voice, if still used in perception, should bias stimulus identification toward yang responses. We tested monosyllabic targets, with stop, fricative, or nasal onsets. The continua we used were constructed from either natural or synthesized stimuli.

2. METHOD

We conducted forced-choice identification tests on the T2-T3 continua constructed from either natural or synthesized stimuli. For both types of stimuli, we contrasted breathy and modal voice. The stimuli were constructed from monosyllabic minimal pairs sharing their segments and differing in tone height: yin tone T2 (34) vs. yang tone T3 (23).

2.1. Participants

Sixteen Shanghai Chinese native speakers (5 males), aged 18-26 years participated in the study. All were born in Shanghai urban area, except one, born in Japan but brought back to Shanghai at age one. All had spent most of their lifetime in Shanghai and reported normal hearing and reading. We excluded the data of one male subject on synthesized stimuli and the data of two male subjects on natural stimuli, due to high missing response rates.

2.2. Stimuli

Stimuli were derived from "base" syllables sharing the /ɛ/ rime, with six different onsets: /∅ (zero), p, t, f, s, m/. Two eight-step tone contour continua (one breathy, one modal) were constructed for each onset from synthesized or natural base syllables. For each step, a contour between T2 and T3 was imposed on a base syllable. Table 1 shows the endpoint syllables of the continua (tones T2 and T3). The synthesized /m/ stimuli sounded unnatural, hence were not used.

Table 1: Endpoint syllables at tones T2 and T3; /mɛ/ was not retained for synthesized stimuli.

	∅	stop	fricative	nasal
T2	ɛ 爱	pɛ 板 tɛ 胆	fɛ 反 sɛ 伞	mɛ 美
T3	ɛ 咸	pɛ 办 tɛ 台	fɛ 烦 sɛ 馋	mɛ 梅

We used a two-mass model with triangular glottis [3] in VocalTractLab 2.1 [17] to synthesize one modal and one breathy base syllable for each onset (modal: default parameters for modal phonation; breathy: lower and upper rest displacement set respectively to 0.60 and 0.55 mm, subglottal pressure to 1300 Pa), using a male voice and a flat F0 contour. H1-H2 was measured to check for voice quality in the synthesized syllables. H1-H2 was consistently higher for breathy than modal base syllables for all five onsets, and across the entire syllable.

T3 vs. T2 natural syllables served as natural base syllables for the breathy vs. modal continua, respectively. A Shanghai Chinese female native speaker, trained phonetician, aged 26, produced the base syllables with deliberate breathy voice for T3 vs. modal voice for T2 syllables for all six onsets. Breathy phonation was checked from H1-H2, which was indeed consistently higher for T3 than T2 syllables, for all six onsets, from vowel onset up to ~80% in the vowel; the differential was largest at vowel onset and decreased throughout the vowel, as observed in naive Shanghai speakers' productions [11]. For each onset, the T2 and T3 syllables were time-scaled so that they had the same onset and vowel durations, since duration patterns influence the perception of yin vs. yang identity (cf. [10]).

All base syllables, synthesized or natural, were intensity-equalized at 80 dB SPL. A continuum of eight equidistant, stylized tone contours between T2 and T3 were imposed on each base syllable, using the Praat [4] implementation of PSOLA [16]. This yielded 12 natural and 10 synthesized continua (2 voice qualities × 6 or 5 onsets). For synthesized continua, the endpoint T2 and T3 contours were taken from the productions of a 74-year-old male speaker. For the natural continua, they were those produced by the trained phonetician for each onset.

2.3. Procedure

The identification test was run using E-Prime 2.0. Participants were tested individually in a quiet room, seated in front of a laptop. Stimuli were presented through professional quality headphones. On each trial, which began with a fixation cross, an auditory stimulus was presented and two Chinese characters (a T2-T3 minimal pair: see Table 1) for the two possible responses appeared on the left and right sides of the screen. Each stimulus was presented twice, switching the side of the T2 and T3 responses.

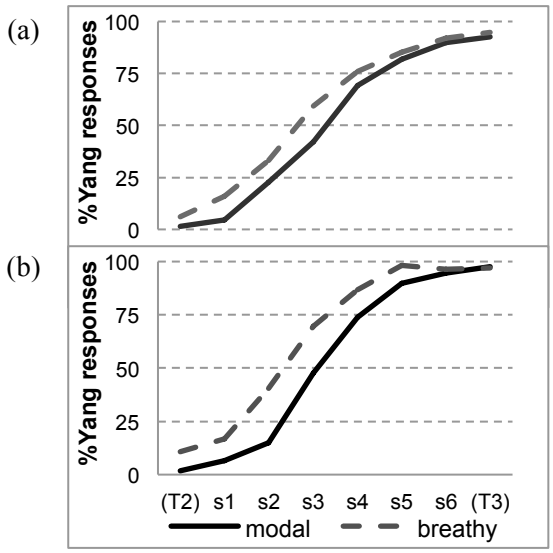
The stimuli were blocked by construction type. Half the participants were tested first on the natural stimuli, and the other half on the synthesized stimuli. Within each block, the stimuli were presented in

random order. The experiment was preceded by a training phase of six trials, with syllable stimuli bearing canonical yin or yang tones. Participants received accuracy and response time feedback during the training but not during the test.

3. RESULTS

Fig. 1 shows the identification functions in terms of "yang" (i.e., T3) response rate, according to voice quality and construction type. The functions are quite categorical, and shifted toward the yin endpoint for breathy relative to modal voice quality.

Figure 1: Identification curves according to voice quality, for (a) synthesized and (b) natural stimuli.



3.1. 50% yang boundary location

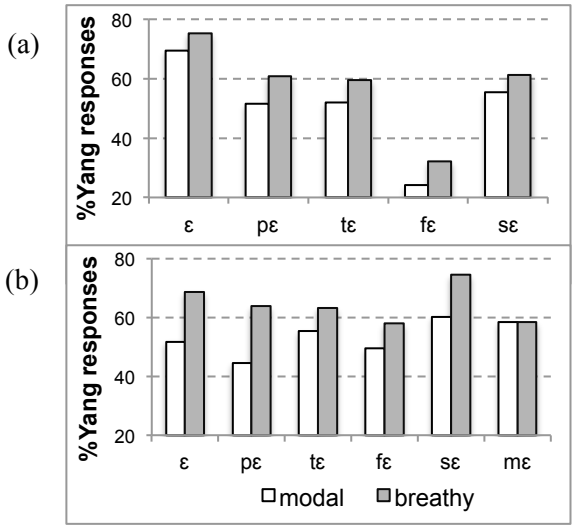
We first estimated the 50% boundary locations, or intercepts, on each continuum for each subject from probit analyses fitting short ogive Gaussians to the data, yielding intercepts and slopes [2]. For synthesized stimuli, the intercepts averaged to 2.60 vs. 3.21 (step#) for breathy vs. modal continua, respectively. For natural stimuli, similar values obtained: 2.51 vs. 3.06. These differences indicate a shift toward yang responses for breathy compared to modal voice continua. We ran by-subject ANOVAs on the intercept data separately for synthesized and natural stimulus data, with *Voice-quality* (modal vs. breathy) and *Onset* (/∅, p, t, f, s/, plus /m/ for natural stimuli) as within-subject factors.

Voice-quality was significant for both natural and synthesized stimulus data, with a boundary closer to the yin endpoint for breathy than modal syllables (natural: $F(1,13)=22.1$; synthesized: $F(1,14)=23.3$; ps<.0005). *Voice-quality* × *Onset* interaction was not significant for synthesized stimuli, $F(4,56)=1.1$, $p=.34$. It was significant for natural stimuli, $F(5,65)=2.8$, $p<.05$: paired comparisons showed that *Voice-quality* was significant for /∅, p, s/ at least at the $p<.05$ level, with breathy–modal differences ranging from 0.65 to 1.14 steps; *Voice-quality* was not significant for /t, f, m/, with breathy–modal differences ranging from 0.11 to 0.36 steps.

3.2. Overall percentage of yang responses

In the synthesized stimulus data, the overall rate of yang responses was 57.8% for breathy syllables vs. 50.6% for modal syllables. In the natural stimulus data, these rates were 64.5% for breathy syllables vs. 53.3% for modal syllables. Fig. 2 shows the yang response rate averaged across continuum steps, as a function of onset and voice quality. For /m/, there is no advantage for breathy over modal voice quality.

Figure 2: Overall yang response rates by onset, for (a) synthesized and (b) natural stimuli.



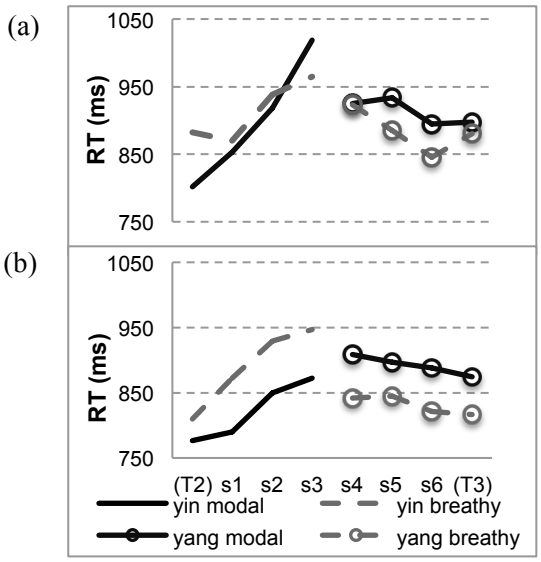
A series of generalized linear models (GLM) were fit to the binomial yin/yang response data, using the *lme4* package [1] in R [18], separately for synthesized and natural stimuli. The fixed effects were *Voice-quality* (modal vs. breathy), *Onset* (/∅, p, t, f, s/ plus /m/ for natural stimuli) and *Step* (0-7). The random effect structure included by-subject random slopes for *Voice-quality* and random intercepts. We used likelihood ratios to compare the full model, which included all the fixed effects and the *Voice* × *Onset* interaction, to models with one fixed effect or interaction missing. We will only discuss the main predictors: *Voice-quality* and *Voice-quality* × *Onset*. *Voice-quality* was significant for both natural and synthesized stimuli (natural: $\chi^2(1)=22.6$, $p<.0001$; synthesized: $\chi^2(1)=13.9$, $p<.0005$). The *Voice-quality* × *Onset* interaction was not significant for synthesized stimuli, $\chi^2(4)=1.1$, $p=.89$, indicating similar effects for the five onsets /∅, p, t, f, s/, but it was for natural stimuli,

$\chi^2(5)=29.7$, $p<.0001$): This significant interaction was due to the onset /m/, which differed from the others by the non-significant effect of *Voice-quality*, as shown by comparing models with and without *Voice-quality* on the data subset restricted to /m/, $\chi^2(1)=0.004$, $p=.95$.

3.3. Response time

Fig. 3 shows the mean response times (RTs) in the identification test according to *Voice-quality*, and for dominant response (yin or yang) only: the dominant response is yin in steps 0-3 and yang in steps 4-7, as can be seen in Fig. 1. For natural stimuli, RTs for the yin response in steps 1-3 were shorter for modal than breathy stimuli, whereas RTs for the yang response in steps 4-7 were shorter for breathy than modal stimuli, suggesting that modal quality facilitates yin responses and breathy quality yang responses. The results were less clear-cut for synthesized stimuli, but with similar trends.

Figure 3: Mean RTs for dominant responses: yin in steps 0-3 (no marks) and yang in steps 4-7 (circles), according to modal vs. breathy stimuli (plain vs. dashed lines), for (a) synthesized and (b) natural stimuli.



We ran by-subject ANOVAs on the RT data separately for synthesized and natural stimulus data. Instead of *Voice-quality*, we used a *Congruence* factor: modal voice stimuli were "congruent" in steps 0-3 and "incongruent" in steps 4-7, and vice versa for breathy voice stimuli. That is, "congruent" meant that stimulus voice quality matched that of the dominant response. *Onset* was defined as in 3.1. All two factors were within-subject factors.

RTs were shorter for congruent than incongruent stimuli, as shown by the effect of *Congruence*: this factor was marginally significant for synthesized stimuli (876<914 ms), $F(1,14)=4.3$, $p=.057$, and significant for natural stimuli (833<893 ms), $F(1,13)=9.6$, $p<.01$. The *Congruence* × *Onset* interaction was significant for natural stimuli only, $F(5,65)=2.9$, $p<.05$, reflecting different effect sizes for different onsets: in particular, the effect size was 0 ms for the /m/ onset, whereas it was 74 ms in average for the other onsets. The RT data are thus quite parallel to the overall yang response rate data, with the /m/ onset standing out for natural stimuli.

4. DISCUSSION

Our study shows that breathy voice biases tone perception in young Shanghai listeners toward the yang category. This conclusion is supported by the shift toward yang responses, for most onsets, induced by breathy compared to modal voice in the identification of T2-T3 continua, as shown both by the shift of the intercepts and by the overall increase in yang response rate, as well as by the RT data. Overall, the results were less clear-cut for synthesized than natural stimulus data, although the latter included /m/ onset stimuli, for which breathy voice had no effects.

The trend toward loss of yang tone breathy voice in production thus would not apply to perception, with the notable exception of /m/ onset syllables.

It may not be purely accidental that nasal onset syllables are among the first to be affected by the loss of a phonation difference between high and low tones. Indeed, such a situation is not unknown. The production data of eight Yue dialects examined by Tsuji [15] (cited in [19]) illustrate different stages in the loss of breathy voice for low tone syllables. Rongxian preserved breathy voice, whereas Cangwu lost it in low tone for all onsets. Between these two stages, which we may view as initial and final, Cenxi lost breathy voice in low tone syllables but only for nasal onset low tone syllables, representing an intermediate stage.

Our perception data suggest that Shanghai Chinese might be on the same tracks as the Cenxi dialect, the loss of breathy voice being more clearly achieved in production, and emerging with nasal onset syllables in perception.

5. ACKNOWLEDGEMENTS

We are very grateful to Peter Birkholz for his help with the VocalTractLab software. This study was supported by LabEx EFL.

5. REFERENCES

[1] Bates, D., Maechler, M., Dai, B. 2008. lme4: Linear mixed-effects models using S4 classes. Version 1.1-7.

[2] Best, C. T., Strange, W. 1992. Effects of phonological and phonetic factors on cross-language perception of approximants. *J. Phon.* 20(3), 305-330.

[3] Birkholz, P., Kröger, B. J., Neuschaefer-Rube, C. 2011. Synthesis of breathy, normal, and phonation using a two-mass modal with a triangular glottis. *Proc. 12th Interspeech*, Florence, 2681-2684.

[4] Boersma, P., Weenink, D. 1992-2015. Praat: doing phonetics by computer, version 5.4.01.

[5] Cao, J. 1987. The ancient initial "voiced" consonants in modern Wu dialects. *Proc. 11th ICPhS*, Tallinn, 169-172.

[6] Cao, J., Maddieson, I. 1992. An exploration of phonation types in Wu dialects of Chinese. *J. Phon.* 20, 77-92.

[7] Chao, Y. 1928. *Studies in the modern Wu dialects.* Peking: Tsinghua College Research Institute.

[8] Chao, Y. 1930. A system of tone letters. *Le Maître Phonetique*, 45, 24-27.

[9] Chen, Y. 2011. How does phonology guide phonetics in segment-f0 interaction? *J. Phon.* 39(4), 612-625.

[10] Gao, J., Hallé, P. 2013. Duration as a secondary cue for perception of voicing and tone in Shanghai Chinese. *Proc. 14th Interspeech*, Lyon, 3157-3161.

[11] Gao, J., Hallé, P. 2013. Are young male speakers losing Tone 3 breathiness in Shanghai Chinese? An acoustic and electroglottographic study. *Proc. 2nd ICPLC*, Hong Kong, 163-166.

[12] Liu, F. 1923. 守温三十六字母排列法之研究 [*Studies on the 36 initials of Shouwen*]. *Guoxue Jikan,* 1(3), 451-464.

[13] Ren, N. 1988. A fiberoptic and transillumination study of Shanghai stops. Paper presented at International Conference on Wu Dialects, Hong Kong, China.

[14] Ren, N. 1992. *An acoustic study of Shanghai stops.* Ph.D thesis, University of Connecticut.

[15] Tsuji, N. 1977. *Eight Yue dialects in Guangxi Province, China, and reconstruction of proto-Yue phonology.* Ph.D thesis, Cornell University.

[16] Valbret, H., Moulines, E., Tubach, J. P. 1992. Voice transformation using PSOLA technique. *Speech Communication* 11(2), 175-187.

[17] VocalTractLab 2.1 http://www.vocaltractlab.de/

[18] R Core Team, R: A language and environment for statistical computing, version 3.0.1.

[19] Yip, M. 1980. *The tonal phonology of Chinese*. Ph.D thesis, MIT.