Original Article

Language *and* Speech

Breathy voice and low-register: A case of trading relation in Shanghai Chinese tone perception?

Language *and* Speech
1–26

DOI: 10.1177/0023830919873080
journals.sagepub.com/home/las
SAGE

Jiayin Gao
Sophia University, Tokyo, Japan
Japan Society for the Promotion of Science, Tokyo, Japan
Laboratoire de Phonétique et Phonologie, CNRS–Université Paris 3, France
Laboratoire Langues et Civilisations à Tradition Orale, CNRS–Université Paris 3, France

Pierre Hallé
Laboratoire de Phonétique et Phonologie, CNRS–Université Paris 3, France
Laboratoire Mémoire et Cognition, INSERM–Université Paris 5, France

Christoph Draxler
Institute of Phonetics and Speech Processing, Ludwig-Maximilian University, Munich, Germany

Abstract

In Shanghai Chinese as well as many other Wu dialects, breathy voice is a well-documented accompaniment of the low-register tone syllables with obstruent as well as sonorant onsets. But Shanghai Chinese is rapidly changing and the breathy voice associated with low-register tones tends to disappear in young speakers' productions. In this study, we asked whether breathy voice is nevertheless still perceived and whether it pushes tone identification toward low-register tones. We conducted forced-choice tone identification tests on young native listeners of Shanghai Chinese, using low–high register tone continua—from tone T3 (23) to tone T2 (34)—imposed on base syllables with either modal or breathy voice quality, and beginning with various onset consonants. We used continua constructed from either naturally produced or synthesized syllables. Our results show that breathy voice does bias tone identification responses toward the low-register tone T3. This result held for both synthesized and natural stimuli, except for the /m/-onset stimuli derived from naturally produced syllables. We propose that the phonetic change at issue—loss of breathiness in production—is not due to misperception but reflects the ever-stronger influence of Standard Mandarin Chinese. In other words, this particular case of sound change seems to be led by production rather than perception. It remains an open question whether this kind of sound change is only determined by sociolinguistic factors (here, the dominance of Mandarin Chinese) or is independently motivated by phonetic and/or phonological factors.

Corresponding author:
Jiayin Gao, Sophia University, 7–1 Kioi-chô, Chiyoda-ku, Tokyo, 102-8554, Japan.
Email: jiayin.gao@gmail.com

Keywords
Shanghai Chinese, tone perception, voice quality, breathy voice, trading relation

Introduction

Although pitch is the primary cue to tone identity in many tone languages, in a large number of East and Southeast Asian languages, it has been suggested that tone is multidimensional in nature, comprising several simultaneous phonetic properties, including pitch and non-pitch cues (Abramson & Luangthongkum, 2009; Rose, 1982). Non-pitch cues include, for example, voicing and duration of initial consonant, rime duration, intensity profile, vowel quality, and, importantly for the present study, voice quality.

The perceptual reality of a multidimensional tone specification is strongly suggested by the contribution of non-pitch cues to the perception of tone contrasts in these languages. First, listeners may rely on non-pitch cues to identify tone, especially if the salience of pitch is reduced. In Mandarin Chinese, for example, Whalen and Xu (1992) showed, using "signal-correlated noise" stimuli, that the intensity contour is sufficient for listeners to identify tone (excepted the high-level tone 1) in the absence of pitch information. Secondly, previous studies suggested that, for those languages whose tonal space is crowded in the sense that the language contains few clearly distinct pitch patterns, or, crucially, includes confusable tone-contour pairs, non-pitch cues are used by speakers/listeners to enhance perceptual distinctiveness. Among non-pitch cues, voice quality is frequently found to interact with pitch in tone languages and may be used as a perceptual cue to tone contrasts. In Sgaw Karen, in which the tonal space is "crowded," comprising one slightly rising tone and five moderately falling tones (hence potentially confusable tones), duration and voice quality play central perceptual roles in addition to pitch (Brunelle & Finkeldey, 2011). In Northern Vietnamese, which also has a system of six tones, some of which share similar pitch contours and heights, voice quality is used to characterize some of the tones; rime-end glottalization is even the primary cue to identifying one of the six tones (Brunelle, 2009). Likewise, in Hmong languages, the tone system is usually very dense and voice quality is used to distinguish between tones with similar pitch heights and contours (for Black Miao, Kuang, 2013; for White Miao, Garellek, Keating, Esposito, & Kreiman, 2013). In Cantonese, creakiness contributes to the identification of the low falling tone, although this realization is quite variable between and within speakers (Yu & Lam, 2014). In contrast, the creakiness often produced with the low tone 3 in Mandarin Chinese (Belotel-Grenié & Grenié, 1997, 2004) is tied to low pitch targets (Kuang, 2017) and appears to have little effect on tone perception (Belotel-Grenié & Grenié, 1997; Gårding, Kratochvil, Svantesson, & Zhang, 1986): it seems that the creakiness of tone 3 is not motivated by perception and rather is a mechanical consequence of low pitch production. Yu and Lam (2014) interpreted this cross-linguistic difference with reference to tonal inventory. In Mandarin, pitch alone is sufficient to distinguish the four tones, whereas in Cantonese, some tones are very similar in pitch contour and may need other cues than pitch to be properly identified. Other than Sino-Tibetan languages, the coexistence between tone and voice quality is also found in other language families (for a review, see Gordon & Ladefoged, 2001), notably in Oto-Manguean languages (e.g., for Jalapa Mazatec, Silverman, Blankenship, Kirk, & Ladefoged, 1995; for Zapotec languages, see Esposito, 2013, p. 12ff, for an overview).

Shanghai Chinese belongs to the Wu family of Chinese dialects and, as the other Chinese dialects, is a tone language in which pitch is the primary cue to tonal contrasts. In this study, we investigate tone identification in the variety of Shanghai Chinese spoken in the urban area of Shanghai (hereafter Shanghai Chinese). There are five lexical tones in Shanghai Chinese: three

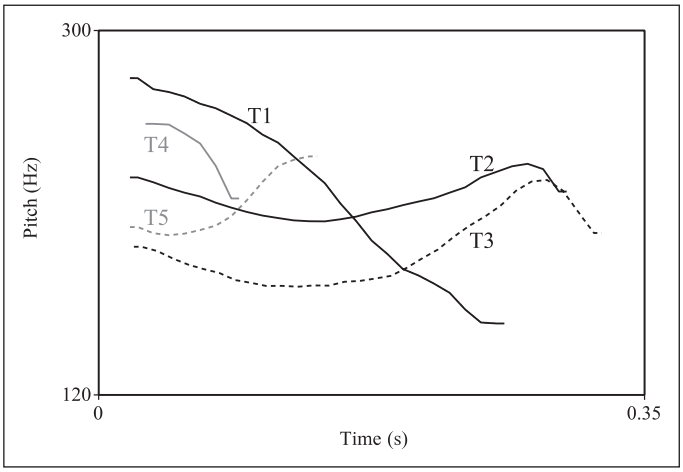

Figure 1. F0 contours of the five lexical tones, derived from electroglottographic signals: [tɛ] (T1–3) and [tɐʔ] (T4, T5) produced by a young female speaker (from the data in Gao & Hallé, 2017).

unchecked tones (tone 1: 52 ˥˧; tone 2: 34 ˧˦; tone 3: 23 ˨˧) and two checked, shorter tones (tone 4: 55 ˥; tone 5: 12 ˨), referred to as T1, T2, T3, T4, and T5, respectively, throughout this paper. The tone values we indicate were proposed by Xu and Tang (1988, p. 8); slightly different values have been proposed by others, such as 13 instead of 23 for T3 (Zee, 2003; Zhang & Yan, 2018; Zhu, 1999). T1 to T5 are illustrated by the following words: [pi˥˧] "border," [pi˧˦] "compare," [pi˨˧] "leather," [pɪʔ˥] "pen," [pɪʔ˨] "nose." T3 and T5 are low, or *yang* tones as opposed to the high, or *yin* tones, T1, T2 and T4. In Shanghai Chinese, tone interacts with voice quality. The low-register tones (or *yang* tones) T3 and T5 are produced with breathy voice. This is not the case for the high-register tones T1, T2, and T4. According to the early impressionistic descriptions of Karlgren (1915–1926), Chao (1928, p. xii), and Liu (1923, p. 459), the stop onsets of *yang* tone syllables are produced with breathy release. These descriptions later received support from phonetic investigations conducted on stop onsets, comparing *yang* to *yin* tone syllables (Cao & Maddieson, 1992; Chen, 2011; Gao, 2016; Gao & Hallé, 2017; Ren, 1992; Tian & Kuang, 2016). Cao and Maddieson (1992) found that breathiness only affects the release portion of the stops. Gao and Hallé (2017) found that breathiness, when present, also affects the first half of syllable's rime. In several studies (Gao, 2016; Gao & Hallé, 2017; Tian & Kuang, 2016), cross-generational comparisons were conducted with both acoustic and EGG (open quotient) measurements in order to examine the breathy voice produced with *yang* versus *yin* tones by two generations. Their results consistently showed that the breathy voice associated with *yang* tones is less systematically produced by young speakers than elderly speakers, suggesting a trend toward disappearance of breathy voice as a redundant cue to low tones, possibly reflecting the growing influence of Mandarin, in which breathy voice is unknown (see 3.3 for more discussion).

Figure 1 illustrates the F0 contours for each of the five tones of Shanghai Chinese produced in citation form. Similar patterns are reported in Chen and Gussenhoven (2015), except for the duration of T1, which they found shorter. Some linguists consider T4 and T5 as the checked variants of unchecked T2 and T3 (Yip, 1980, p. 201), which reduces the tonal inventory to three tones. If we surmise that checked tones are clearly distinct perceptually from unchecked tones, based on presence of a glottal-stop coda, duration, and vowel quality, the three unchecked tones T1–3 then occupy a rather non-crowded tonal space. Yet, the tone-letter notation proposed by Xu and Tang

(1988) suggests that T2 (34) and T3 (23) are relatively close to each other. Chen and Gussenhoven (2015) reported an F0 difference at rime onset between T2 and T3 of about 50 Hz. Gao and Hallé (2017) reported T2–T3 differences at rime onset of the same order of magnitude (about 50–60 Hz in average, 4.0 to 5.8 semitones), except for some young male speakers (only 22 Hz, 2.9 semitones), whose tonal space was narrower than young female and senior speakers. These pitch differences are well above those required to perceive a tonal contrast (’t Hart, 1981), except in the case of the young male speakers we mentioned, for which the difference may be only marginally sufficient (cf. Kuang, 2013). Therefore, we may expect tones in Shanghai Chinese to be easily distinguishable based solely on their pitch contour, even T2–T3 produced by non-extreme young males. On that assumption, non-pitch cues would not be *necessary* to help tone identification. In particular, breathy voice might not help perceiving low-register, *yang* tones. That the breathy voice associated with *yang* tones tends to disappear in young Shanghai speakers’ productions would also suggest that breathy voice might not help perceiving the *yang* tones of Shanghai Chinese. Conversely, if breathy voice still is a cue to *yang*-ness in Shanghai Chinese tone perception, it should bias stimulus identification toward *yang* tone perception.

Testing these two opposite predictions motivates the present study. We manipulated breathiness in synthetic and naturally produced syllables in order to find out whether or not breathiness plays a role in *yang* versus *yin* tone identification. If breathiness plays a role in tone identification, breathy voice syllables should tend to be identified as carrying a *yang* tone more often than their modal voice counterpart, everything else being equal.

Thus far, the impact of breathiness on Shanghai Chinese lexical tone perception has been little investigated. Gao and Hallé (2013) showed that listeners primarily rely on pitch to identify *yin* versus *yang* tone syllables. They manipulated naturally produced *yin* or *yang* syllables to create syllables retaining their stylized original tone contour versus syllables carrying the stylized contour of the opposite *yin* or *yang* category. That is, they crossed the F0 and non-F0 information in the original *yin* and *yang* tone syllables in order to test whether a mismatch between F0 contour and other characteristics—which, importantly, include voice quality—would impede identification performance. Mismatch had no effect on identification accuracy except for syllables with a labial fricative onset. The case of labial fricative onsets could be explained by the fact that these onsets are frequently phonetically voiced in *yang* but not *yin* syllables (Gao & Hallé, 2017): voiceless [f] thus does not fit with a *yang* tone, nor voiced [v] with a *yin* tone. Mismatch, however, *did* affect response times: the mismatched stimuli were responded to more slowly than the others. Altogether, Gao and Hallé’s (2013) results showed that Shanghai tone identification is mainly cued by pitch, but also that non-pitch cues somewhat influence listeners’ judgment at a presumably non-conscious level, hence, indirectly indicating that voice quality may play a role.

In a recent study, Zhang and Yan (2018) investigated young Shanghai speakers/listeners’ production and perception of T2 and T3 syllables with stop, fricative, and sonorant onsets, various rimes, and in both monosyllabic and sandhi context. In the following, we summarize their study with respect to the monosyllabic, non-sandhi context. Zhang and Yan analyzed the acoustic cues to the T2–T3 contrast relevant to the onset consonant itself (voicing/voice quality), the rime’s voice quality cue, and the F0 contour cue. They compared the weights of these cues in both production and perception. In production, they found weak evidence for breathy voice in *yang* tone T3 based on either H1*–H2* (H1–H2 corrected for formant frequencies) or Cepstral Peak Prominence, both in the onset and in the rime, and especially so for the sonorant-onset syllables. The main cue they found to tone T2 versus T3 was F0 contour. In their perception study, they started from natural T2 and T3 syllables and constructed stimuli by recombining onset consonant C, vowel V, and F0 contour of each T2–T3 pair in all possible ways (eight combinations including the two original syllables), using cross-splicing and F0 contour superimposition. C was /b, p/ (stops), /v, f/ (fricative), or

/m̥, m/ (sonorant). V was /u, i, ɛ/, F0 contour was 34 (T2) or 13 (T3). For example, they cross-spliced C of *pu34* to V of *bu13* and superimposed on V the F0 contour of *pu34*, producing a *pu34* stimulus in which only voice quality of the rime was changed from modal to breathy. This design allowed to assess separately the role of C, V, and F0 contour properties in T2–T3 categorization. The perception data "were generally consistent" with the production data: "the laryngeal contrast in question [the *yin–yang* contrast in our formulation] was primarily cued by F0 in the non-sandhi context" (Zhang & Yan, 2018, p. 1304). Yet, the contrast was also secondarily cued by C properties (voicing/voice quality) and "Phonation" (rime's breathiness), depending on onset type: "Phonation" affected T2–T3 categorization for both stop- and fricative-onset syllables; C properties only affected fricative-onset syllables, while sonorant-onset syllables were not affected by secondary cues. To sum up, F0 contour was the main cue to distinguish *yang* tone T3 from *yin* tone T2 in monosyllables, in line with Gao and Hallé (2013). Yet, breathiness contributed to *yang* tone identification in stop- and fricative-onset syllables but not in sonorant-onset syllables.

To our knowledge, Ren (1992) was the first study that directly investigated the role of breathiness in identifying Shanghai Chinese tones along a breathiness continuum for three given F0 contours. His findings suggested that breathy voice is used as a perceptual cue to *yang* tone. Zhang and Yan (2018), as we just saw, showed that although F0 contour is the main cue to distinguish *yang* T3 from *yin* T2, breathiness is a secondary cue, at least for syllables with a stop or a fricative onset.

More than 20 years after Ren's pioneering study, we decided to test whether the younger generation of Shanghai Chinese speakers are still exploiting breathy voice to identify *yang* tones, although they tend not to produce it. We, however, took a different approach from both Ren (1992) and Zhang and Yan (2018). We examined whether identification of *yin–yang* tone contour continua could be systematically affected by presence/absence of breathiness, that is, whether there is some kind of trading relation between breathiness and *yin–yang* tone register. Our approach thus requires constructing *yin–yang* tone contour continua on a given tone pair in the first place. We chose unchecked tones T2 (*yin*) and T3 (*yang*) because they share a similar rising contour shape, although the contour may be somewhat flatter in T2 than T3 (Figure 1), and mainly differ by tone height: high- versus low-register for T2 versus T3. Because T2–T3 (mid- vs. low-rising) is more confusable than T1–T3 (high-falling vs. low-rising), T2–T3 is perceptually more susceptible to voice quality influence than is T1–T3. In other words, breathiness and low-register cues are more likely to trade for T2–T3 than for T1–T3. We thus constructed tone contour continua from *yin* T2 to *yang* T3. As detailed more extensively in section 2.1.2, we superimposed the F0 contours of T2–T3 continua on monosyllabic T2–T3 minimal pairs, which were pronounced with breathy voice in tone T3 and modal voice in tone T2 (e.g., /bɛ/ in T3 vs. /pɛ/ in T2: both are realized with a voiceless [p]). We used the main types of syllable onset consonants: zero onset, stop, fricative, and nasal onsets. This makes it possible to compare our data with that recently obtained by Zhang and Yan (2018), who contrasted stops, fricatives and sonorants. Whereas the main goal of Zhang and Yan's study was to assess the weights of the onset, rime, and F0 cues to the T2–T3 contrast, our study focuses on the role of rime's breathiness. We thus did not orthogonally vary the onset and rime properties. For fricatives onsets, although we know they tend to be produced with phonetic voicing in T3 especially by young speakers (Gao & Hallé, 2017), we nevertheless constructed or recorded phonetically voiceless [f]s and [s]s: Our goal was to examine whether breathy voice could bias the T2–T3 categorization toward T3 in spite of the likely influence of phonetic voicing: voiceless [fe] and [se] are likely more often categorized as *yin* T2 syllables than syllables with voiceless stop onsets (Zhang & Yan, 2018). Finding a breathy-voice bias in such adverse conditions thus would be strong evidence for the bias. In the case of nasal-onset syllables, we had our speaker intentionally produce a clear breathiness distinction, whereby voice quality co-varied in the onset and the

rime. That is, our yang tone T3 /mɛ/ was (clearly) breathy in both /m/[1] and /ɛ/ ([m̤ɛ̤]), although we know it is usually produced with no or little breathy voice in both tones T3 and T2 (Gao, 2015; Zhang & Yan, 2015, 2018). Zhang and Yan (2018) data suggest that "Phonation" does not affect categorization of their /mɛ/ stimuli, which differed in breathiness in the rime but presumably not in the nasal onset itself (given their production data), unlike in our study. The co-variation we introduced should, if anything, enhance the T3 bias in the breathy T2–T3 continuum categorization. Finding no T3 bias with breathy /mɛ/ in our study would thus add support to Zhang and Yan's findings.

Because the stimuli derived from natural T2 or T3 syllables possibly retained non-tonal characteristics of the original syllables other than voice quality, we also constructed fully synthetic stimuli, in which only F0 contour and voice quality were manipulated and other characteristics kept constant. The stimuli were presented to young Shanghai listeners for identification (forced choice among two T2–T3 minimal pair words). If breathy voice still is a cue to T3 identification, the identification functions for breathy- versus modal-voice syllables should differ, with more *yang* T3 responses for breathy than modal voice.

2 Experiment

We ran a 2AFC identification test on onset+/ɛ/ syllables varying along a T2–T3 continuum and produced with either modal or breathy voice. We used both modified natural and fully synthesized stimuli. In each trial, the two choices proposed were two minimal-pair single syllable words in tones T2 and T3 (e.g., 反 [fɛ34] "reverse" and 烦 [fɛ23] "annoying").

If breathiness is a cue to *yang*-ness in perception, we expect that breathy voice will bias responses toward the T3, *yang* choice. However, because the modified natural *yang* stimuli may still retain acoustic characteristics *other* than breathy voice, for example intensity contour (Rose, 1982; Zhu, 1999), the possibly observed bias may not be solely due to breathy voice. The synthetic stimuli, in which only voice quality and F0 contour were manipulated, allow controlling for the possible role of such characteristics. The synthetic stimuli were modeled after real syllables and differed from the natural stimuli in terms of voice gender, tone contour steepness, and durations. The motivation for such variation was to test the robustness of the putative breathy voice effect on tone identification across variable phonetic realizations. Whether such variation is relevant to tone identification remains an empirical issue to which our data may provide at least a partial answer.

2.1 Methods

2.1.1 Participants. Sixteen native speakers of Shanghai Chinese (5 male), aged 18–26 (mean 22), participated in the experiment. All were born in the Shanghai urban area, except one female who was born in Japan and came to Shanghai at the age of one. All had spent most of their lifetime in Shanghai and were living there at the time of the experiment. All spoke Standard Chinese and had learned English. Twelve of the participants were recruited from Tongji University in Shanghai where they majored in English, Japanese, or engineering; they received payment for their participation. The other four participants volunteered to participate in the experiment. No participant reported any hearing or reading disorder. At the time of the experiment, all the participants were naïve to its purpose. They were simply told the study was about word identification in Shanghai Chinese.

2.1.2 Materials. The stimuli were based on six T2–T3 minimal pairs of single-syllable words sharing the [ɛ] rime, such as [pɛ34]–[pɛ23] (Table 1). The pairs of Chinese characters for identification

Table 1. T2–T3 minimal pair monosyllabic words used as endpoints of the T2–T3 continua. ([m] onset syllables were not synthesized; there is no [nɛ34] syllable in Shanghai Chinese.).

	∅	stop	fricative	nasal
T2 (yin)	ɛ 爱 "love"	pɛ 板 "broad" tɛ 胆 "gallbladder"	fɛ 反 "reverse" sɛ 伞 "umbrella"	mɛ 美 "beautiful"
T3 (yang)	ɛ 咸 "salty"	pɛ 办 "handle" tɛ 台 "desk"	fɛ 烦 "annoying" sɛ 馋 "greedy"	mɛ 梅 "plum"

Table 2. VocalTractLab settings used to manipulate voice quality.

	Arytenoid area	Lower rest displacement	Upper rest displacement	Subglottal pressure	Aspiration strength
modal	0 mm^2	0.15 mm	0.10 mm	1000 Pa	−40 dB
breathy	2 mm^2	0.60 mm	0.55 mm	1300 Pa	0 dB

responses were selected based on high and similar subjective frequencies collected from 17 native speakers of Shanghai Chinese who did not participate in the experiment. The [mɛ34]–[mɛ23] pair was used for the natural stimuli only, due to synthesizer difficulty with nasals. There were thus 12 base syllables for the natural stimuli: six were produced in tone T2, with intended modal voice, and six in tone T3, with intended breathy voice. We call these syllables "base syllables" because they served as the starting endpoints of the "natural" modal- and breathy-voice T2–T3 continua. They were produced by the first author, a female native speaker of Shanghai Chinese, well trained in phonetics, aged 26 at the time of the recording. She recorded the speech materials in a soundproof room at University Paris 3, with an AKG C520L headband microphone through an Edirol Audio Interface connected to a laptop computer. The sampling rate was 44.1 kHz, with 16-bit resolution. For the synthesized stimuli, the same syllables (minus [mɛ]) were synthesized with a constant F0 of 120 Hz, with either a modal or breathy voice (see below), thus making 10 base syllables. We first provide more detail on the synthesized stimuli. We then explain how the tone-contour continua were constructed and present acoustic measurements showing that our *yang* tones were breathier than our *yin* tones. Finally, we present naturalness-rating data collected for all the stimuli.

Construction of the synthesized syllables. We used the VocalTractLab 2.1 software (Birkholz, Kröger, & Neuschaefer-Rube, 2011: http://www.vocaltractlab.de/) to synthesize breathy and modal syllables. VocalTractLab is an articulatory speech synthesizer, using a two-mass model triangular glottis, in which each vocal fold is modeled by two clamp-shaped masses opening along the dorso-ventral axis. This model can simulate incomplete closure of the glottis, thus different degrees of breathiness. For modal syllables, we used the default gestural score for modal phonation. For breathy syllables, we used settings entailing incomplete closure of the glottis (larger arytenoid area and "articulator" displacements) and stronger airflow through glottis (subglottal pressure and aspiration strength), as shown in Table 2. These settings, chosen based on prior pilot listening, were kept constant on the entire syllable. They resulted in two different configurations of the two-mass model triangular glottis, with complete versus incomplete closure for modal- versus breathy-voice quality.

Construction of the tone continua. All the base syllables, synthesized or natural, were equalized for mean intensity at 80 dB SPL. Onset and rime durations differ according to tone register (Shen,

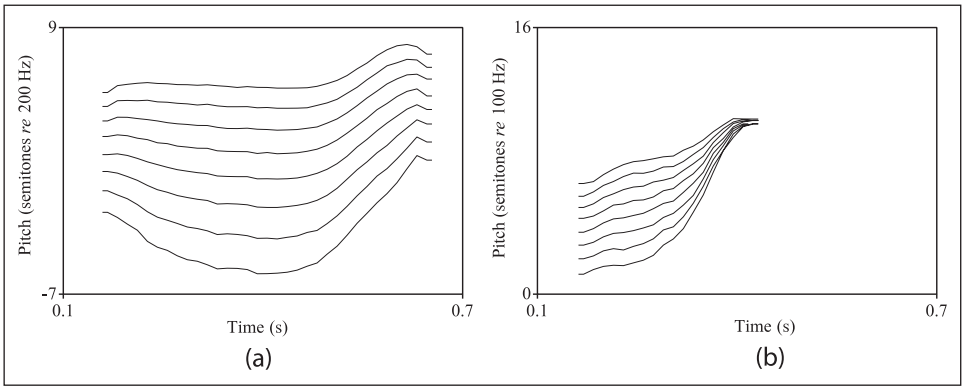

Figure 2. T2–T3 tone continua for the /ɛ/ (zero onset) syllable: (a) natural; (b) synthesized.

Wooters, & Wang, 1987) with longer onsets and slightly shorter vowels for *yin* than *yang* syllables. This difference may bias tone perception (Hallé & Gao, 2016). We therefore decided to keep onset and rime durations constant in all the T2–T3 base syllable pairs. For all synthetic base syllables, rime duration was set to 285 ms; onset duration was set to 175 ms for [s, f]; [ɛ] base syllables (zero onset) were synthesized with no onset. For natural base syllables, [s, f, m] onset syllables were piece-wise time-scaled[2] so that onset and rime durations were 100 and 500 ms, respectively; for the other onsets (Ø and [p, t]), only the [ɛ] rime was time-scaled to 500 ms. These duration values are averages taken from the model syllables that served for synthesized syllables or from the original versions of the natural base syllables. After time-scaling was applied, a continuum of eight F0-equidistant F0 contours between T2 (step 0) and T3 (step 7) was imposed[2] on each base syllable, synthesized or natural. This resulted in breathy- versus modal-voice T2–T3 continua according to whether the base syllable was naturally produced in tone T3 versus T2, that is, presumably, with breathy versus modal voice, or synthesized with breathy versus modal voice. For modified natural stimuli, the endpoint F0-contours of the continua were those, time-scaled, of the original T2 and T3 natural syllables of any given pair (e.g., F0-contours of natural T2 [se34] and T3 [se23] for the [se34]–[se23] continuum). For synthetic stimuli, the endpoint F0-contours were those measured from "model" syllables produced by a typical male speaker (one of those used in Gao & Hallé, 2017). Figure 2 shows the "natural" and "synthetic" continua for the Ø-onset /ɛ/ syllable. As can be seen, the tone contours substantially differ for the two continua. The main difference lies in the final F0 of T3, which is indistinguishable from that of T2 in the synthetic-based continuum (as in Rose, 1993) but not in the natural-based T2–T3 continuum. Taking Figure 1 as a reference, the tone contours for the natural continua are less typical than those for the synthetic continua. Rime-initial F0s, however, follow the same pattern of a clearly lower pitch for *yang* T3 than *yin* T2 for both types of continua, which might be critical in perception. This difference was in average 90 and 39 Hz (7.12 and 5.36 semitones) for natural and synthetic stimuli, respectively.

Stimulus voice quality. We first checked whether the natural as well as the synthesized base syllables did exhibit the targeted voice quality on the stimulus vowel-rime. To assess the intended breathy-voice quality, we used the H1–H2 measure of spectral tilt, which is thought to robustly correlate with breathiness (for a survey of the acoustic correlates of breathiness, see Zhang & Yan, 2018, p. 1294, and references therein). Because all the stimuli had the same rime [ɛ], we did not correct for formant variation (cf. Shue, Keating, Vicenik, & Yu, 2011). Since F0 may influence

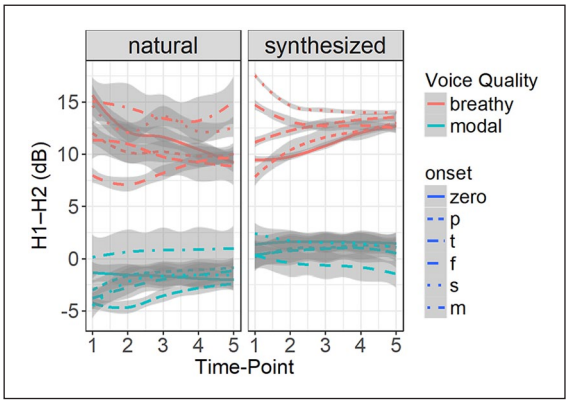

Figure 3. H1–H2 computed on the [ɛ] rime of the stimuli at five time points, according to intended voice quality, stimulus type, and onset. The shaded zones correspond to a 95% confidence interval (variation across continuum steps) and the lines to H1–H2 averaged across continuum steps.

H1–H2 for the same vowel (Chen, Park, Kreiman, & Alwan, 2014; Swerts & Veldhuis, 2001), we measured H1–H2 across the entire continua (i.e., for all F0 contours). We measured H1–H2 in the natural as well as the synthesized base syllables' rime [ɛ], at all eight steps of the stimulus continua, at five equidistant time points from 10% to 90% within the rime. We used a Praat script estimating H1 and H2 amplitudes from short-term spectra with high frequency resolution, whereby H1 and H2 frequency estimation was guided by F0 estimate (Gao, 2015). We found that H1–H2 was consistently larger for T3 than T2 syllables (i.e., voice quality was breathier for T3 than T2) for all the onsets used, for both natural and synthesized stimuli, for all five time points, and for all eight steps of the continua. This is illustrated in Figure 3. We ran mixed-model linear regressions on the H1–H2 data, using the *lmer* and *anova* functions of the *lme4* (Bates, Maechler & Dai, 2018) and *car* (Fox & Weisberg, 2011) packages in R (R Core team, 2016) for the natural and synthesized data, with Onset intercept as a random variable, intended Voice quality (breathy vs. modal), stimulus Type (natural vs. synthesized), continuum Step (1–8), and Time point (1–5) as fixed effects. Voice was significant for both the natural and synthesized, $\chi^2(1) = 4983.60$ and $\chi^2(1) = 6428.82$, respectively, $ps < .0001$, with much larger H1–H2 values for (intended) breathy than modal voice quality ($11.9 > -0.5$ dB), at each step and each time point, $ps < .0001$. Time point was not significant overall. Yet, for natural breathy stimuli, H1–H2 decreased with time toward the end of the rime, $ps < .0001$, as reported in prior production studies (Cao & Maddieson, 1992; Gao & Hallé, 2017). The other variations and their discussion are beyond the scope of this study. We do not discuss them further.

Naturalness ratings. Recall that young speakers of Shanghai Chinese, especially female speakers, tend not to produce breathy voice in *yang* tones (e.g., Gao & Hallé, 2017). Because a young female speaker (the first author) produced the natural base syllables, one may worry about her breathy-voice productions, even though she is a trained phonetician. As we reported, her breathy-voice, *yang* tone T3 syllables were indeed breathy in terms of H1–H2, as intended, compared to her modal-voice, *yin* tone T2 syllables. But how natural did these breathy-voice syllables sound? Stimuli that do not sound natural—that is, sound "odd"—might affect categorization in unpredictable ways, blurring out the expected pattern of results. We therefore checked for this potential issue and collected naturalness ratings for all the stimuli to be used in the experiment. Another

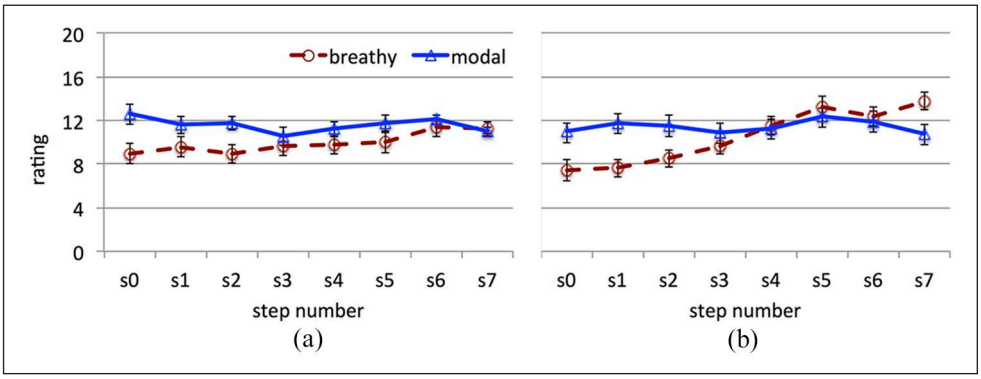

Figure 4. Naturalness ratings for (a) synthesized and (b) natural stimuli, as a function of step and voice quality. Error bars stand for Standard Errors.

concern is the use of pitch-scaling with PSOLA (see footnote 2). The larger the F0 change from the base contour, the larger alteration of stimulus quality, hence, perhaps, the lower rating might be expected. On this view, the naturalness ratings would decrease from step 0 (T2) to 7 (T3) for modal-voice natural stimuli; conversely, they should decrease from step 7 to 0 for breathy-voice natural stimuli. As for the synthetic stimuli, for which the base syllables were synthesized with a flat 120 Hz F0 contour, the naturalness ratings would decrease from step 7 (T3) to 0 (T2) because rime-onset F0 is the lowest and closest to 120 Hz in tone T3 (about 110 Hz) and rime-offset F0 is 180 Hz for both T2 and T3.

The rating data were collected using the "Percy" web-based experimental setting (Draxler, 2014). Listeners were recruited mainly in Shanghai city: 17 subjects (7 males, aged 18–39, mean age 28 y) rated the natural stimuli and 17 others (5 males, aged 19–34, mean age 28 y) rated the synthesized stimuli. None of them participated in the categorization experiment. They listened to each stimulus once and gave it a rating (under no time pressure) using a graphic ruler, whose chosen position was then converted into a 0–20 scale. The instructions were given in Shanghai Chinese. A reasonable assumption is that subjective naturalness correlates with familiarity or frequency of exposure in terms of phonetic-acoustic characteristics—that is, a familiar, frequently heard syllable should be rated high. Figure 4 shows the averaged rating profiles for synthesized and natural stimuli as a function of step. As can be seen, the natural breathy T3 (step 7) stimuli sounded slightly more natural than the natural modal T2 (step 0) stimuli (13.8 vs. 11.0), not *less* natural, as could have happened, had our young speaker produced odd-sounding breathy-voice tone T3 syllables. This finding answers the first concern about our young speaker producing breathy tone T3.

We further explored the rating data by running an analysis of variance, with Subject as the random factor, stimulus Type (natural vs. synthesized) as a between-subject factor, and Step (0–7) and Voice quality (modal vs. breathy) as within-subject factors. (For the sake of simplicity, we only report p values in the following.) The two Types of stimuli exhibited similar profiles: Type was not significant, overall as well as for either modal- or breathy-voice stimuli, Fs < 1, and did not interact with Voice. Voice was significant overall with higher ratings for modal- than breathy-voice stimuli, $p < .05$, and interacted with Step: The modal voice advantage was significant in the first half (steps 0–3) of the continua (ps $< .05$) but not in the second half in steps 4–7 (ps $> .13$). Step was significant for breathy- but not for modal-voice stimuli (ps $> .4$). For modal-voice stimuli, the rating profile was indeed essentially flat, with a weak dip at mid-continuum (see Hallé, Chang, & Best, 2004, for a similar observation: mid-continuum stimuli receive lower ratings, presumably

due to intrinsic ambiguity between endpoints). In contrast, for breathy-voice stimuli, there was a significant Step effect for both the synthesized and natural types (natural: $p < .0001$; synthesized: $p < .05$): rating correlated positively with step number, with lower ratings for the T2 than the T3 endpoint of the continua. This trend was more marked for natural than synthesized stimuli, as indicated by the significant Type × Step interaction in the breathy-voice stimulus rating data, $p < .0001$. These rating data thus do not follow the patterns that could be expected on the assumption that F0 contour modifications consistently alter stimulus quality. This alleviates the second concern one might have with F0 modification using PSOLA (also see Hallé et al., 2004).

The patterns we found rather suggest that breathy-voice stimuli fared better in terms of naturalness in the low than high pitch range of the continua, whereas ratings for modal-voice stimuli were insensitive to pitch range. The steady rating profile for modal-voice syllables and the tilted profile for breathy-voice syllables probably show that only high pitch breathy-voice syllables sound less natural than the others. Indeed, the breathy-voice syllables our young listeners may hear in their environment, presumably in the speech of older speakers of Shanghai Wu, usually carry a low tone.

2.1.3 Procedure. The experiment was conducted using the E-Prime 2.0 software. Participants were tested individually in a quiet room, in front of a laptop. They were presented with the stimuli through professional quality headphones. Each of the $96 = 12 \times 8$ natural and $80 = 10 \times 8$ synthesized stimuli was presented twice during the test phase. Each test phase trial consisted of the following events: (a) a fixation cross appeared at the center of the screen; (b) 500 ms after trial onset, the auditory stimulus was presented; (c) at stimulus offset, the fixation cross disappeared and was replaced with two Chinese characters on the left and right side of the screen, representing the two possible responses to the trial. The T2 and T3 response sides were counterbalanced across the two repetitions of each stimulus so that each character appeared the same number of times at either side of the screen. On each trial, participants were asked to indicate the Chinese character whose reading best matched the auditory stimulus by pressing one of the two labeled keys on the left and right of the keyboard, as quickly and accurately as possible. Response times were measured from the onset of the display of the two response choices (exactly 150 ms after audio stimulus acoustic offset). The response time-out was set to 2.5 seconds. Stimuli were blocked by construction type: half of the participants were tested on the natural stimuli first (192 trials), and the other half on the synthesized stimuli first (160 trials). Listeners took a short break between the two blocks. Within each block, the stimuli were presented in a different randomized order for each participant. The test phase was preceded by a training phase of six trials, with syllable stimuli bearing unambiguous *yin* or *yang* tones. Participants received accuracy and response time feedback during the training but not during the test phase. The instructions to the participants were given orally in Shanghainese.

After the experiment, which lasted about 40 minutes, participants were asked to comment on which block (i.e., natural or synthesized stimuli) was easier and/or more natural.

2.2 Results

The average rate of missing responses (no response before time-out) was 0.53% (0.62 and 0.43% for natural and synthetic stimuli, respectively). One 19-year-old male participant, S02, was an outlier in this respect with 6.8% missing responses for natural stimuli (although only 1.25% for synthetic stimuli). Moreover, for the synthesized stimuli, S02 perceived as *yin* tone T2 syllables most of the stimuli in the modal *and* breathy T2–T3 continua for fricative onsets: *all* the [fɛ] stimuli as well as all the [sɛ] stimuli but the T3 endpoint. Because fricative onsets, especially labials, are often *phonetically voiced* in *yang* syllables, as mentioned earlier, a general trend for phonetically

voiceless fricative onset syllables to be perceived as *yin* syllables is not surprising—and all the syllables constructed for the T2–T3 [fɛ] and [sɛ] continua *were* phonetically voiceless. However, S02 was the only participant with such extreme data for synthesized [fɛ, sɛ] and a relatively high rate of missing responses for natural stimuli. We therefore discarded the entire data of S02.

Except for S02, we retained all the available binary response data (subsumed by rate of *yang*, T3 responses) as well as the available RT data for both *yin* (T2) and *yang* (T3) responses, provided sufficient response rate; this condition was met for *yin* responses in the first half (steps 0–3) of the continua and for *yang* responses in the second half (steps 4–7), as the identification data will make clear: *yin* (T2) responses were largely dominant in the half beginning with the T2 endpoint and *yang* (T3) responses in the half ending with the T3 endpoint.

The data from a few participants is qualitatively informative. For the natural stimuli, one 21-year-old male participant, S14, perceived as *yang* tone T3 syllables about 80% of the stimuli in the breathy T2–T3 continua: *all* the [sɛ] and [mɛ] stimuli, as well as all the [ɛ] stimuli except the T2 endpoint. Two female participants, S05 and S15 (aged 22 and 24 years) identified *all* the stimuli of the breathy [ɛ] continuum as *yang* tone T3 [ɛ]. Finally, one female participant, S09 (aged 19 years) identified all the stimuli of the breathy [sɛ] continuum, except the T2 endpoint, as *yang* tone T3 [sɛ]. That is, a few participants relied *exclusively* or almost exclusively on voice quality to identify some T2–T3 minimal pair syllables, with the notably extreme case of S14, who identified three out of six breathy natural stimulus continua in this way. As pointed out by an anonymous reviewer, this bias was possibly due to a ceiling effect: the breathy–modal differentials being quite large in our stimuli, breathiness outweighed F0 as a cue to *yang* tone for these participants. The other participants' reliance on voice quality to identify syllables was far from that extreme. Note that there was no symmetrical bias for the natural stimuli to be responded *yin* (T2) based on the sole modal voice quality.

As for the post-test comments, participants overall found the identification task relatively easy to complete. Among the 15 participants whose data were retained, 10 reported that the natural stimuli were easier to identify, three reported that the synthesized stimuli were easier to identify, and two gave unclear responses. (They were not told which stimuli were "natural" and which were "synthesized": they only reported which block—first or second—they found easier.)

2.2.1 Identification data. Figure 5A–B shows the identification curves (*yang* response rate × continuum step), averaged across the [Ø, p, t, f, s] onsets and participants, according to stimulus construction type and voice quality. These curves are typical of categorical identification, with a steep slope at the 50% boundary. For both natural and synthesized stimuli, a shift toward the *yin* endpoint (T2) can be seen for breathy relative to modal stimuli, suggesting a bias towards *yang* responses for breathy stimuli. Individual identification curves show the same tendency for all participants. Identification curves for each onset also show the same shift, except for [m] (Figure 5C).

One straightforward way to substantiate these observations is to compare the overall *yang* response rates in the modal and breathy conditions: is there a significant bias toward more *yang* responses in the breathy than modal condition? In principle, probit analysis can also be used to directly estimate this bias in terms of categorical boundary shift. Yet, probit analyses, whereby cumulative Gaussians are fitted to the individual raw identification data, require that the identification data cross the 50% level (Best & Strange, 1992; Hallé, Best, & Levitt, 1999; Hallé, Chang, & Best, 2004). As we noted earlier, this requirement was not always met for four subjects who identified entire continua of breathy-voice stimuli as *yang* tone T3. Because the opposite pattern was not observed in the participants we retained, probit analyses thus would unilaterally exclude a substantial part of the intercept data that could point to breathy voice biasing identification toward *yang* responses. We therefore first focused on the *yang* response rate analysis.

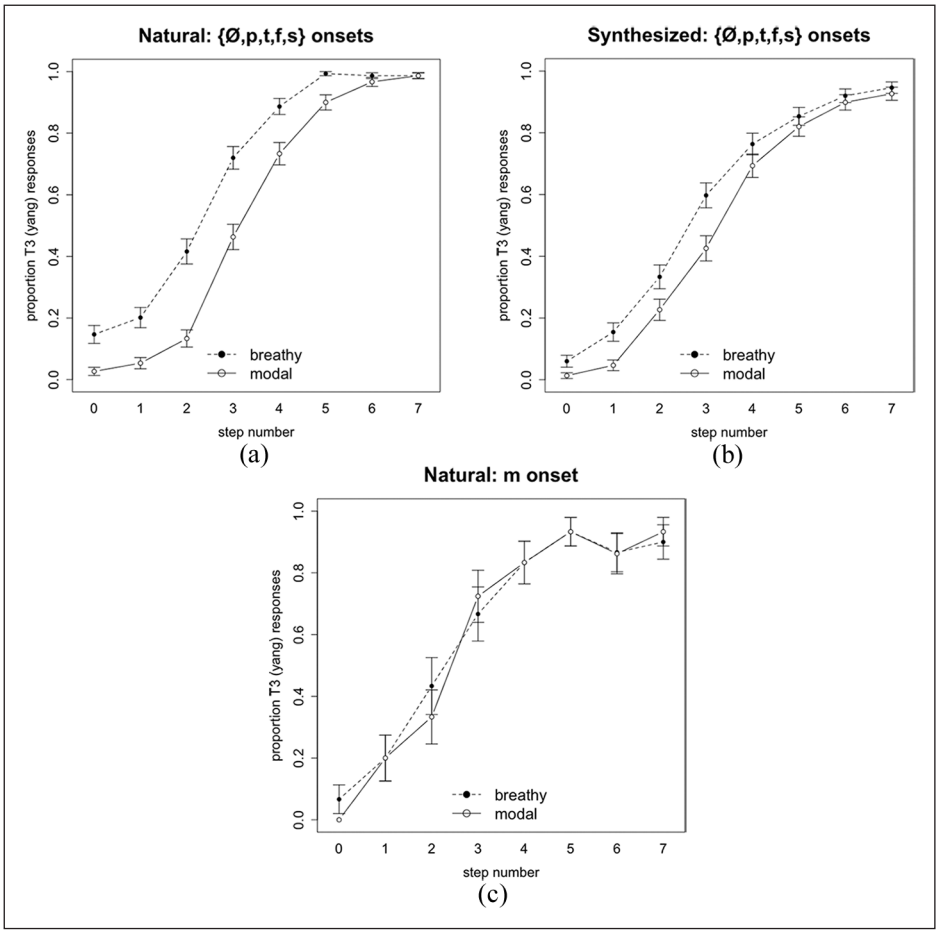

Figure 5. *yang* response rate according to voice quality and continuum step, averaged across subjects and onsets: (a) natural and (b) synthetic non-[m] onset stimuli; (c) [m]-onset natural stimuli.

In the natural stimulus data, the overall rate of *yang* responses was 64.5% for breathy versus 53.4% for modal stimuli. In the synthesized stimulus data, these rates were 57.8% for breathy versus 50.6% for modal stimuli. Table 3 shows these rates detailed by onset.

A series of log-odds regression generalized linear mixed-models (GLMM) were fitted to the binomial response data, using the *lme4* package (Bates et al., 2018) in R. *Yang* response was dummy-coded as 1 and *yin* response as 0. The following predictors were included as factors in the initial model: stimulus Type (natural vs. synthetic), block Order (natural vs. synthetic block first), Repetition (first vs. second occurrence), Voice (modal vs. breathy), Onset (∅, p, t, f, s, m), and stimulus Step (as an ordered factor from step 0 to step 7, the T2 and T3 endpoints). The reference levels for the above predictors were: natural, natural block first, first occurrence, breathy voice, ∅ onset, and step 0. The three-way interaction Type × Voice × Onset improved the model fitting as estimated by likelihood ratio tests using the *anova* function, $\chi^2(8) = 186.62$, $p < .0001$. (The interaction between Voice and the other factors, did not improve the model, suggesting similar effects of Voice in different conditions involving Order, Repetition, or Step.) We did not include

Table 3. *Yang* response rate (%) averaged across continuum steps and subjects, according to voice quality and onset, for natural and synthesized stimuli. Significance levels (Δ columns): * for $p < .05$, ** for $p < .01$, *** for $p < .001$.

Onset	Natural			Synthesized			Average		
	Modal	Breathy	Δ	Modal	Breathy	Δ	Modal	Breathy	Δ
Ø	50.4	67.5	***17.1	69.6	75.4	*5.8	60.0	71.5	***11.5
p	44.2	62.5	***18.3	51.7	60.8	**9.1	48.0	61.7	***13.7
t	54.6	62.9	**8.3	52.1	59.6	**7.5	53.4	61.3	***7.9
f	48.8	56.7	***7.9	24.2	32.1	**7.9	36.5	44.4	***7.9
s	59.6	72.1	***12.5	55.4	61.3	*5.9	57.5	66.7	***9.2
Means	51.5	64.3	***12.8	50.6	57.8	***7.2	51.1	61.1	***10.0
m	58.3	57.5	−0.8	—	—	—	—	—	—

Table 4. Results of post-hoc pairwise comparisons for the breathy–modal contrast, by onset and stimulus type. Positive coefficients indicate higher likelihood of *yang* responses.

onset	natural				synthetic			
	Estimate	*SE*	*z*	*p*	Estimate	*SE*	*z*	*P*
∅	2.00	0.33	5.98	< .0001	0.66	0.34	1.96	0.050
p	2.11	0.33	6.43	< .0001	1.02	0.33	3.13	0.002
t	0.86	0.33	2.61	0.009	0.82	0.33	2.50	0.013
f	0.94	0.32	2.90	0.004	0.76	0.31	2.47	0.014
s	1.72	0.34	5.03	< .0001	0.67	0.33	2.04	0.041
m	0.13	0.33	0.40	0.687				

Gender as a fixed effect because the number of female versus male participants was far too much imbalanced: 11 versus 4. (Numerically, however, the rate of *yang* responses was very similar for female versus male participants: 0.569 vs. 0.576.) We retained intercept for subjects as a random effect, and added slope on Voice as a random effect which improved the model fitting, $\chi^2(2) = 21.66, p < .0001$.

According to the result of the selected GLMM model, the main effect of interest, Voice, was found significant (*Estimate* = -2.00, *SE* = 0.33, *z* = -5.98, $p < .0001$), suggesting higher *yang* (T3) response probability for the breathy than modal condition. On average, the T3 response rate was 0.623 for breathy condition and 0.519 for modal condition. Repetition was also significant (*Estimate* = 0.37, *SE* = 0.09, *z* = 4.06, $p < .0001$), with more *yang* responses in the second than the first repetition (0.587 > 0.555). Marginal effects were found for Type (*Estimate* = 0.56, *SE* = 0.32, *z* = 1.75, *p* = 0.08) and Order (*Estimate* = -0.59, *SE* = 0.29, *z* = -1.94, *p* = .0528). However, Voice interacted with Type (*Estimate* = 1.36, *SE* = 0.44, *z* = 3.02, *p* = 0.003), suggesting a lesser Voice effect (weaker breathy-voice bias) in the synthesized compared to natural stimuli. Step was obviously significant (linear function: *Estimate* = 7.87, *SE* = 0.23, *z* = 4.06, $p < .0001$), indicating higher *yang* response probability when step increases from 0 to 7, that is, from T2 endpoint to T3 endpoint.

Post-hoc pairwise comparisons were made on the above GLMM model, using the *emmeans* package (Lenth, 2018) in R, for each onset under each stimulus type (Table 4). When broken down by onset, the Voice effect was significant for each of the five onsets under scrutiny, for either

Table 5. Average intercepts (steps 0–7) according to voice quality, onset, and stimulus type. Means exclude the [m] onset. Shaded Δ cells indicate non-significant breathy-modal differences.

Onset	Natural			Synthesized			Average		
	Modal	Breathy	Δ	Modal	Breathy	Δ	Modal	Breathy	Δ
Ø	3.68	2.68	1.00	1.91	2.04	−0.13	2.80	2.36	0.43
p	3.59	2.73	0.87	3.14	2.46	0.68	3.37	2.59	0.77
t	2.91	2.64	0.27	3.23	2.64	0.59	3.07	2.64	0.43
f	3.90	3.23	0.67	5.40	4.90	0.50	4.65	4.06	0.59
s	2.55	1.82	0.73	3.37	2.82	0.55	2.96	2.32	0.64
Means	3.33	2.62	0.71	3.41	2.97	0.44	3.37	2.80	0.57
m	2.55	2.77	−0.23	—	—	—	—	—	—

natural or synthesized stimuli, with only one exception: for [m] onset (natural stimuli), Voice had no effect.

To summarize the overall *yang* response rate data, Voice was significant for both stimulus types (more *yang* T3 responses for breathy- than modal-voice stimuli), whether constructed from natural or from synthesized base stimuli. The only exception lies in the [m] onset continua, for which listeners' T2–T3 categorization was not biased by stimulus voice quality.

We also conducted a probit analysis, fitting short ogive cumulative Gaussians (cf. Best & Strange, 1992; Hallé et al., 1999, 2004) to *yang* identification data where possible. This analysis should confirm the finding of breathy voice biasing responses toward *yang* identification, using a direct estimation of the modal-to-breathy shift observed in the identification functions (Figure 5). We had to discard the data of the four participants discussed above, whose *yang* identification data were above the 50% level for some breathy-voice continua, making impossible the estimation of crossover location (intercept). Table 5 shows the intercept data detailed by onset, as estimated by fitting short ogive Gaussians to three raw data points encompassing the 50% identification crossover for each subject in each condition.[3] As can be seen, the intercept data largely paralleled the *yang* identification rate data, although we excluded from the analysis the four subjects that exhibited a very strong *yang* bias for breathy-voice stimuli. Indeed, the intercepts were generally located earlier in the continua for breathy than modal voice, reflecting the *yang* bias for breathy-voice stimuli, with the notable exception of /mɛ/. The category boundary shift can be estimated to 0.57 versus 0.71 steps in average for synthesized versus natural stimulus continua.

We conducted an analysis of variance on the intercept data for the five onsets common to natural and synthesized stimuli, that is, excluding [m], with Subject as the random factor, Voice (modal vs. breathy), Onset (∅, p, t, f, s), and stimulus Type (natural vs. synthesized) as within-subject factors. Voice was significant, $F(1, 10) = 50.70$, $p < .0001$, reflecting smaller intercept values for breathy than modal continua ($2.80 < 3.37$), that is, categorical boundaries closer to *yin* endpoint. The trend toward a Voice × Type interaction did not reach significance, $F(1, 10) = 3.16$, $p = .107$. Voice was indeed significant for both natural, $F(1, 10) = 26.62$, and synthetic stimuli, $F(1, 10) = 32.32$, both ps $< .0005$. Onset was significant, $F(4, 40) = 28.22$, $p < .0001$, indicating variation of the intercept values across onsets (see Table 5). Onset interacted with Type, $F(4, 40) = 18.08$, $p < .0001$, indicating that intercept variation differed in natural and synthetic stimuli. (We return to the intercept variations across onsets in the discussion.) The Voice × Onset interaction was not significant, $F(4, 40) < 1$, indicating that the effect of Voice was overall stable across onsets other than [m].

Voice was indeed significant for each onset (data pooled across Type) at least at the $p < .01$ level, except [s], for which Voice was marginally significant, $p = .06$. (More variation was found in each Type taken separately, as shown in Table 5.) The intercept data for onset [m] (natural stimuli) were analyzed separately: the intercepts for breathy and modal voice continua did not differ (2.77 vs. 2.55), $t(10) < 1$.

Slopes (more precisely, the fitted Gaussians' standard deviations, inversely proportional to the identification function slopes at category boundary) did not differ in the breathy versus modal conditions (natural: 0.33 vs. 0.34; synthesized: 0.33 vs. 0.31; ts < 1). We do not discuss them further.

Therefore, in spite of their lesser statistical power due to discarding the data of four subjects, the probit analysis intercept data also showed that breathy voice biased responses toward the *yang* response, in a perhaps more direct way than the overall *yang* response rate data.

2.2.2 Response time data. The identification data of the natural stimuli show that the proportion of *yin* responses fell below 0.15 (breathy voice) or 0.32 (modal voice) after step 3. Likewise, the proportion of *yang* responses fell below 0.39 (breathy) or 0.13 (modal) before step 3. A similar pattern obtained for the synthetic stimuli. We thus restricted RT data analyses to the portions of the continua in which the proportion of responses was substantial: steps 0–3 for *yin* (T2) responses and steps 4–7 for *yang* (T3) responses. The *yin* response RT data in steps 4–7, as well as the *yang* response RT data in steps 0–3 were taken as not reflecting a sufficient number of observations and were not examined.

Figure 6A–B shows the identification response times (RT) according to voice quality and stimulus construction type, averaged across participants and the [Ø, p, t, f, s] onsets. Figure 6C shows RTs for the [m] onset natural stimuli. Only *yin* response RTs in steps 0–3 and *yang* response RTs in steps 4–7 are shown. For the natural stimuli, *yin* response RTs in steps 0–3 were shorter for modal- than breathy-voice stimuli, whereas *yang* response RTs in steps 4–7 were shorter for the breathy- than modal-voice stimuli. That is, *yin* responses were faster with modal than breathy voice and vice versa for *yang* responses in the regions of interest of the continua. The results were less clear-cut for the synthesized stimuli, but the pattern was similar, as summarized in Figure 6A–B. The RT data thus suggest that *yin* tone is congruent with modal voice and *yang* tone with breathy voice.

To substantiate these findings, we analyzed the RT data as follows. For each continuum, we computed (a) the *yin* response RT averaged across steps 0–3 for the modal (*yin*-congruent) versus breathy (*yin*-incongruent) voice quality, and (b) the *yang* response RT averaged across 4–7 for the breathy (*yang*-congruent) versus modal (*yang*-incongruent) voice quality. These RT measures were numerically shorter for congruent than incongruent voice quality as shown in Figure 6D.

We took the log-transformed RT, ln(RT), as the dependent variable (DV) so that the DV distribution was closer to a Gaussian normal distribution. Yet, for the sake of clarity, we report RT values in ms in the following. We ran an analysis of variance on ln(RT), with the following within-subject factors: voice quality Congruence (congruent vs. incongruent), continuum Half (steps 0–3 vs. steps 4–7), Onset (∅, p, t, f, s), stimulus Type (natural vs. synthetic), and Repetition (1–2). Subject was the random factor. Congruence was significant, with shorter RTs for congruent than incongruent stimuli (836 < 891 ms), $F(1, 14) = 30.97$, $p < .0001$, indicating an overall "voice quality" bias. Half was not significant, $F(1, 14) = 1.20$, $p = .29$, nor was Repetition, $F(1, 14) < 1$, and neither of these interacted with Congruence. Type was marginal, $F(1, 14) = 3.31$, $p = .09$, but did not interact with Congruence, $F(1, 14) < 1$, suggesting the voice quality bias was not significantly larger for natural than synthetic stimuli in spite of the observed numerical trend (natural: 68 ms bias; synthetic: 41 ms bias). Congruence was indeed significant for both the natural, $F(1, 14) = 9.84$,

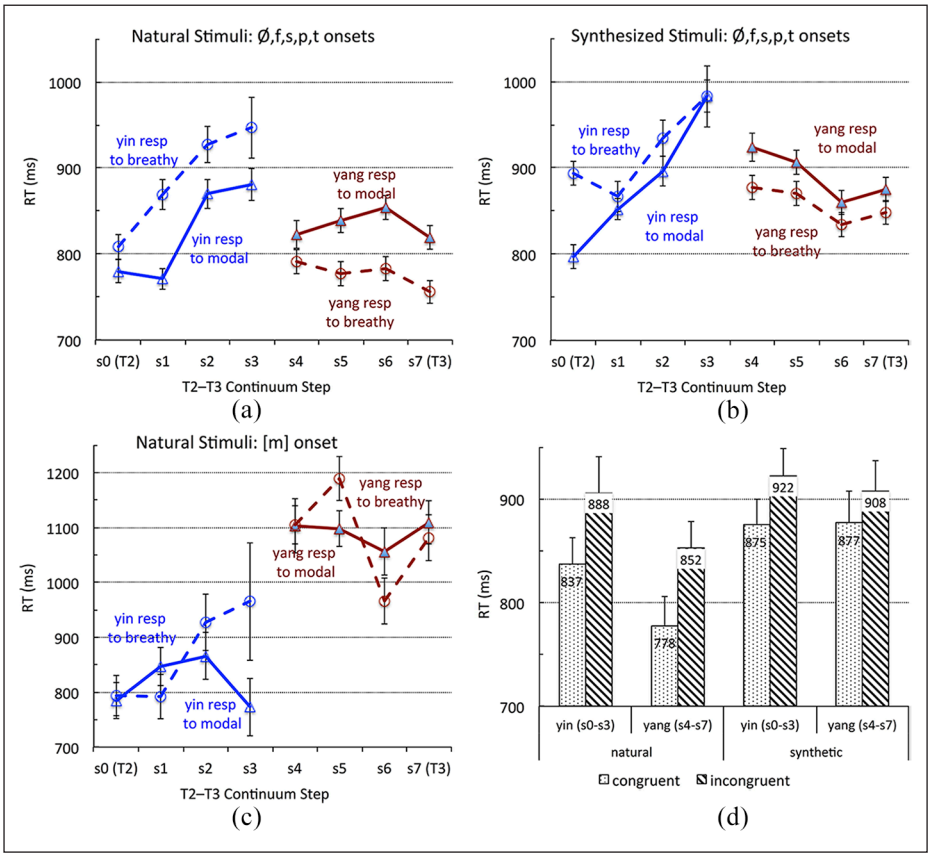

Figure 6. *yin* response RTs in steps 0–3 and *yang* response RTs in steps 4–7, averaged across subjects according to voice quality for (a) natural and (b) synthesized stimuli ([Ø, p, t, f, s] onsets), and (c) for natural [m]-onset stimuli; (d) pooled according to stimulus type and dominant response.

Table 6. Response time data according to Congruence for each of the five onsets Ø, [p], [t], [f], [s], and averaged across them, and for [m], which was used for natural stimuli only.

	Onset						
	Ø	p	t	f	s	mean(Ø-s)	m
Congruent	720	993	886	878	705	836	979
Incongruent	836	1003	940	911	763	891	983
Difference	117	9	54	33	59	54	4
Mean RT	778	998	913	895	734	863	981

$p < .01$, and the synthetic stimuli, $F(1, 14) = 6.72, p < .05$. Onset was significant, $F(4, 56) = 22.86$, $p < .0001$, indicating variation in mean RT value among the five onsets (see Table 6). Onset interacted significantly with Congruence, $F(4, 56) = 2.88$, $p < .05$, indicating the Congruence effect varied across onsets: it was indeed consistently significant for onsets Ø and [t], less so for [f] and

[s], and non-significant across the board for [p]. (We examined the [m] onset separately.) However, as shown in Table 6, the Congruence effect affected RTs in the same way (at least numerically) for all onsets: RTs were shorter in the congruent than non-congruent condition.

We ran a separate analysis, still on ln(RT), for the natural stimuli with the [m] onset, with Congruence, Half, and Repetition as within-subject factors. Because S14 only gave *yang* T3 responses to *all* the breathy [mɛ] stimuli, his *yin* response RTs were missing in the breathy (incongruent) condition: we replaced them with his *yin* response RTs in the modal (congruent) condition, a most conservative solution for testing the Congruence effect. In doing so, the differential between the congruent and incongruent conditions amounted to 4 ms (Table 6), against 1.5 ms when discarding S14's RT data altogether. At any rate, such differentials are negligible. Indeed, for [m], Congruence was found non-significant, $F(1, 14) < 1$. Half was significant, $F(1, 14) = 9.98$, $p < .01$, reflecting longer RTs for *yang* responses (steps 4–7) than *yin* responses (steps 0–3), as can be seen in Figure 6C. Repetition was significant, $F(1, 14) = 5.27$, $p < .05$, reflecting shorter RTs in repetition 2 than 1 (928 < 1039 ms). Yet, no effect of Congruence was found in either repetition 1 or 2, Fs < 1.

To summarize, the RT data show that breathy-voice speeds up *yang* responses, whereas modal-voice speeds up *yin* responses, at least in the continua regions of interest, where these responses are largely dominant.

3 General discussion

In this study, we showed that breathiness does affect Shanghai tone perception in biasing listeners' identification towards the *yang* category. First, there were more *yang* T3 responses for breathy- than modal-voice stimuli for the *same* T2–T3 tone-contour continua, resulting in a clear shift in categorical boundary for breathy- relative to modal-voice T2–T3 continua. The shift was estimated to about 0.71 versus 0.57 steps in average for the natural versus synthetic continua, corresponding to a 0.72 versus 0.46 semitones pitch difference at tone contour onset, respectively. Second, the *yang* bias was supported by the RT data, with generally faster RTs when voice quality was congruent with the dominant tone response: in steps 0–3, where *yin* responses were dominant, RTs were faster for modal- than breathy-voice stimuli; in steps 4–7, where *yang* responses were dominant, RTs were faster for breathy- than modal-voice stimuli. To summarize, in spite of the current trend in the young generation of Shanghai native speakers to produce little or no breathiness with *yang* tones (Gao, 2016; Gao & Hallé, 2017; Tian & Kuang, 2016; Zhang & Yan, 2018), breathiness still plays an important role in perception. The participants of the current study and the native speakers in the cited production studies belonged to the same age group of the same population, thus the perception–production discrepancy is virtually observed in one person: the young Shanghai Chinese speaker. This speaker is clearly biased toward categorizing breathy voice syllables as *yang* syllables. That is, his/her perception of tone exhibits a trading relationship between pitch and breathiness, at least for the T2–T3 contrast.

3.1 Natural versus synthesized stimuli

It might be objected that the "breathiness effect" found in the continua constructed from natural stimuli was due to the breathy stimuli continua retaining cues to *yang* tone T3 other than breathiness: in particular, intensity profile and, perhaps, phonetic characteristics of the onset. (Note that one potentially influential cue—relative onset and rime durations—was neutralized: see 2.1.2.) However, essentially the same pattern of breathiness effect obtained for the synthesized stimuli, in which all the other-than-voice-quality cues were, by construction, completely neutralized. We may

thus safely conclude that these other cues did not play a substantial role in the natural stimuli, as suggested in a previous study (Gao & Hallé, 2013). That is, the breathiness effect obtained with natural stimuli is not due to unwanted confounds. Yet, the breathy–modal differentials were larger for natural than synthetic stimuli in terms of *yang* response rate (0.128 vs. 0.073), for intercept (0.71 vs. 0.57 steps) and response times (68 vs. 41 ms). Even though this advantage in breathiness effect for natural over synthetic stimulus continua might still be attributed to the presence of non-neutralized cues in natural stimuli, the crucial observation is that the breathiness effect obtains in the absence of such cues.

That the breathiness effect was observed for both natural and synthesized stimuli also shows that the categorization of these continua was not much influenced by substantial differences between the natural and synthesized continua in terms of F0 contour shape (cf. Figure 2A–B) as well as in onset–rime duration patterns (see 2.1.2). In spite of these differences, the endpoints of both natural and synthetic continua were consistently identified as T2 (step 0) and T3 (step 7), and the categorical boundary was shifted towards T2 when voice quality was switched from modal to breathy. We may only note a slight, non-significant trend for a later categorical boundary for synthesized than natural continua (3.08 vs. 2.97 steps), $F(1, 10) = 2.64$, $p = 0.135$. This trend, however weak, is in line with the smaller rime-to-onset duration ratio for synthesized than natural stimuli, that is, a timing pattern pointing to lesser voicedness.

Natural and synthesized stimuli also differed in a somewhat larger variability across the five onsets [Ø, p, t, f, s] of the breathiness effects induced by natural than synthesized stimuli as shown by the standard deviations across onsets (*yang* response rate: 5.0% vs. 1.4%; intercept: 0.28 vs. 0.32 steps; RT: 62.6 vs. 35.6 ms). Also, in the *yang* rate data, Onset interacted with Voice only for natural, not for synthesized stimuli. In the following section, we shall discuss the breathiness effect detailed by stimulus onset.

3.2 Variation across onsets

The breathy-voice effect varied according to syllable onset, as consistently shown across all the statistical analyses of *yang* response rate, intercept, and response time. We will discuss here the onsets that departed from the average trends reported above, and/or differed according to the type of stimulus (natural vs. synthesized). Onset [m] departed the most from average: no breathy-voice effect whatsoever was observed in all the three diagnostic measures we ran (*yang* response rate, intercept, RT). We will discuss the [m] onset the last because this discussion involves many different linguistic aspects—including diachronic ones—as well as non-linguistic aspects. For the moment, we focus on Ø and [f], which are far less complicated cases than [m].

The zero onset did not depart from the average "onset" in the natural stimuli, but it did in the synthesized stimulus data: the *yang* response rate and intercept data were very close to average for the natural but not the synthesized stimuli. For the latter type, *yang* response rate was higher than the average (72.5 vs. 54.2%). Likewise, the intercepts for both modal and breathy voice were closer to the *yin*, T2 endpoint than for other onsets (1.98 vs. 3.50 steps). We believe that more research is needed for this particular onset, which may be produced as a glottal closure in *yin* syllables but as a weak voiced friction for *yang* syllables (Chen & Gussenhoven, 2015). We, however, synthesized a truly zero onset, yielding plain [ɛ] syllables. The natural stimuli were probably produced closer to listeners' expectations and indeed yielded *yang* response rates similar to those for the other onsets: natural *yang* /ɛ/ was produced with some leading aspiration, whereas modal, *yin* /ɛ/ was produced with an abrupt vowel onset, hence somewhat akin to glottal stop.

Another onset that deviated from the general trend is [f], in the synthesized but not the natural stimulus data. For this onset and for synthesized stimuli, the *yang* response rate was lower than the

average (28.1 vs. 54.2%) and the intercepts for both modal and breathy voice were farther from the *yin* T2 endpoint than for other onsets (5.15 vs. 2.70 steps). This pattern likely reflects the difference in production, at least in the young generation of Shanghai Chinese speakers, between the labial fricatives and the other onsets. In word-initial position in *yang* tone syllables, while stop onsets are almost always phonetically voiceless, fricatives—especially labial ones—are often phonetically voiced. Gao and Hallé (2017) measured voicing-ratio (proportion of voicing during consonant) for fricatives, and reported, on average, higher voicing-ratio for fricatives in *yang* tones than *yin* tones ($0.41 > 0.08$), as well as higher voicing-ratio for labial than dental fricatives in *yang* tone ($0.61 > 0.25$). Zhang and Yan (2018) reported that 100% of word-initial labial fricatives in T3 syllables (in a carrier sentence) were classified as phonetically voiced, while this number was only 50% for dental counterparts. For young Shanghai listeners, phonetic voicing thus must be a cue to *yang* tone for syllables with a labial fricative onset (Gao & Hallé, 2013; Zhang & Yan, 2018). Since both *yin* and *yang* tone labial fricative onsets were synthesized as voiceless [f] in our experiment, the young Shanghai participants probably tended to judge both the *yin* and *yang* tone labial fricative onset syllables as *yin* tone syllables. (One participant, S02, whose data were not retained, even perceived the entire modal and breathy *yin-yang* [fɛ34]–[fɛ23] continua as *yin* tone [fɛ].) To sum up, the data for synthesized stimuli with a labial fricative onset suggest that participants expected *yang* tone /vɛ/ to be produced with a phonetically voiced [v]. Since they were synthesized with voiceless [f], participants were strongly biased to perceive them as *yin* tone /fɛ/. In contrast, the natural *yang* tone /vɛ/ was perhaps produced with some trace of voicing[4] although the speaker strived to pronounce a voiceless [fɛ], thereby explaining the absence of deviation from the general trend for natural [fɛ].

Finally, a major departure from the breathiness effect found with most onsets was the special case of onset [m] in the natural stimulus data. Virtually no breathiness effect was found with [m]. This result was also found by Zhang and Yan (2018). It is all the more robust that, in our stimuli, both the [m] onset itself and the rime were produced with more breathiness in *yang* tone T3 than *yin* T2. The special pattern obtained with [m] and, perhaps, with any nasal onset, might be motivated by two reasons. The first reason is based on acoustic–auditory constraints. The effect of breathy voice on vowels is similar to that of nasalization: higher formants (especially F1), increased bandwidth of the formants, lower amplitude, presence of anti-resonances in the spectrum (Ohala, 1975), leading to possible confusion between nasality and breathy voice. Perception experiments indeed showed that breathy voice sound nasal to listeners in synthetic stimuli (Arai, 2006). Similarly, phonetically trained listeners perceive oral vowels adjacent to voiceless fricatives as nasalized vowels in synthetic speech, in English or Spanish (Ohala & Amador, 1981), as well as in Hindi (Ohala & Ohala, 1992). Therefore, breathy voice misinterpreted as nasalization might not contribute to the perception of a *breathy* low tone. The misperception may also occur in the opposite direction: listeners may misperceive nasal consonants as being breathier than oral consonants (Garellek, Ritchart, & Kuang, 2016). One way or the other, as suggested by Berkson (2016, 2019) based on production and perception data in Marathi, "phonation contrasts in sonorants [including nasals] are 'vulnerable' in a way that those in obstruents are not" (Berkson, 2016, p. 9). The second reason has a functional basis. The distinctive function of breathiness must be quite weak for nasal onsets because a large majority of them co-occur with *yang* tones. Taking the [ɛ] rime in the first syllable of a word as an example, the "Shanghai dialect Dictionary" (Xu & Tao, 1997) contains 12 entries for [mɛ] with *yin* tones (6 with T1 and 6 with T2), compared with 38 entries for *yang* T3; 1 entry for [nɛ] with *yin* T1, compared with 25 entries for T3; no entry for [ŋɛ] with *yin* tones, compared with 43 with T3. The clear dominance of *yang* tone syllables also holds for liquid-onset syllables: 1 entry for [lɛ] with *yin* T1, compared with 51 entries for [lɛ] with *yang* T3. (Since the "Shanghai dialect Dictionary" somewhat departs from the present-day urban Shanghai Chinese,

we applied a few corrections in our counts: we counted as /ɛ/ rimes the Dictionary /e/ rimes that merged with /ɛ/, but not those that changed to /ø/.) The counts are much more balanced for the obstruent onsets we used in the present study; for example, there are 25 entries for [tɛ] in *yin* tones (18 with T1 and 7 with T2), compared to 30 entries for [tɛ] in *yang* T3. To sum up, breathiness can only be weakly informative as to the *yin* versus *yang* identity of nasal-onset syllables (and perhaps, more generally, sonorant-onset syllables) compared to obstruent-onset syllables. This could account for the absence of a breathiness effect with the [m] onset in our data. Another possible explanation would be that, unlike *yang* syllables with obstruent onsets, *yang* syllables with nasal onsets were never produced with breathiness since Middle Chinese. Some linguists, however, transcribe nasal onsets in *yang* tone syllables with an [ɦ] segment or diacritic (e.g., [nɦ], Xu & Tang, 1988, p. 7). Although this might only be a transcriptional convention, the phonetic reality of breathiness has also been reported by Rose (1989, 2002), who claimed to perceive breathiness in nasal onsets in several Wu dialects. Worthy of note, however, breathy voice is very weak in the production of nasal onsets (Gao & Hallé, 2017; Zhang & Yan, 2015, 2018), suggesting that the nasal onset is "leading" the disuse of breathiness in low tone syllables. In Shanghai Chinese, the loss of *yang* tone breathiness is mostly achieved with nasal onset syllables in perception and in production. It is not unknown that nasal onset syllables are among the first to be affected by the loss of a voice quality difference in production. In another Chinese dialect group, Yue dialects spoken in Guangdong and Guangxi, breathiness also accompanied low tones. The production data of eight Yue dialects examined by Tsuji (1977), illustrate different stages in the loss of breathiness for low tone syllables. The Rongxian dialect preserved low tone breathiness in syllables with all onset types, whereas the Cangwu dialect lost low tone breathiness for all onset types. Between these two extremes, which we may view as an initial and final stage, respectively, some Yue dialects, including Cenxi, preserve low tone breathiness except for nasal and liquid onsets, illustrating an intermediate stage, presumably similar to the pattern found in young Shanghai Chinese speakers.

3.3 Proposed accounts of the breathiness effect

We began this study by asking whether young native speakers of Shanghai Chinese, who reportedly tend not to produce breathiness in *yang* tones, as older speakers still do, nevertheless exploit the breathy-voice cue to categorize T2–T3 continua. And we found that they do. This pattern seems puzzling in the light of the widely accepted scenario of phonetic change in many languages. According to Ohala (1993), most phonetic changes originate in misperception of the produced speech. Misperception would induce a change in phonetic parsing, which, however subtle, may lead to a change in pronunciation. What we observe here runs in the opposite direction. A plausible account is that young speakers of Shanghai Chinese, especially female speakers (Gao, 2016) are influenced by the increasingly dominant variety of Chinese spoken throughout China, especially in urban areas: Standard Mandarin Chinese, that is, *putonghua* (普通话). In *putonghua*, there is indeed no breathy-voice cue to low tones. Young speakers of Shanghai Chinese are very much exposed to *putonghua* and use it quite often in their daily life. In this case, the classic account proposed by Ohala (1981, 1993) as a general, default scenario does not seem to apply. Young speakers are on the way to losing breathiness in production but still use it in perception. This is perhaps because the variety of Shanghai Chinese they are most often exposed to is that spoken by the speakers of the older generation (e.g., their own parents), who still produce breathiness as a redundant cue to *yang* tones T3 and T5. As a result, their phonetic parsing of Shanghai Chinese tones still integrates the breathiness cue.

In other words, breathiness still is a component of the phonetic make-up of *yang* tones in young speakers' perception and we may understand the breathiness effect in our data as a case of

trading relation: a rise in F0-contour can be "offset" by a change from modal to breathy voice (or the other way round) so that "phonetic quality" is preserved (Parker, Diehl, & Kluender, 1986). Breathiness and F0-contour register are trading. Note that our data are neutral with respect to the claim made by earlier research that trading relations reflect the output of a special phonetic mode of speech perception (see Repp, 1982) as opposed to general auditory effects. Both views have received experimental support (pros: Best, Morrongiello, & Robson, 1981; cons: Parker et al., 1986) and this debate, however fascinating, is clearly out of the scope of our study. We simply observe that young Shanghai Wu listeners integrate *trading* F0-contour and breathiness cues in tone perception.

Another explanation of the breathiness effect observed across the board, with the exception of the natural [mɛ] stimuli, is that breathiness, or more precisely, falling spectral tilt (maximum amplitude in the lowest harmonics), biases perception toward a low pitch, as a general auditory phenomenon. Indeed, recent studies by Kuang and Liberman (2015) suggest that falling spectral tilt biases F0 discrimination of non-speech complex tones so that the stimuli with tilted spectrum are confused with stimuli with flat spectrum and lower F0. Interestingly, the interdependence of breathiness and pitch height, proposed by Kuang and Liberman (2015) as a general auditory mechanism, fits quite well with Yip's (1993) analysis of the Shanghai Chinese tone system. Yip initially proposed (Yip, 1980) a pitch register feature [upper] to distinguish *yin* from *yang* tones. For example, *yin* T1 (52) would be represented as [+upper, HL] and *yang* T3 (23) as [-upper, LH]. Yip later proposed (Yip, 1993) that this pitch register feature is also a phonation register or voice quality feature—she called it [murmur]—specifying breathiness. Yip thus seemed to propose that pitch and phonation depend on one another, in line with the auditory "equivalence" of low pitch and breathiness (or at least, falling spectral tilt) suggested by Kuang and Liberman (2015). However appealing, the general auditory account is a bit at odds with at least one study showing that listeners from native languages differing in the phonetic or phonemic use of breathy voice do perceive breathiness differently, suggesting a linguistic rather than auditory locus for the integration of breathiness and pitch height (Esposito, 2010). In one experiment using vowel stimuli from 10 languages using breathy voice either phonologically or allophonically, non-native listeners were asked to evaluate the similarity between breathy and modal vowels, and then to categorize the vowels they were presented with. Listeners were from three language groups: Gujarati, which has a phonological contrast between breathy and modal voice, English, which has allophonic breathy voice, and Mexican Spanish, which has neither phonological nor allophonic breathy voice. Gujarati listeners showed a much greater sensitivity to non-native breathiness than English and Mexican Spanish listeners. This enhanced sensitivity of Gujarati listeners to the breathy–modal contrast strongly suggests that breathiness is not solely processed at an auditory, psychoacoustic level but may engage language-specific phonetic processing. The data from Esposito (2010), however, are quite indirect evidence against the general auditory account that breathiness induces perception of low pitch. More studies are needed to understand the seemingly contradictory findings we just summarized.

We tentatively propose that Shanghai Chinese listeners parse spoken syllables into breathiness and F0-contour cues, which they then integrate into a single tonal percept. When these cues are manipulated orthogonally, the perceptual categorization data that obtains is typical of a trading relation between F0-contour and breathiness. Should we postulate, as proposed by Yip (1993), a [murmur] feature that would trade with the L and H tonal features? Such a voice quality feature could be viewed as a remnant of a previous state of the Shanghai Chinese phonological system. However, in contemporary Shanghai Chinese, breathiness is a secondary, redundant *cue* to *yang* tone identity. In other words, the main determinant of the *yin–yang* tonal distinction

remains tone height in terms of pitch contour specification, that is, the [upper] feature initially proposed by Yip (1980).

The coexistence in mental representations of pitch-height (e.g., [upper]) and voice quality (e.g., [murmur]) features to define the *yin–yang* contrast may be specific to languages in the course of an ongoing, as yet unachieved diachronic change. We may consider there is a continuum between languages that use voice quality distinctively and pitch contour as a redundant or optional feature for tonal specification, such as Mon (Shorto, 1967), and languages in which pitch contour is distinctive and voice quality is redundant, such as Mandarin Chinese. Some languages seem to lie in-between these two extremes. Very often, such languages are in an intermediate stage of evolution toward a tonal height contrast. Tamang is a good example of an "in-between" language, in which distinctive and redundant features, including voicing of word-initial consonant, pitch, and voice quality are difficult to tease apart (Mazaudon & Michaud, 2008). Other examples of languages at an intermediate stage of evolution are Suai and Khmu'. These two Mon-Khmer languages seem to be evolving from "voice-register" languages, with a modal versus breathy contrast, towards a stable tonal language (Abramson & Luangthongkum, 2009). In Shanghai Chinese, more advanced in this evolution path than these "in-between" languages, pitch height clearly is a distinctive feature, whereas breathiness may be viewed as redundant (and disappearing) but still is parsed and integrated at the phonetic perception level. We therefore propose that both pitch height and breathiness are imprinted in the "phonetic grammar" of Shanghai speakers.

To sum up, our perception data suggest that Shanghai tones have a multidimensional feature specification. More generally, as proposed by Rose (1982) a long time ago, tones require a multidimensional specification. Shanghai Chinese would be no exception, with tones characterized by a bundle of phonetic cues. We have examined here the psychological reality of the breathiness cue in perception and may safely conclude it is quite robust, contrary to what is observed in production, whereby the breathiness distinction seems to be fading away in the Shanghai Chinese spoken by the young generation, especially by young female speakers, who, as observed elsewhere, would "lead the change" (Labov, 2001, p. 293; Gao, 2016), a change toward the pragmatically more useful and prestigious Standard Mandarin Chinese. Hence, in this case, production rather than perception is leading a sound change. An open issue is of whether perception-led sound changes are the neutral, default case and production-led changes are only motivated by sociolinguistic factors.

Acknowledgements

Part of the results has been presented in a Discussant session of the 18th International Congress of Phonetic Sciences. We wish to express our gratitude to Douglas Whalen, Andrea Levitt, and two anonymous reviewers for their constructive comments on the first version of this paper. We are also grateful to Hongwei Ding and Jing Peng for helping to recruit participants and providing test facilities at Tongji University of Shanghai. Finally, thanks to all the participants in the experiments run in Shanghai.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was conducted as part of a doctoral dissertation by the first author. It was supported by LabEx EFL (ANR/CGI), a doctoral fellowship to the first author, and the LexPhonSem (IDEX) grant to the second author.

ORCID iD

Jiayin Gao https://orcid.org/0000-0001-7572-4482

Notes

1. The H1–H2 values for /m/ in our modal versus breathy /mɛ/ stimuli, averaged across the eight steps of the continua and five equidistant time points, were 5.2 versus 22.5 dB, respectively.
2. For time-scaling and pitch-scaling, we used Praat scripts based on the Praat implementation of the Pitch Synchronous Overlap Add (PSOLA) method (Praat: Boersma & Weenink, 2018; PSOLA: Valbret, Moulines, & Tubach, 1992). Piece-wise time-scaling (different time-warping factors for different speech intervals) avoids cross-splicing the speech signal.
3. We also fitted cumulative Gaussian distributions to all the eight data points of each individual continuum and found similar intercept values. We only report here the intercepts computed for short ogive Gaussians.
4. Possibly supporting this speculation, we found a clearly lower spectral center of gravity in [f] for T3 than T2 natural base stimuli (293 vs. 1120 Hz); such a large differential was not found in synthesized [f] (2341 vs. 2528 Hz).

References

Abramson, A. S., & Luangthongkum, T. (2009). A fuzzy boundary between tone languages and voice-register languages. In G. Fant, H. Fujisaki, & J. Shen (Eds.), *Frontiers in phonetics and speech science* (pp. 149–155). Beijing, China: The Commercial Press.

Arai, T. (2006). Cue parsing between nasality and breathiness in speech perception. *Acoustical Science and Technology*, *27*, 298–301.

Bates, D., Maechler, M., & Dai, B. (2018). *lme4: Linear mixed-effects models using Eigen and S4*. Version 1.1–7, http://CRAN.R-project.org/package=lme4

Belotel-Grenié, A., & Grenié, M. (1997). Types de phonation et tons en chinois standard. *Cahiers de Linguistique Asie Orientale*, *26*, 249–279.

Belotel-Grenié, A., & Grenié, M. (2004). The creaky voice phonation and the organisation of Chinese discourse. *TAL-2004*, 5–8.

Berkson, K. (2016). Production, perception, and distribution of breathy sonorants in Marathi. *Proceedings of 6th FASAL (Formal Approaches to South Asian Languages)*, 4–14.

Berkson, K. (2019). Acoustic correlates of breathy sonorants in Marathi. *Journal of Phonetics*, *73*, 70–90.

Best, C. T., Morrongiello, B., & Robson, R. (1981). Perceptual equivalence of acoustic cues in speech and nonspeech perception. *Perception and Psychophysics*, *29*, 191–211.

Best, C. T., & Strange, W. (1992). Effects of phonological and phonetic factors on cross-language perception of approximants. *Journal of Phonetics*, *20*, 305–330.

Birkholz, P., Kröger, B. J., & Neuschaefer-Rube, C. (2011). Synthesis of breathy, normal, and phonation using a two-mass modal with a triangular glottis. *Proceedings of Interspeech 2011*, 2681–2684. *Florence, Italy*.

Boersma, P., & Weenink, D. (2018). Praat: Doing phonetics by computer [Computer program]. Version 6.0.40, retrieved 12 May 2018 from http://www.praat.org

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

Brunelle, M., & Finkeldey, J. (2011). Tone perception in Sgaw Karen. *Proceedings of the 17th International Congress of Phonetic Sciences*, 372–375. *Hong Kong, China*.

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

Chao, Y. R. (1928). *Studies in the modern Wu dialects*. Monograph No.4. Peking: Tsinghua College Research Institute.

Chen, G., Park, S. J., Kreiman, J., & Alwan, A. (2014). Investigating the effect of F0 and vocal intensity on harmonic magnitudes: Data from high-speed laryngeal videoendoscopy. *Proceedings of Interspeech 2014*, 1668–1672. *Singapore*.

Chen, Y. (2011). How does phonology guide phonetics in segment–*f0* interaction? *Journal of Phonetics*, *39*, 612–625.

Chen, Y., & Gussenhoven, C. (2015). Shanghai Chinese. *Journal of the International Phonetic Association*, *45*, 321–337.

Draxler, C. (2014). Online experiments with the Percy software framework – experiences and some early results. *Proceedings of LREC*, 235–240. Reykjavik, Iceland.

Esposito, C. M. (2010). The effects of linguistic experience on the perception of phonation. *Journal of Phonetics*, *38*, 306–316.

Fox, J., & Weisberg, S. (2011). *An R companion to applied regression*. Thousand Oaks, CA: Sage. http://z.umn.edu/carbook

Gao, J. (2016). Sociolinguistic motivations in sound change: On-going loss of low tone breathy voice in Shanghai Chinese. *Papers in Historical Phonology*, *1*, 166–186.

Gao, J., & Hallé, P. (2012). Caractérisation acoustique des obstruantes phonologiquement voisées du dialecte de Shanghai. *Proceedings of JEP-TALN-RECITAL*, 145–152.

Gao, J., & Hallé, P. (2013). Duration as a secondary cue for perception of voicing and tone in Shanghai Chinese. *Proceedings of Interspeech 2013*, 3157–3161.

Gao, J., & Hallé, P. (2017). Phonetic and phonological properties of tones in Shanghai Chinese. *Cahiers de Linguistique Asie Orientale*, *46*, 1–31.

Gårding, E., Kratochvil, P., Svantesson, J. O., & Zhang, J. (1986). Tone 4 and Tone 3 discrimination in modern standard Chinese. *Language and Speech*, *29*, 281–293.

Garellek, M., Keating, P., Esposito, C. M., & Kreiman, J. (2013). Voice quality and tone identification in White Hmong. *Journal of the Acoustical Society of America*, *133*, 1078–1089.

Garellek, M., Ritchart, A., & Kuang, J. (2016). Breathy voice during nasality: A cross-linguistic study. *Journal of Phonetics*, *59*, 110–121.

Hallé, P., Best, C. T., & Levitt, A. (1999). Phonetic vs. phonological influences on French listeners' perception of American English approximants. *Journal of Phonetics*, *27*, 281–306.

Hallé, P., Chang, Y-C., & Best, C. (2004). Categorical perception of Taiwan Mandarin Chinese tones by Chinese versus French native speakers. *Journal of Phonetics*, *32*, 395–421.

Hallé, P., & Gao, J. (2016). *Shanghai Chinese obstruent durations vary with voicing: A phonological or phonetic effect? Communication at the 15th Laboratory Phonology Conference*, *Ithaca, USA*.

Karlgren, B. (1915–1926). Études sur la phonologie chinoise. *Archives d'études orientales*, 15. Uppsala: K. W. Appelberg; Leiden: E. J. Brill.

Kuang, J. (2013). The tonal space of contrastive five level tones. *Phonetica*, *70*, 1–23.

Kuang, J. (2017). Covariation between voice quality and pitch: Revisiting the case of Mandarin creaky voice. *Journal of the Acoustical Society of America*, *142*, 1693–1706.

Kuang, J., & Liberman, M. (2015). Influence of spectral cues on the perception of pitch height. *Proceedings of the 18th International Congress of Phonetic Sciences*. Paper no. 435, 1–5.

Labov, W. (2001). *Principles of linguistic change, Volume 2: Social factors*. Oxford, UK: Blackwell.

Lenth, R. (2018). emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.3.0.

Liu, F. (1923). Study on the 36 initials of Shouwen. *Guoxue Jikan*. *1*, 451–464.

Mazaudon, M., & Michaud, A. (2008). Tonal contrasts and initial consonants: A case study of Tamang, a "missing link" in tonogenesis. *Phonetica*, *65*, 231–256.

Ohala, J. J. (1975). Phonetic explanations for nasal sound patterns. In C. A. Ferguson, L. M. Hyman, & J. J. Ohala (Eds.), *Papers from a symposium on nasals and nasalization* (pp. 289–316). Stanford, CA: Language Universals Project.

Ohala, J. J. (1981). The listener as a source of sound change. In C. S. Masek, R. A. Hendrick, & M. F. Miller (Eds.), *Papers from the parasession on language and behavior* (pp. 178–203). Chicago, IL: Chicago Linguistic Society.

Ohala, J. J. (1993). Sound change as nature's speech perception experiment. *Speech Communication*, *13*, 155–161.

Ohala, J. J., & Amador, M. (1981). Spontaneous nasalization. *Journal of the Acoustical Society of America*, *69*, S54–S54.

Ohala, J. J., & Ohala, M. (1992). *Nasals and nasalization in Hindi*. Communication at 3rd International Symposium on Language and Linguistics: Pan-Asiatic Linguistics. Bangkok, Thailand: Chulalongkorn University.

R Development Core Team (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org

Ren, N. (1992). *An acoustic study of Shanghai stops*. PhD dissertation, University of Connecticut, USA.

Repp, B. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Psychological Bulletin*, *92*, 81–110.

Rose, P. (1982). Acoustic characteristics of the Shanghai-Zhenhai syllable-types. In D. Bradley (Ed.), *Papers in Southeast Asian Linguistic, No 8: Tonation (Pacific Linguistics, Series A, no. 62)*, 1–53, Canberra, Australia: Australian National University.

Rose, P. (1989). Phonetics and phonology of Yang tone: Phonation types in Zhenhai. *Cahiers de Linguistique Asie Orientale 18*(2), 229–245.

Rose, P. (1993). A linguistic-phonetic acoustic analysis of Shanghai tones. *Australian Journal of Linguistics*, *13*, 185–220.

Rose, P. (2002). Independent depressor and register effects in Wu dialect tonology. *Journal of Chinese Linguistics*, *30*, 39–81.

Shen, Z. W., Wooters, C., & Wang, W. Y (1987). Closure duration in the classification of stops: A statistical analysis. *Ohio State University Working Papers in Linguistics*, *35*, 197–209.

Shorto, H. L. (1967). The register distinctions in Mon-Khmer languages. *Wissenschaftliche Zeitschrift der Karl-Marx-Universität, Leipzig, Gesellschafts-und sprachwissenschaftliche Reich*, *1*(2), 245–248.

Shue, Y.-L., Keating, P., Vicenik, C., & Yu, K. (2011). VoiceSauce: A program for voice analysis. Available at http://www.ee.ucla.edu/~spapl/voicesauce/.

Swerts, M., & Veldhuis, R. (2001). The effect of speech melody on voice quality. *Speech Communication*, *33*, 297–303.

't Hart, J. (1981). Differential sensitivity to pitch distance, particularly in speech. *Journal of the Acoustical Society of America*, *69*, 811–821.

Tian, J., & Kuang, J. (2016). Revisiting the register contrast in Shanghainese. *Proceedings of the 5th Symposium on the Tonal Aspects of Languages*, 147–151.

Tsuji, N. (1977). *Eight Yue dialects in Guangxi Province, China, and reconstruction of proto-Yue phonology*. PhD dissertation, Cornell University, USA.

Valbret, H., Moulines, E., & Tubach, J. P. (1992). Voice transformation using PSOLA technique. *Speech Communication*, *11*, 175–187.

Whalen, D. H., & Xu, Y. (1992). Information for Mandarin tones in the amplitude contour and in brief segments. *Phonetica*, *49*, 25–47.

Xu, B., & Tang, Z. (1988). *Shanghai Shiqu Fangyan Zhi* [Urban Shanghai dialect records]. Shanghai, China: Shanghai Education Press.

Xu, B., & Tao, H. (1997). *Shanghaishi Fangyan Cidian* [Dictionary of Shanghainese]. Jiangsu, China: Jiangsu Education Press.

Yip, M. (1980). *The tonal phonology of Chinese*. PhD dissertation, Massachusetts Institute of Technology, USA.

Yip, M. (1993). Tonal register in East Asian languages. In H. van der Hulst & K. Snider (Eds.), *The phonology of tones* (pp. 245–268). Berlin, Germany: Mouton de Gruyter.

Yu, K. M., & Lam, H. W. (2014). The role of creaky voice in Cantonese tonal perception. *Journal of the Acoustical Society of America*, *136*, 1320–1333.

Zee, E. (2003). Shanghai phonology. In G. Thurgood & R. J. LaPolla (Eds.), *The Sino-Tibetan languages* (pp. 131–138). London, UK: Routledge.

Zhang, J., & Yan, H. (2015). Contextually dependent cue weighting for a laryngeal contrast in Shanghai Wu. *Proceedings of the 18th International Congress of Phonetic Sciences*. Paper no. 362, 1–5.

Zhang, J., & Yan, H. (2018). Contextually dependent cue realization and cue weighting for a laryngeal contrast in Shanghai Wu. *Journal of the Acoustical Society of America*, *144*, 1293–1308.

Zhu, X. (1999). *Shanghai tonetics* (Vol. 32). Lincom Europa.