MUSIC PERCEPTION INFLUENCES PLOSIVE PERCEPTION IN WU DIALECTS

Marjoleine Sloos[1], Jie Liang[2], Lei Wang[2]

[1]Aarhus University, [2]Tongji University
marj.sloos@gmail.com, liangjie56@163.net, leiwang1987@126.com

ABSTRACT

Wu is a dialect group of the Chinese branch of Sino-Tibetan languages. Wu dialects are known for having plain, aspirated as well as voiced stops. Crucially, voiced plosives always co-occur with low-register tones. We investigated the perception of voicing distinction among phonetically and phonologically trained Wu native speakers by superimposing different tones on syllables starting with originally plain, aspirated, and voiced stops. The results show that recognition of the voicing contrast turned out to be largely inaccurate, and the subjects mostly relied on lexical tone rather than on phonation itself.

Subsequently, we examined the perception of music improved the recognition of the phonation distinction. Although the perception of the voicing distinction did not become more accurate, it turned out that listening to musical fragment in between the language fragments led to a different classification of the lexical tones. This, in turn, led to a different perception of the plosives.

Keywords: Wu, biased perception, lexical tone, music perception, phonation.

1. LANGUAGE AND MUSIC TRANSFER

The two main auditory domains—language and music—do not only show structural similarities but also similarities in cognitive processing (see [1-2] among many others). In a broader sense, musically trained people appear to have an advantage across a range of skills, like phonemic awareness, reading, and mathematics—and even showed a higher than average IQ [3,4]. Transfer from music to language processing is specifically studied by comparing listeners with and without musical education. This kind of research repeatedly showed enhancement in perception and production of intonation and lexical tone in second language acquisition among musically trained subjects (e.g. [5-9]).

Short term transfer effects, like the immediate effect on language perception by listening to music, have less often been investigated (apart from the more general and hotly debated Mozart Effect [10]). Nevertheless, short term effects are relevant for individual speech sound perception, in second language acquisition, as well as in native language perception. Speech perception, after all, not only relies on incoming stimuli, but is also considerably influenced by other factors, like the native phoneme inventory [11-12], sociolinguistic factors (e.g. perceived age and social class), and the overall perception of the variety and expectations about the pronunciation of that variety [13], and—also among linguists—knowledge and expectations about the variety to which one is exposed [14].

In this contribution, we explore the possibility of the influence of music perception transfer to the perception of individual speech sounds, in relation to listeners' expectations. We concentrate on the perception of the voicing contrast in Wu plosives among Wu linguists. Finding that their ability to distinguish the original voicing contrast based on phonation is rather poor, we repeated the experiment in which the language stimuli alternated with musical fragments. Although overall accuracy did not improve under the music condition, we observed a remarkable difference: under the music condition, voicing was attributed to plosives that co-occurred with a rising tones, whereas under the non-music condition, voicing was most likely to be attributed to plosives that co-occurred with mid tones.

2. WU PLOSIVES AND LEXICAL TONES

Wu is the second largest dialect group in China in terms of the number of speakers, after Mandarin [15]. It is spoken in the southeast of China, including Shanghai. Two of its main features are a three-way phonation contrast among plosives and a more complex tone system than Mandarin, including a distinction between a low and high tonal register. These two factors (tone and phonation) are related. We will discuss phonation in section 2.1 and lexical tone in section 2.2.

2.1. Plosives

Wu dialects have plain, aspirated, and voiced stops. Each of these natural classes combines with three places of articulation: labial [p^h p b], coronal [t^h t d], and velar [k^h k g]. Unlike the other plosives, voiced stops only occur in initial and medial position but not in the syllable coda. However, voiced stops are only really voiced in medial position. In initial position, they surface as breathy voiced [16-19]. Breathiness spreads from the plosives to the following vowel [16-17]. The acoustic properties of this breathiness can be defined by the difference between the first and second harmonic: H1−H2 is higher if the vowel is preceded by a plain stop; and lower if the vowel is preceded by a breathy voiced stop (at least for the beginning and the medial parts of the rhyme) [18]. Crucially, voiced consonants always co-occur with low register lexical tones [20].

2.2. Lexical tones

Tonal systems in Wu differ drastically across dialects, but in general two registers are distinguished. The total number of lexical tones varies from five in Shanghainese [21] to eight in e.g. Shaoxing [22] or Wenzhou [23]. Checked tones (in which the syllable ends in a glottal stop) may occur and are shorter than other tones. We did not implement these in our study and will therefore not discuss them here further.

The acoustic cue for breathiness (namely H1−H2, see section 2.1) is not as robust as voice onset time (VOT) which is the acoustic parameter that corresponds to the plosive distinction, either in terms of aspiration or voicing. The acoustic description of breathiness given above largely depends on the *vowel* quality rather than on the plosive. Can breathiness of the consonant be perceived independently of lexical tone? Given the correspondence between low-register tones and voiced consonants, can perception be only dependent on lexical tone? Or, alternatively, could it be the case that both phonation and lexical tone contribute to the distinction between voiced and plain stop (similar to the equivalent contribution of tenseness and length in the distinction between long tense vowels and short lax vowels in Dutch [24])? The key question is thus: what is cue weighting of phonation and tone in voicing distinction in Wu dialects? We investigate this for Shanghainese by presenting subjects with syllables in which we combined all three phonation types with four different pitch contours.

3. METHODOLOGY

3.1. Subjects

Ten native Wu speakers, fluent in Mandarin as well, without reported hearing disorders, participated in the study. All subjects were phonetically and phonologically trained as to ensure that they were aware of the three phonation types and the correspondence between low register tones and voiced plosives. They were unaware of the purpose of the research. All subjects were paid for their participation.

3.2. Design

The experiment is part of a larger perception study among Wu and Mandarin speakers. This part consisted of four sessions. Each participant took part in all sessions, with intervals of approximately two weeks. During one session, 144 stimuli were presented: 4 blocks ∗ 4 tones * 9 different syllables. The four blocks were separated by a break of 63.0 seconds. The order of the stimuli was quasi-randomized within each block such that the same tone did not occur more than twice in a sequence, and each subsequent stimulus had a different plosive than the previous one. Thirty-six stimuli were separated by intervals of 7.0 seconds to provide time to note down the stimulus and were presented in a different order within each block. After each set of 6 stimuli, a sine sound of 440Hz (default in Praat [25] speech processing software) with a duration of 400ms was included, in order to help the subjects keep track of the experiment, since they had to fill in their responses in an Excel sheet.

During the third and fourth session, identical stimuli were used in identical order, but this time the stimuli alternated with musical fragments. Each block started with 63.0 seconds of a musical fragment and instead of an interval of silence after 36 stimuli, we presented the first 7 seconds of the audio clip used at the beginning of the block. All audio fragments were faded out at the end.

3.3. Material

The language stimuli were taken from the Asian English Speech Corpus Project of the Chinese Academy of Social Sciences [26]. We selected nine syllables /p^{h}a pa ba/ /t^{h}a ta da/ and /k^{h}a ka ga/ as pronounced by a Shanghainese female speaker. The syllable /t^{h}a/ differed from the other syllables because it had a centralized vowel. In order to arrive at a comparable set of stimuli, we therefore cut and concatenated the onset of /t^{h}a/ with the vowel of the syllable /k^{h}a/. Subsequently, we created four different tones with a bandwidth of 150-250Hz: high-level 55, rising 24, mid-level 33, and falling 51, using the Praat speech processing software [25].

These pitch contours were superimposed on all nine syllables, thus resulting in 36 stimuli.

In order to avoid effects of differences in music exposure among the subjects, we selected musical fragments that were presumably unfamiliar to all subjects, belonging to the genre of jazz. The subjects were asked to pay close attention to the instruments and were requested to indicate which instruments they perceived, to attract their attention to the music as much as possible. Given their unfamiliarity with western musical instruments, a sheet of paper with pictures in full colour of all instruments used was provided. We used the following musical fragments (all live recordings):

- Block 1: "All Blues" (Bobby Ramirez – flute, Kiki Sanchez – piano, Ivan Velasquez – drums, Jose Velasquez – bass). Recorded 12-11-2006.
- Block 2: “Melancholy Blues” (The Hot Five: Kid Ory – trombone, Johnny Dodds – clarinet, Johnny St. Cyr – banjo, Lil Armstrong – piano, Louis Armstrong - cornet or trumpet). Recording: Okeh 8496, 1927.
- Block 3: “Slow” (Earl Swope – trombone, Stan Getz, Zoot Sims - tenor sax, Al Cohn - tenor saxophone, arranger, Duke Jordan – piano, Jimmy Raney – guitar, Mert Oliver – bass, Charlie Perry – drums). Recorded: NYC, May 2, 1949, Savoy 967.
- Block 4: “Autumn leaves” (Retaw Boyce, violin) Online release: 25-04-2009.

3.4. Procedure

The experiment took place in a quiet room at Tongji University or in the sound insulated room of Fudan University (Shanghai). The sound file was presented to the subjects auditorily via a laptop over a Sennheiser HD201 headphone. The subjects filled in the perceived plosives in a column in a Microsoft Excel file. During the musical exposure they indicated the musical instruments they heard.

4. RESULTS

Regarding the aspirated plosives, the subjects performed at ceiling. This was not the case for plain and voiced consonants, however. We first address the accuracy of the subjects’ distinction between breathy voiced consonants and plain consonants. The results show a very weak correlation between original and reported phonation (φ = 0.049). Under the music condition, performance was only slightly more accurate, with a correlation of φ = 0.057 (Table 1).

To investigate the factors that played a role in the perception of the voicing distinction, we conducted a logistic repeated measures regression test with a within subjects design using the lme4 package [27] in the R statistical environment [28]. The dependent variable was the perceived phonation (voiced or plain) and the independent variables were original phonation, tone, music, and place of articulation. Random effects were subject and session. Negative estimates and *z*-values should be interpreted as a higher number of reports as plain consonants and positive estimates and *z*-values values should be interpreted as a higher number of reports as voiced consonants. The results are provided in Table 2.

Table 1: Confusion matrix of phonation.

		Response phonation	
	Original Phonation	Plain	Voiced
Non-music	Plain	765	194
	Voiced	726	233
Music	Plain	790	169
	Voiced	747	213

Table 2: The estimates, Standard Error, z-value, and p-value of music, tone, original voicing, and place of articulation. Significance at the 95% confidence interval level is indicated by asterisks.

	Est.	S.E.	*z*-value	*p*-value
(Intercept)	−6.780	0.862	−7.862	$<0.001^{*}$
Orig.Voicing	0.758	0.139	5.472	$<0.001^{*}$
T 55	2.353	0.485	4.856	$<0.001^{*}$
T 33	6.993	0.492	14.204	$<0.001^{*}$
T 24	2.267	0.486	4.661	$<0.001^{*}$
Music	1.274	0.574	2.218	$<0.001^{*}$
Labial	0.598	0.172	3.478	$<0.001^{*}$
Velar	1.153	0.172	6.700	$<0.001^{*}$
Music:T 55	−3.464	0.687	−5.042	$<0.001^{*}$
Music:T 33	−6.653	0.597	−11.154	$<0.001^{*}$
Music:T 24	3.370	0.578	5.836	$<0.001^{*}$

The results shows a significant correlation between original and reported phonation (z = 5.472, $p < 0.001$). But tone had a stronger effect, in that mid tones 33 were more likely to be reported as voiced than other tones (z = 14.204, $p < 0.001$); but it interacted with music perception ($z = -11.154$, $p < 0.001$). In general, listening to the musical fragments correlates with fewer reported voiced consonants (z = 2.218, $p < 0.001$). Further, compared to coronal plosives (here the reference level), labial and velar stops were more likely to be reported as voiced.

What is the nature of the interaction between tone and music? Figure 1 shows a clear difference between the tones regarding their effect on the perception of the voicing distinction under the music and non-music conditions. Under the non-music version, 71% of the stops that co-occurred with a mid tone were reported as voiced. But only a small number of the stops that co-occurred with the other tones were reported as voiced (falling: 1%, high: 9%, rising: 8%). Even more surprisingly, we observed that in the music version the pattern for mid and rising tones was reversed: only 5% of the plosives that co-occurred with mid tones were reported as voiced, whereas 70% of the stops that co-occurred with rising tones were reported as voiced.

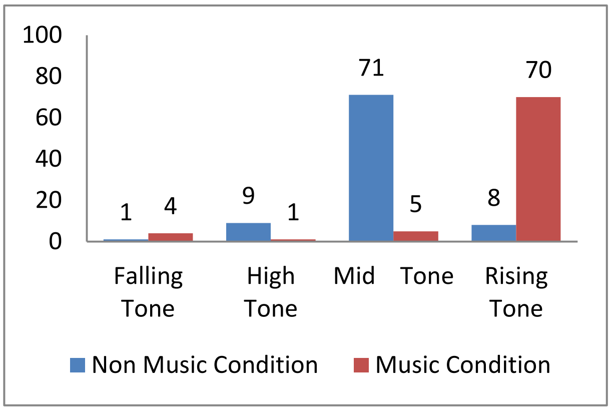

Figure 1: Percentage of stops reported as voiced, divided by tone under the music and non-music condition.

5. DISCUSSION

We investigated the perception of voicing among native Wu linguists and ran an experiment with and without alternations between linguistic and musical stimuli. In general, perception of voicing turned out to be strongly dependent on lexical tone. Under both conditions, perception of phonation was highly inaccurate, but the results were surprisingly different for the music version than for the non-music version. If syllables had a falling or a high tone, almost no voicing was reported. However, plosives that co-occurred with mid tones were most likely to be reported as voiced in the non-music version but as plain in the music version. In contrast, plosives that co-occurred with rising tones were most likely to be reported as plain in the non-music version but as voiced in the music version.

Voicing in Wu always co-occurs with low-register tones: either tones that are entirely low, or rising tones with a low onset. The fact that almost no voicing was reported for (originally voiced) plosives that co-occurred with tones that clearly belong to the high tone register (viz. 55 and 51) shows that perception of voicing relies almost fully on lexical tone.

Mid tones, in this sense, are ambiguous. Under the non-music condition, the majority of the plosives that co-occurred with mid tones were reported as voiced—even those which were originally plain. This seems to indicate that mid tones were very often regarded as low-register tones. Interestingly, in the music versions of the experiment, these mid tones were *not* perceived as low-register tones and the number of reported voiced plosives dropped dramatically. Even more surprisingly, for stops that co-occurred with rising tones we observed the opposite pattern. Under the non-music condition, the number of plosives perceived as voiced was equally low as that for high tones, but under the music condition, this was 70%. Apparently, the rising tone was considered as a low-register tone under the music condition but as a high register tone under the non-music condition. Let us speculate on the reason why this might happen.

In Shanghainese, both registers have a rising tone (34 and 13), so 24 could be perceived as ambiguous by the Wu subjects, like the mid 33 tone. The task of transcribing the plosives as aspirated, plain, or voiced is likely to convince listeners that voiced plosives do occur in the experiment. Since they co-occur with low-register tones, the question is which tone(s) are perceived as low-register ones. We think that in the non-music version, the tones 55, 24, 33, 51 were perceived as similar to Standard Mandarin, respectively 55, 35, 312, 51. Tone 3 is the only tone that could be considered as belonging to the low register. It is likely that in the music condition subjects paid more attention to the exact pitch, and in that case 24 is the only tone that starts with a low onset, thus the only one that could be considered as low-register.

6. CONCLUSION

The perception of the voicing contrast in Wu by native speakers of Wu dialects who are linguistically trained turned out to be highly inaccurate. The perception of phonation largely relied on lexical tone. If the lexical tone was perceived as a low-register tone, phonation was more likely to be perceived as voiced.

The most important finding of the present study is that a short term effect of listening to music may influence speech sound perception in a subtle and intricate manner: reclassification of particular tones in either the low or the high register, thus leading to different perception of plosive phonation.

We conclude that the interaction between linguistic and musical perception is a field which we are only beginning to understand and which also requires investigation in a much more detailed way than before.

7. ACKNOWLEDGEMENTS

This research has been supported by seed funding from the Interacting Minds Centre, Aarhus University, which is gratefully acknowledged. We are grateful to Jeroen van de Weijer for comments on a previous version of this paper.

8. REFERENCES

[1] Patel, A. D. 2010. *Music, Language, and the Brain*. Oxford: Oxford University Press.

[2] Lerdahl, F., Jackendoff, R. 1985. *A Generative Theory of Tonal Music.* MIT press.

[3] Anvari, S. H., Trainor, L. J., Woodside, J., Levy, B. A. 2002. Relations among musical skills, phonological processing, and early reading ability in preschool children. *Journal of Experimental Child Psychology* 83(2), 111–130.

[4] Schellenberg, E. G. 2004. Music lessons enhance IQ. *Psychological Science* 15(8), 511–514.

[5] Thompson, W. F., Schellenberg, E. G., Husain, G. 2004. Decoding speech prosody: Do music lessons help? *Emotion* 4(1), 46–64.

[6] Besson, M., Schön, D., Moreno, S., Santos, A., Magne, C. 2007. Influence of musical expertise and musical training on pitch processing in music and language. *Restorative Neurology and Neuroscience* 25(3), 399-410.

[7] Wong, P. C. M., Skoe, E., Russo, N. M., Dees, T., & Kraus, N. 2007. Musical experience shapes human brainstem encoding of linguistic pitch patterns. *Nature Neuroscience* 10(4), 420-422.

[8] Chobert, J., Besson, M. 2013. Musical expertise and second language learning. *Brain Sciences* 3(2), 923-940.

[9] Slevc, L. R., Miyake, A. 2006. Individual differences in second-language proficiency: Does musical ability matter? *Psychological Science* 17(8), 675-681.

[10] Hetland, L. 2000. Listening to music enhances spatial-temporal reasoning: Evidence for the" Mozart effect". *Journal of Aesthetic Education* 34(3/4), 105-148.

[11] Best, C. T., McRoberts, G. W., Goodell, E. 2001. Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system. *J. Acoust. Soc. Am*. 109(2), 775-794.

[12] Kuhl, P. K. 1992. Infants' perception and representation of speech: Development of a new theory. *Proc. ICSLP,* 449-456.

[13] Drager, K. 2010). Sociophonetic variation in speech perception. *Language and Linguistics Compass* 4(7), 473-480.

[14] Sloos. 2015. Misperception as a result of accent-induced coder bias. *Review of Cognitive Linguistics* 13(1) 59-80.

[15] Lewis, M. P. (2009). *Ethnologue: Languages of the world* (16th ed.). Dallas: SIL International.

[16] Cao, J., & Maddieson, I. (1989). An exploration of phonation types in Wu dialects of Chinese. *UCLA Working Papers in Phonetics* 72, 139-160.

[17] Chen, Z. 2010. 吴语清音浊流的声学特征及鉴定标志—以上海话为例. An acoustic study of voiceless onset followed by breathiness of Wu (吴) dialects: Based on the Shanghai (上海) dialect. *Studies in Language and Linguistics* 30(3), 20-34.

[18] Gao, J., & Hallé, P. 2012. Caractérisation acoustique des obstruantes phonologiquement voisées du dialecte de Shanghai. Acoustic properties of phonologically voiced obstruents in Shanghai dialect. *Actes De JEP-TALN-RECITAL* 145-152.

[19] Gao, J., & Hallé, P. A. 2013. Duration as a secondary cue for perception of voicing and tone in Shanghai Chinese. *Interspeech* 3157-3161.

[20] Duanmu, S. 2000. *Phonology of Chinese (Mandarin)* (2nd ed.). Oxford: Oxford University Press.

[21] Zee, E., & Maddieson, I. 1979. Tones and tone sandhi in Shanghai: Phonetic evidence and phonological analysis. *UCLA Working Papers in Phonetics* 45, 93-129.

[22] Zhang, J. 2006. The phonology of Shaoxing Chinese. PhD dissertation. Leiden University.

[23] Rose, P. 2002. Tonal complexity as conditioning Factor–More depressing Wenzhou dialect disyllabic lexical tone sandhi. *Proc. 9th Australasian International Conference on Speech Science and Technology,* 64-69.

[24] van Heuven, V. J. 1986. Some acoustic characteristics and perceptual consequences of foreign accent in Dutch spoken by Turkish immigrant workers. In J. van Oosten, & J. F. Snapper. (eds.), *Dutch linguistics at Berkeley, Dutch linguistics colloquium,* 67-84. Berkeley: The Dutch Studies Program, U. C. Berkeley.

[25] Boersma, P., Weenink, D. 2010. *Praat: Doing phonetics by computer.* [computer program]

[26] Tseng, C. 2011. Phonotactic and discourse aspects of content design in AESOP (Asian English speech cOrpus project). *Oriental COCOSDA,* 24-29.

[27] Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., et al. 2014. *Linear mixed-effects models using eigen and S4.* CRAN repository.

[28] R Development Core Team. 2009. *R: A language and environment for statistical computing.* Vienna: R Foundation for Statistical Computing.