Duration as a secondary cue for perception of voicing and tone in Shanghai Chinese

Jiayin Gao\textsuperscript{1} & Pierre Halle\textsuperscript{1,2} \\
\textsuperscript{1} Laboratoire de Phonétique et Phonologie – CNRS / Université Paris 3 \\
\textsuperscript{2} Laboratoire Mémoire et Cognition – Université Paris 5 \\
jiajin.gao@univ-paris3.fr, pierre.halle@univ-paris3.fr

Abstract

Previous studies have reported phonetic characteristics of the Shanghai Chinese phonological voicing contrast, which co-occurs with a tonal contrast. In stressed word-initial position, phonetic voicing is neutralized and replaced with a tonal register contrast: high ‘yin’ tones for (phonologically) voiceless and low ‘yang’ tones for voiced obstruents. Furthermore, breathy vs. modal voice quality, and low vs. high C/V duration ratio accompany voiced vs. voiceless obstruents. In two syllable identification experiments, we explored the impact of these characteristics on the perception of underlying phonological voicing. In Experiment 1, we manipulated tone contour (‘yin’ vs. ‘yang’) while maintaining other phonetic properties, including duration pattern. Syllable identification was mainly determined by the imposed contour, except for syllables with a voiced labial fricative onset. However, response times tended to increase when the imposed contour differed from the original one. In Experiment 2, we manipulated duration pattern and created tone contour continua from a ‘yin’ tone to a ‘yang’ tone. The duration pattern manipulation influenced identification in that high C/V duration ratios induced more frequent and faster ‘yin’ identification (phonologically voiceless onset syllable). This result only held for unchecked syllables. We conclude that duration pattern contributes to the perception of phonological voicing in Shanghai Chinese.

Index Terms: Shanghai Chinese, perception, voicing, duration pattern, tone

1. Introduction

Shanghai Chinese belongs to the Wu family and refers to the language spoken in Shanghai, although different dialectal variants exist in the surrounding suburban areas [1]. It is characterized, phonologically, by its three-way laryngeal contrast between voiceless unaspirated, voiceless aspirated, and voiced, which are respectively ‘quanqing 全清’ (fully clear), ‘ciqing 次清’ (secondarily clear), and ‘quanzhuo 全浊’ (fully muddy), as labeled by Chinese linguists (e.g., [2]).

Multiple acoustic and articulatory features (F0, intensity, voice quality, duration, etc.) correlating with this contrast have been suggested in earlier impressionistic descriptions and studied in recent instrumental investigations. For sake of clarity, “voicing” without further qualification refers in the following to phonological voicing; phonetic voicing is specified otherwise.

Syllables with a voiced vs. voiceless onset bear a tone of the low vs. high tone register. The diachronic explanation goes like this: late Middle Chinese (end of Tang dynasty) underwent tone split, a tonal development shared by languages of a vast geographical zone of South-east Asia, in which voiceless syllable onsets produced a high tone register called ‘yin’, whereas voiced ones produced a low tone register called ‘yang’ [3]. The voicing contrast was transphonologized into a yin–yang tone register contrast in many Chinese dialects. In Shanghai Chinese, tone split was also achieved and produced ‘yin’ tones (T1: 53, T2: 34, T4: 5) and ‘yang’ tones (T3: 23, T5: 2), but the voicing contrast was maintained. The tone register yin-yang contrast applies to stressed word-initial syllables, whose onset is now phonetically devoiced. Note that phonetic voicing for ‘yang’ fricative onsets in stressed word-initial syllables has been observed in [6]. Phonetic voicing applies to the onsets of unstressed syllables in non-initial position, where the tonal contrast is neutralized due to tone sandhi (see [4], among others), although [5] found that, even in this condition, voicing affects tone contour: F0 is lowered by voiced onsets but is raised by voiceless ones.

Another laryngeal feature, voice quality, has often been proposed by linguists as accompanying phonological voicing in northern Wu dialects (including Shanghai Chinese): syllables with a voiced obstruent onset are produced with breathy voice. This was first described as ‘qingyin zhuoliu’ (clear sounds followed by muddy breathing) in [6] and [2] and was recently substantiated by acoustic data [8] as well as physiological investigations ([9] for fiberoptic data; but see [10] for ePGG data). This feature may be traced back to late Middle Chinese: in the course of transphonologization from voicing to tone register contrast, all voiced initial consonants might have developed a breathy quality, as suggested by the traditional Chinese term “muddy” for the voiced series [11]. We can still observe the breathy quality associated with voiced initial consonants in several languages that underwent the tone split, such as Chinese Northern Wu, Mon-Khmer languages [12],[13], some Tamang languages [14], etc.

Duration pattern is found to be another robust feature that distinguishes the voiceless and voiced series. In intervocalic position, when voiced series are phonetically voiced, voiced stops have shorter closure duration than voiceless stops [15],[4],[6], and voiced fricatives have shorter duration than voiceless fricatives [6]. Vowel duration also varies according to the voicing of the following stop [4]: long before a voiced stop and short before a voiceless stop, as is observed in many other languages (e.g., [16] for English). Besides, vowel duration is conditioned in the same way by a preceding obstruent (long after a voiced obstruent and short after a voiceless obstruent, other things being equal) [4]. In word-initial position, when voiced obstruents are phonetically voiceless, duration pattern seems to maintain, with short voiceless consonants followed by short vowels and short “voiced” consonants followed by long vowels, for stops [15] as well as for fricatives [6]. There is a similar trend in Korean: tense obstruents shorten the following vowel but lax ones do not, as found in [17].

It has been proposed in [18] that, in English, the fricative voicing contrast, just like the stop voicing contrast, is at least in part a duration contrast, namely a contrast of frication duration. The duration of frication and that of the preceding vowel are not only acoustic correlates of fricative voicing.
Thirty-six natural CV monosyllabic words were recorded in 2.2. experiment. All were naive as to the purpose of the study. Fifteen native speakers of Shanghai Chinese (8 males and 7 females) aged from 21 to 29 years (mean 25) participated in Experiment 2.1. Participants were 3158. In Experiment 1, we used a two-fold forced-choice syllable identification task. The stimuli were presented. The identification test was conducted using the E-Prime software. Participants were presented with the stimuli through professional quality headphones; they were tested individually in a quiet room, in front of a computer. Each trial consisted of the following events: at trial onset, a fixation cross was displayed at the center of the screen; 500 ms after trial onset, one of the stimuli was presented; 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 character whose reading matched the auditory stimulus, that is, the correct response character, appeared either on the right or on the left of the screen. The correct response side was counterbalanced across the two or four repetitions of the same stimulus. The Chinese characters for identification responses were based on the subjective frequencies (on a 1-5 scale) collected from 17 native speakers of Shanghai Chinese who did not participate in the perception experiments. They rated 107 characters (mixed with 10 distracters). For each minimal pair or triplet, we selected from each group of homophone characters two or three characters with the closest ratings. For each stimulus, participants were asked to indicate the character whose reading matched best the stimulus by pressing one of the two possible responses to the trial. The character whose reading matched the auditory stimulus, that is, the correct response character, appeared either on the right or on the left of the screen. The correct response side was counterbalanced across the two or four repetitions of the same stimulus. The Chinese characters for identification responses were based on the subjective frequencies (on a 1-5 scale) collected from 17 native speakers of Shanghai Chinese who did not participate in the perception experiments. They rated 107 characters (mixed with 10 distracters). For each minimal pair or triplet, we selected from each group of homophone characters two or three characters with the closest ratings. For each stimulus, participants were asked to indicate the character whose reading matched best the stimulus by pressing one of the two possible responses to the trial. The character whose reading matched the auditory stimulus, that is, the correct response character, appeared either on the right or on the left of the screen. The correct response side was counterbalanced across the two or four repetitions of the same stimulus. The Chinese characters for identification responses were based on the subjective frequencies (on a 1-5 scale) collected from 17 native speakers of Shanghai Chinese who did not participate in the perception experiments. They rated 107 characters (mixed with 10 distracters). For each minimal pair or triplet, we selected from each group of homophone characters two or three characters with the closest ratings. For each stimulus, participants were asked to indicate the character whose reading matched best the stimulus by pressing one of the two possible responses to the trial. The character whose reading matched the auditory stimulus, that is, the correct response character, appeared either on the right or on the left of the screen. The correct response side was counterbalanced across the two or four repetitions of the same stimulus. The Chinese characters for identification responses were based on the subjective frequencies (on a 1-5 scale) collected from 17 native speakers of Shanghai Chinese who did not participate in the perception experiments. They rated 107 characters (mixed with 10 distracters). For each minimal pair or triplet, we selected from each group of homophone characters two or three characters with the closest ratings. For each stimulus, participants were asked to indicate the character whose reading matched best the stimulus by pressing one of the two possible responses to the trial. The character whose reading matched the auditory stimulus, that is, the correct response character, appeared either on the right or on the left of the screen. The correct response side was counterbalanced across the two or four repetitions of the same stimulus.
A three-way ANOVA on the accuracy data, with Congruence (two levels), Place of articulation (stop vs. fricative) and Manner of articulation (dental vs. labial) as within-subject factors, showed a significant effect of Congruence, \( F(1,14)=60.8, p<.0001 \). Thus, the overall trend described above (97.9% vs. 90.4%) was significant. The significant Congruence x Place x Manner interaction, \( F(1,14)=18.2, p<.001 \), suggested differences in the effect of Congruence across syllable onsets. The effect of Congruence was significant only for the labial fricative onsets, \( F(1,14)=105.4, p<.0001 \), as can be seen in Figure 1.

Whereas the accuracy data suggested the effect of Congruence is limited to the labial fricative syllables, the RT data for correct responses suggested otherwise.

Table 1. Identification RTs (ms) for congruent vs. incongruent syllables (significance levels: \( * p<.05 \), \( ** p<.0001 \)).

<table>
<thead>
<tr>
<th></th>
<th>Labial stop</th>
<th>Dental stop</th>
<th>Labial fric.</th>
<th>Dental fric.</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>congrn</td>
<td><strong>861</strong></td>
<td>929</td>
<td>871</td>
<td>868</td>
<td><strong>833</strong></td>
</tr>
<tr>
<td>incongrn</td>
<td></td>
<td></td>
<td>1164</td>
<td>769</td>
<td>936</td>
</tr>
</tbody>
</table>

Table 1 shows the RT differences between congruent and incongruent syllables (103 ms in average) for all types of syllables. We excluded labial fricative syllables (light-grey cells) from statistical analyses because the number of observations was unbalanced relative to the other syllables. The accuracy data for labial fricatives were anyway sufficiently clear-cut to show that listeners did not solely rely on F0 contour to identify the labial fricative syllables.

To summarize, we found that Shanghai Chinese listeners rely heavily on tone contour to identify yin-yang minimal pairs. However, they seem to additionally use other phonetic-acoustic cues (which remain to be qualified), especially for /v/-onset syllables. These syllables tend to be produced with some initial phonetic voicing (cf. [15]), which could explain why they are still identified as ‘yang’ after their tone contour is switched to ‘yin’. For the other onsets, listeners’ use of non-tonal cues is revealed by slower RTs in case of incongruent tonal and segmental information.

3. Experiment 2

Experiment 1 found that non-tonal characteristics influence syllable identification. But which non-tonal characteristics? Experiment 2 explores the possibility that duration patterns influence syllable identification. We manipulated the C/V duration ratio and looked for its possible influence on tone categorization, using T3-T2 and T5-T4 tone continua.

3.1. Participants

Twenty-six native speakers of Shanghai Chinese aged 18 to 34 years (mean 22.7), born and raised in urban areas of Shanghai, participated in Experiment 2. None reported hearing or reading disorder. All were naive about the purpose of the experiment.

3.2. Materials, design, and procedure

The stimuli were constructed from eight natural CV(?) monosyllabic words pronounced by the first author (a 25-year-old native speaker of Shanghai Chinese). C was a labial or dental fricative (\( v, v, s, z \)). Because we wanted to compare the possible role of durations in checked versus unchecked syllables, the syllable rime was /\( \text{ɛv} \)/ or /\( \text{æt} \)/. /\( \text{fɛ} \)/ and /\( \text{sɛ} \)/ were produced with tone T2 (34), /\( \text{vɛ} \)/ and /\( \text{sɛ} \)/ with tone T4 (5), /\( \text{væ} \)/ and /\( \text{æt} \)/ with tone T3 (23), and /\( \text{vɛ} \)/ and /\( \text{æe} \)/ with tone T5 (2); T2 and T4 are ‘yin’ tones, T3 and T5 are ‘yang’ tones. The speaker pronounced the syllables in word-initial position. She produced phonetically voiceless onsets across the board.

Two duration patterns were then applied to the original syllables using PSOLA [21]: LS (long onset and short rime) and SL (short onset and long rime). In the LS pattern, original onset duration was increased by 30% and original rime duration decreased by 30% (glottal stop and creaky voice portions were not altered for sake of naturalness); opposite changes were applied to produce the SL pattern syllables.

An eight-step yin-yang T2-T3 or T4-T5 tone contour continuum was imposed on each of the /\( \text{ɛv} \)-rime or /\( \text{æt} \)-rime syllables, respectively. Eight yin-yang continua were thus constructed for each LS and SL duration pattern. The endpoint contours were taken from the original ‘yin’ and ‘yang’ syllables. Figure 2 shows a T2-T3 continuum whose endpoints are taken from /\( \text{vɛ} \)/ (T2) and /\( \text{væ} \)/ (T3).

Figure 2: T2-T3 continuum on the rime /\( \text{ɛv} \)/ (with labial fricative onset): the T2 and T3 endpoints are the highest and lowest contours, respectively.

The test phase consisted of 256 trials (4 onsets x 2 rimes x 2 duration patterns x 8 steps x 2 repetitions) presented to participants in pseudo-random order (stimuli sharing onset and
rime could not appear in succession). It was preceded by a training phase of 6 trials in which feedback was provided.

The same two-fold forced choice identification procedure as in Experiment 1 was used. Yet, since the notion of “correct” response could not be used with stimulus continua, subjects’ responses were simply recorded as ‘yin’ or ‘yang’. The characters used as possible identification responses were chosen based on the subjective frequencies collected for Experiment 1.

### 3.3. Results and discussion

Participants’ ‘yin’ response rate and response times for ‘yin’ and ‘yang’ responses (yin-RTs and yang-RTs) were analyzed. Figure 3A-B shows ‘yin’ identification curves according to duration pattern and rime. For /ɛ/-rime syllables, category boundaries for the LS and SL patterns differ: For LS syllables, this boundary is shifted toward the ‘yang’ endpoint, that is, there are more ‘yin’ responses than for SL syllables. This perceptual shift is not observed for /ɐʔ/-rime syllables.

![Figure 3: (A) /ɛ/-rime and (B) /ɐʔ/-rime syllables. Step 0 corresponds to the ‘yin’ (T2 or T4) endpoint.](image)

Overall, 55.4% of the syllables were identified as ‘yin’ for the LS pattern. This percentage went down to 50.6% for the SL pattern. A three-way ANOVA was conducted on the ‘yin’ response rate data, with Duration pattern (LS vs. SL), Onset (/f, v, s, z/), and Rime (/ɛ vs. /ɐʔ/) as within-subject factors. Duration pattern had a significant effect overall, $F(1,25)=20.5, p<.0001$ (more ‘yin’ responses for the LS pattern). Duration x Rime interaction was significant, $F(1,25)=6.4, p<.05$, reflecting the larger effect of Duration pattern for /ɛ/- than /ɐʔ/-rime syllables, although the effect was significant for both rimes, $r_{es}<.05$. Duration x Onset interaction was significant, $F(1,25)=18.5, p<.001$, reflecting the larger effect of Duration pattern for labial than dental fricatives (labial: $F(1,25)=18.5, p<.001$; dental: $F(1,25)=3.50, p=.07$).

Figure 4 shows identification RTs according to duration pattern for /ɛ/-rime syllables: yin-RTs are plotted for steps 0 to 3 (where ‘yin’ responses are dominant) and yang-RTs for steps 4 to 7 (where ‘yang’ responses are dominant). The yin-RTs at steps 2 and 3 (close to the ambiguous region of the continuum) were longer for the SL than LS pattern, suggesting again that the LS pattern facilitates ‘yin’ decisions for /ɛ/-rime syllables. The converse held for the yang-RTs at steps 4 and 5 (close to the ambiguous region of the continuum), suggesting that the SL pattern facilitates ‘yang’ decisions for /ɛ/-rime syllables.

To summarize, Experiment 2 showed that C/V duration pattern influences tone categorization in a yin-yang tone continuum. High C/V duration ratios favor ‘yin’ categorization compared to low C/V ratios. This result is specific to Shanghai Chinese in that its ‘yin’ vs. ‘yang’ tones are associated with phonologically voiceless vs. voiced onsets, respectively. It is universal, however, in that C duration is cross-linguistically related to C voicing. Our data suggest that the C/V duration patterns observed in the production of modern Shanghai Chinese syllables contribute to Shanghai listeners’ perception of phonological voicing, even in the absence of phonetic voicing.

![Figure 4: (A) /ɛ/-rime and (B) /ɐʔ/-rime syllables: black for LS vs. grey for SL.](image)

### 4. Discussion and conclusion

Experiment 1 showed that F0 contour is the dominant cue to syllable identification for yin-yang minimal pairs but that the congruence between tonal and non-tonal characteristics speeds up syllable identification. Syllables with labial fricative onset, especially /v/, differed markedly from the others: switching their tone contour from ‘yin’ to ‘yang’ or vice versa, affected correct tone identification. Non-tonal characteristics thus influenced syllable identification. In the case of /v/ onset syllables, the non-tonal characteristic is presumably phonetic voicing in word-initial /v/, as observed in [15], especially with young speakers of Shanghai Chinese. We may therefore suggest that phonetic voicing is not completely neutralized in word-initial position, neither in production nor in perception. Other non-tonal characteristics than phonetic voicing only affect response time without affecting response accuracy.

Experiment 2 explored the role of duration pattern. We found that duration pattern influences categorical perception of tone in yin-yang tone continua. But this influence was only found for /ɛ/-rime syllables: high C/V duration ratios induce faster and more frequent ‘yin’ identifications. For /ɐʔ/-rime syllables, C/V duration pattern does not seem to play a role. This might be explained by the fact that we avoided to time-scale the creaky part of the /v/ rimes. This might have made less perceptible the duration pattern contrasts on the whole syllable. Alternatively, the presence of a glottal stop might reduce the perceptual impact of vowel and onset durations.

As Phil Rose put it, tonal investigation requires a “polydimensional approach” “contra the prevalent monodimensional stance which ignores from the outset all parameters except F0/pitch variation” [23], see also [24]). In line with this stance, we showed that non-tonal characteristics, in particular C/V duration pattern, influence syllable identification in Shanghai Chinese. Besides, the voicing contrast assumed to neutralize in word-initial position may surface occasionally in the production of labial fricatives [6]. For these onsets, the voicing contrast interferes with the tonal contrast in syllable identification. But in all cases, C/V duration pattern seems a robust secondary cue to voicing in both production and perception.

### 5. Acknowledgements

We are very grateful to Cécile Fougeron for her invaluable suggestions to an earlier draft of this paper, to Thibault Fux for his help in the Matlab implementation of PSOLA, and to Weilin Shen for his help in E-Prime design.
6. References