Relative importance of tone and segments for the intelligibility of Mandarin and Cantonese

Wang, H.; Zhu, L.; Li, X.; Heuven, V.J. van; Zee E, Lee W-S

RELATIVE IMPORTANCE OF TONE AND SEGMENTS FOR THE INTELLIGIBILITY OF MANDARIN AND CANTONESE

Hongyan Wang[a], *Ligang Zhu*[a], *Xiaotong Li*[a] *& Vincent J. van Heuven*[b]

[a]Department of English, Shenzhen University, China;
[b]Phonetics Lab, Leiden University Centre for Linguistics (LUCL), the Netherlands
wanghongyan0069@hotmail.com; V.J.J.P.van.Heuven@hum.leidenuniv.nl

ABSTRACT

This study aims to establish the relative importance of segmental and word-prosodic properties for the intelligibility of spoken Mandarin and Cantonese. Mandarin has a relative small inventory of lexical tones (four) while Cantonese has a richer tone inventory (at least seven). Word prosody is normally redundant relative to segmental properties so that word recognition does not crucially depend op prosody. In poor quality speech, however, the relative importance of word prosody (stress, tone) will increase since (word) prosody is more robust against noise and distortion than segmental properties. We test two predictions. First, the importance of lexical tone will be greater in Cantonese than in Mandarin as the former has a richer tone inventory and will therefore rely more on tone for lexical contrasts than the latter, and second, that the relative importance of lexical tone will increase as segmental quality is degraded. Possibly, speech quality and tone inventory interact, if the effect of segmental degradation is superadditive to that of the size of the tone inventory.

Keywords: segmental properties, word prosody, lexical tone, Mandarin, Cantonese, intelligibility

1. INTRODUCTION

The words in a language are mainly distinguished by their segmental make up. If a strict CV language has 10 consonant phonemes and 5 vowels, it has in principle 50 different syllables. If the language allows words up to three syllables in length, the number of possible words would be on the order of 50 × 50 × 50 = 125,000, which by itself would be more than enough to fulfill the requirements of most lexicons. About half of the languages in the world have stress, i.e. the property that one syllable within the word is stronger than any of the others.[1] Stress is a so-called culminative property, which means that only one syllable within the word can be the strongest. Therefore, stress in the hypothetical language introduced above would raise the number of potentially different words by a factor 3, i.e. 375,000. The other half of the world's languages has lexical tone.[2] Ideally, and in its simplest form, in a tone language any syllable within a word can be pronounced on either a high or a low pitch. In the hypothetical language this mechanism raises the number of potential word forms to 100 × 100 × 100 = 1,000,000. The conclusion follows that lexical tone is more efficient in increasing the number of different words in a language than stress. When a language has a predilection for monosyllabic words, such as in the languages within the Sino-Tibetan family, clearly, the most efficient way to increase the size of the lexicon would be to use lexical tone.

The number of different tones in Sinitic languages varies considerably. Mandarin languages, including Putongua or Standard Chinese, has four different tones, conventionally transcribed as 55 (high level), 35 (mid rising), 214 (low dipping), and 51 (high falling). The dialect of Guangzhou (or: Cantonese) is said to have seven phonemically distinct tones, transcribed as 55 (high level, 53 (high falling), mid rising (35), mid level (33), low rising (23), low level (22) and low-falling (21) (see e.g. [7] for a comprehensive survey of tone systems in Chinese language varieties.

Stress and tones are prosodic properties, which characterize linguistic units larger than individual vowels and consonants. As a result prosodic properties vary relatively slowly over time. Also, prosody is apparent predominantly in acoustic parameters of pitch (periodicity, in Hz or semitones), loudness (weighted intensity, in Sones) and duration rather than spectral distribution of energy (e.g. formants) [3]. The latter property is characteristic of individual vowels and consonants. There seems to be a division of work between prosodic and segmental properties in speech communication. The fast-changing spectral

properties are capable of differentiating between many vowels and consonants but are vulnerable to background noise and distortion. The slowly varying prosodic properties (pitch and duration) contribute less to the differentiation of segments but are highly robust against noise and distortion. As a result the relative importance of segmental and prosodic properties to speech intelligibility varies depending on the communicative circumstances. If segments are poorly defined, e.g. due to noise in the communication channel (noise, electronic distortion, computer speech or foreign accent), the importance of prosody will increase. Earlier research [9] has shown that recognition of Mandarin tones was close to ceiling no matter what kind of filtering had been applied to the signals (whether low pass or high pass) while correct identification of segments (vowels, consonants) was severely affected. This indeed shows that tone, like other prosodic features, is a highly robust property in speech communication. When melodic properties were removed from the stimuli (using resynthesis with noise excitation or excited by a monotonised sawtooth wave), word recognition scores dropped to 24 and 16 percent, respectively; while sentence intelligibility was at 24 and 33 percent, respectively. When the sawtooth excitation was given its original melody, word and sentence scores rose to 50 and 73 percent correct; adding noise excitation (during obstruents) to the frequency-modulated sawtooth source yielded word and sentence scores of 60 and 90 percent correct.

In the present research we will compare the relative contribution of segments and word prosody (i.e. lexical tones) for the intelligibility of two Sinitic languages, i.e. Mandarin and Cantonese, which differ in the size of the tonal inventories. Moreover we will present materials in three conditions, viz. one in which the segments are optimally defined (so that the role of prosody is secondary or redundant), one in which in which segmental information has been removed from the signal (so that lexical tone is the only source of information remaining in the signal), and one intermediate condition in which the quality of the segments is degraded such that prosody (lexical tones), which is normally redundant, may provide crucial information to word recognition in spite of degraded segmental quality. Our hypothesis is that the relative importance of lexical tone will be higher in Cantonese (with its richer tone inventory) than in Mandarin, all else being equal.[3]

2. METHOD

Speech intelligibility was measured by an adapted version of the Speech Perception in Noise test (SPIN) developed for American English in 1977 by [4]. A set of 60 short, simple everyday sentences was selected and translated to Mandarin and to Cantonese. The materials were then read by one male and one female native speaker per language, speaker pairs hailing from Beijing and Guangzhou, respectively, and recorded on digital audio tape using a Shure SM10 close-talking microphone. These materials were developed as part of a larger research project on the mutual intelligibility of 15 Chinese dialects (see [5, 6]). The SPIN sentences are constructed in such a way that the sentence-final word is highly predictable if the earlier words in the sentence are correctly recognized. The listener's task is to simply write down the final word (keyword) in each sentence presented, as in *She wore her broken arm in a sling* (keyword underlined).

Each utterance was then manipulated in two ways, i.e. tonally and spectrally. In the tonal dimension the natural fundamental frequency was replaced by a constant f0 of 100 Hz, using PSOLA analysis-and-resynthesis as implemented in the Praat speech processing software [1]. This manipulation removes all melodic information from the utterances, although secondary correlates of lexical tone, such as differences in duration and intensity contour are not affected. In the second dimension, the original recordings were low-pass filtered using a digital filter with a cut-off frequency of either 1000 Hz or 300 Hz (default smoothing constant of 100 Hz. As a result of filtering at 1000 Hz, the intelligibility of the segmental information is severely reduced although most of the words remain intelligible. LP-filtering at 300 Hz obliterates all segmental information rendering the remaining utterance practically unintelligible. Systematic combination of tonal (melody versus monotone) and spectral manipulation (full bandwidth, LP-1000 Hz, LP-300 Hz) creates six stimulus conditions.

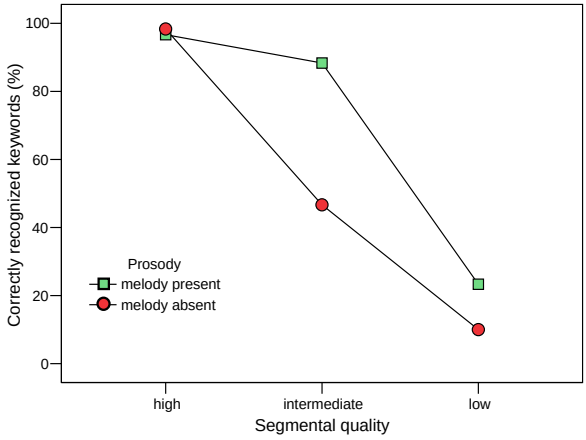
The stimuli were then divided into six blocks of ten sentences, where each set of ten were generated in one of the six different signal conditions. The 60 sentences (in six blocks of ten) were then presented to listeners in a Latin Square design such that each listener heard each sentence only once (irrespective of the signal manipulation), and that each of the 60 sentences was presented in each of

the six conditions an equal number times. In one stimulus list, half of the sentence types were spoken by the male speaker and one half by the female speaker of the language at issue; in a second stimulus list, the same sentence types were spoken by the other speaker, so that speaker sex was blocked across stimulus lists. The Mandarin sentences were presented to 24 Mandarin native listeners, while the Cantonese materials were offered to 18 native listeners of Cantonese. All listeners were students at Shenzhen University. Stimuli were played to groups of three listeners at a time, over good quality loudspeakers in a quiet room. For each language, half of the listeners were presented stimulus list 1, and the other half responded to list 2.

3. RESULTS

Figure 1 presents the results for the Mandarin part of the experiment. It shows the percentage of correctly recognized keywords in the 60 stimulus sentences, broken down by six conditions defined by the segmental degradation (unfiltered, LP-1000 and LP-300), and by the presence versus absence of f0.

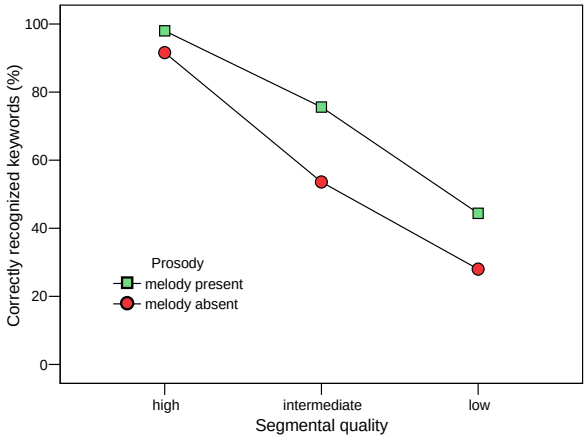
Figure 1: Percent correctly recognized keywords in 60 Mandarin SPIN sentences broken down by segmental quality and presence/absence of fundamental frequency.



When the segmental quality is optimal (no filtering), word recognition is near ceiling (98% correct), irrespective of the presence or absence of fundamental frequency information in the signal. Intelligibility is very poor when segmental information is filtered out of the signal (low segmental quality, i.e. LP filtering at 300 Hz), with a mean of 17 percent correctly recognized keywords. When the segmental quality is intermediate, due to LP filtering at 1000 Hz, the mean intelligibility is at 68%, but here the presence or absence of f0 makes a difference. When f0 is present in the signal, there is hardly any loss of intelligibility relative to the unfiltered condition (88% correct). When melodic information is removed from the signal, the percentage of correctly recognized keywords drops to 47. The main effects of segmental quality and melody are highly significant by a two-way Analysis of Variance (ANOVA), $F(2,54) = 176.6$ ($p < .001$, partial $\eta^2 = .499$) and $F(1,54) = 25.1$ ($p < .001$, partial $\eta^2 = .066$). All three segmental conditions differ from each other (Bonferroni posthoc test with $p < .05$). The interaction between the main effects is also highly significant, $F(2,54) = 12.8$ ($p < .001$, partial $\eta^2 = .067$).

The results of the Cantonese part of the experiment are presented in figure 2, which is analogous to figure 1.

Figure 2: Percent correctly recognized keywords in 60 Cantonese SPIN sentences broken down by segmental quality and presence/absence of fundamental frequency.



The results for Cantonese resemble the general configuration of the Mandarin results, but overall the effects seem weaker in Cantonese. Again, performance is near ceiling when segmental quality is high, with 98 and 92% correct for the conditions with and without f0, respectively (mean = 95%). Segmental degradation yields poorer word-recognition scores, with 76 and 54% for the intermediate, against 44 and 23% for the poorest segmental quality, for versions with and without f0 information, respectively. The main effect of segmental quality is highly significant but smaller than that obtained in the Mandarin part of the experiment, $F(2,54) = 24.1$ ($p < .001$, partial $\eta^2 = .472$). All three segmental conditions differ from each other (Bonferroni posthoc test with $p < .05$). The main effect of melody is also highly

significant but smaller than in the Mandarin data, $F(1,54) = 4.7$ ($p = .034$, partial $\eta^2 = .080$). However, the interaction between the main effects fails to reach significance, $F(1,54) < 1$.

4. CONCLUSIONS

We predicted that the contribution of word prosody would become more important when segmental quality degrades. This prediction is generally borne out by the results of our experiment. The elimination of fundamental frequency information has virtually no effect when the segmental quality is good, whether in Mandarin or in Cantonese: performance is always near ceiling. When segmental quality is degraded through low-pass filtering word recognition is better when melodic information present in the signal than when it is removed. The contribution of fundamental frequency is largest for intermediate segmental quality. It is smaller when the segmental quality is so poor that word recognition is virtually impossible, with or without prosodic information. Crucially, our second prediction was that the contribution of melodic information, i.e. lexical tone or its primary carrier fundamental frequency, should be relatively larger in Cantonese than in Mandarin, on the strength of the argument that Cantonese has a larger tone inventory than Mandarin. This prediction does not seem to be supported by our results. On the contrary, the results show that, especially, at the crucial intermediate segmental quality obtained through low-pass filtering at 1000 Hz, speech is still well understood in Mandarin when fundamental frequency is left intact in the signal, but it drops below the speech reception threshold of 50% intelligibility when fundamental frequency information is removed. The interaction between segmental quality and speech melody is smaller (and statistically absent) in the Cantonese results, so that the conclusion follows that the relative contribution of fundamental frequency does not depend on the size of the lexical tone inventory.

The present study is no more than a pilot investigation of the relative importance of lexical tone for the intelligibility of words and sentences. Future experiments will have to be conducted, on a wider variety of Sinitic and other languages, with a wider range of lexical tone inventories. Until such experiments have been done, it may be too soon to dismiss our basic hypothesis.

5. REFERENCES

[1] Boersma, P., Weenink, D. 1996. Praat, doing Phonetics by computer. *Report nr. 136*. Institute of Phonetic Sciences, University of Amsterdam.

[2] Comrie, B., Dryer, M.S., Haspelmath, M., Gil, D. (eds.). 2005. *World Atlas of Language Structures*. Oxford: Oxford University Press.

[3] van Heuven, V.J., Sluijter A.M.C. 1996. Notes on the phonetics of word prosody. In Goedemans, R., van der Hulst, H., Visch, E. (eds.), *Stress Patterns of the World, Part 1: Background*. HIL Publications vol. 2, Holland Institute of Generative Linguistics, The Hague, 233-269

[4] Kalikow, D.N., Stevens, K.N., Elliott, L.L. 1977. Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *J. Acoust. Soc. Am.* 61, 1337-1351.

[5] Tang, C. 2009. *Mutual Intelligibility of Chinese Dialects*. LOT dissertation series, Utrecht: LOT, 228.

[6] Tang, C., van Heuven, V.J. 2009. Mutual intelligibility of Chinese dialects experimentally tested. *Lingua* 119, 709-732.

[7] Yan, M.M. 2006. *Introduction to Chinese Dialectology*. LINCOM Studies in Asian Linguistics. München: LINCOM.

[8] van Zanten, E., Goedemans, R.W.N. 2007. A functional typology of Austronesian and Papuan stress systems. In van Heuven, V.J., van Zanten, E. (eds.), *Prosody in Indonesian Languages*. LOT Occasional Series 9, Utrecht: LOT, 63-88.

[9] Zhang, J.-l., Qi, S.-Q., Song, M.-Z., Liu, Q.-X. 1981. 汉语声调在言语可懂度中的重要作用 [On the important role of Chinese tones in speech intelligibility]. *声学学报* [Acta Acustica] 4, 237-241.

[1] The World Atlas of Linguistic Structures [2] lists 220 tone languages versus 307 no-tone languages (chapter 13); at the same time it lists 502 stress languages, divided in chapter 14 between 282 with fixed stress (281 in chapter 15) versus 220 with no-fixed stress (219 in chapter 15).

[2] It has been estimated that languages with stress-based word prosody, tone-based systems and languages without word prosody occur in 80, 16 and 4% of the world's languages, respectively ([8], p64).

[3] This *ceteris paribus* condition would seem to be fulfilled in the comparison of Mandarin (22 consonants, 18 vowels) and Cantonese (17 consonants, 23 vowels), see [6], pp. 208-212. The consonant and vowel inventories appear to be of equal size, although the languages may differ to some extent in their combinatorics.