Speech Prosody 2016
31 May - 3 June 2016, Boston, USA

A Comparative Study on Audiovisual Perception of Final Boundaries by Chinese and English Observers

Ran Bi [1,2,3], *Marc Swerts*[3]

[1] Nanjing University, School of Foreign Studies
[2]Jiangsu Normal University, School of Foreign Studies
[3]Tilburg University, School of Humanities, TiCC Research Center
biran1030@163.com, m.g.j.swerts@uvt.nl

Abstract

It has been suggested that conversation partners use and interpret both auditory and visual features as markers of the end of an utterance. Previous work on languages like Dutch and English have shown that speakers and listeners rely on prosodic cues such as boundary tones and eyegaze aversion to pre-signal finality in an utterance. However, little is known about how listeners of different linguistic backgrounds (Chinese and English), when perceiving utterance-finality, make use of these auditory and visual cues as used by speakers of these languages, whether these cues and their use are language-specific. Using naturally elicited stimuli from Chinese and English speakers, this study conducted a perception experiment to measure both Chinese and English participants' reaction time and accuracy in a task of judging whether a speech fragment occurred in utterance-final position or not. The participants were exposed to the same stimuli in three formats: audio-only, vision-only and audiovisual. Results revealed that audiovisual stimuli contributed most in both languages, and showed correlations between the two dependent variables (reaction time and accuracy). Additionally, English and Chinese stimuli differed in how easily and accurately they could be judged by observers from both language groups.

Index Terms: comparative study, audiovisual perception, finality, Chinese, English

1. Introduction

1.1. Audiovisual cues to boundaries

Speakers and listeners have previously been shown to exploit specific prosodic devices to provide their spoken discourse with structurally relevant information that cannot always be derived from the words or the syntax of their utterances. For instance, it has been argued that speakers use such features to indicate that they intend to finish their utterance, while listeners interpret these as signals that a speech unit is indeed about to reach its end. Evidence for the existence of such cues have often been sought through analyses of the turn-taking mechanism, especially in view of the observation that this mechanism has been reported to work remarkably well. The transition between consecutive turns of two different speakers usually proceeds very smoothly, sometimes with delays of only a few milliseconds (e.g. [1]). That suggests that speaker utterances contain features that pre-signal the upcoming final boundary, which would allow listeners to immediately take over a speaking turn. Previous work has shown that these pre-signals could be lexical or syntactic in nature [2], and they indeed could be prosodic (auditory and visual) as well. The current study presents a comparative analysis of the latter type of boundary cues in Chinese and English. In the remainder of this introduction, we discuss results of previous research in this domain, and motivate why the two languages of our investigation present potentially interesting comparative data.

Past research has suggested that these cues could be multimodal in nature. On the one hand, there is research that has shown that speakers use melodic means to signal final boundary [3,4,5,6,7], e.g. falling tone (L-L%) signals the finality of an utterance with the pragmatic meaning of certainty and definiteness [8,9,10,11,12,13,14,15]; in contrast, high or rising tone (H-H%, L-H% or H-L%) is used to differentiate question to statement with the pragmatic meaning of uncertainty, continuity (e.g., in listing) and request [8,16]. On the other hand, other studies have suggested that conversation partners also rely on visual cues, especially in the human face, i.e. eyes movement, eye brow movements and head nods, etc. [17,18,19,20]. There has been a particular interest in how patterns in mutual eye gaze may affect turn-taking. The phenomenon of gaze aversion during a speaking turn has been explained as being a consequence of a speaker's cognitive activity, in that gaze aversion is used to help thinking and avoid cognitive overload or distraction [21,22,23,24].

1.2. Crosslinguistic differences

However, few studies have explored how these two cues contribute to finality interaction. Although [25] investigated how Dutch observers perceive the end of utterance with auditory and visual cues and revealed that audiovisual modality contributes most to the detection of the end of an utterance, their study was based on a single language, namely Dutch. It is not clear whether results of that language would generalize to other languages as well, such as the tone language Chinese, and whether it is language-specific when observers perceive their native and second languages. In the current study, we are particularly interested in comparing tonal language (Chinese) and non-tonal language (English) for at least two reasons. First, Germanic languages like Dutch, English and German, have a very flexible intonation structure, in that pitch information does not serve a lexical function, and can almost exclusively be reserved for marking discourse information such as prominence and boundaries. Chinese,

10.21437/SpeechProsody.2016-173

however, is less flexible in that respect, because melodic information is largely exploited for cueing lexical information, as nicely illustrated with the use of lexical tones to distinguish lexical meanings in words that have an identical segmental structure (e.g., mā-mother, má-hemp, mǎ-horse, mà-scold). In many previous accounts of Chinese intonation, it has been argued that Chinese does not use boundary tones in the way that speakers of English or Dutch would do. Rather than varying pitch in the final part of a speech utterance, speakers of Mandarin have been argued to vary pitch register contrasts over a whole utterance domain [26,27,28,29,30,31,32]. Second, there may be cultural differences as well in eyegaze behavior and visual cues, given that is known that cultures can be markedly different in how they use patterns in eyegaze behavior for regulating the conversational interaction.

1.3. Goals of this study

Therefore, with these given motivation, we conducted a perception experiment in which Chinese and English participants were presented with the same stimuli naturally elicited from Chinese and English speakers. We aim to explore the following questions: (1) how auditory and visual cues affect finality detection in Chinese and English utterance, and whether it is language-specific; (2) how auditory and visual cues affect finality detection among observers of different languages, and whether it is language-specific when observers perceive their native and second languages.

2. Methodology

2.1. Stimuli

The stimuli for our perception experiment were taken from the audiovisual recordings of two Chinese native speakers (1 male, 1 female, average age of 20) and two English native speakers (1 male, 1 female, average age of 26) with a digital video camera when producing responses to a questioner who was asking them a series of questions in a semi-spontaneous interview. They were not informed about the purpose of the experiment.

The speakers were required to answer twenty questions in their native language (Chinese or English). The questions designed in this experiment were variants of the questions used by [25], constructed such that they would elicit target answers with variable lengths, consisting of either three or five words. Typical examples of such questions are: “Please name three ball games”, “Please list five fruits you love eating”. Then both long (2 words long) and short (1 word long) stimuli were extracted using Adobe Premiere ™. Half of the fragments had occurred in utterance-final position and half in a non-final position. Crucially, the resulting utterances would consist of simple lists of words, and care was taken that the ones chosen for our perception experiment did not contain potential syntactic or lexical cues to utterance endings (e.g., like the filler “uh” or a conjunction like “and” preceding the final word). For each speaker, two short pairs (final/non-final) and two long pairs (final/non-final) of fragments were created. Both Chinese and English fragments were stored in three ways: audio-only (AO), vision-only (VO) and audiovisual (AV). Therefore, in total 48 stimuli (Chinese and English) were created: 4 speakers × 2 lengths (short/long) × 2 types (final/non-final) × 3 conditions (AO, VO, AV). Figure 1 shows some representative stills of a Chinese and English speaker when producing a three-word utterance.

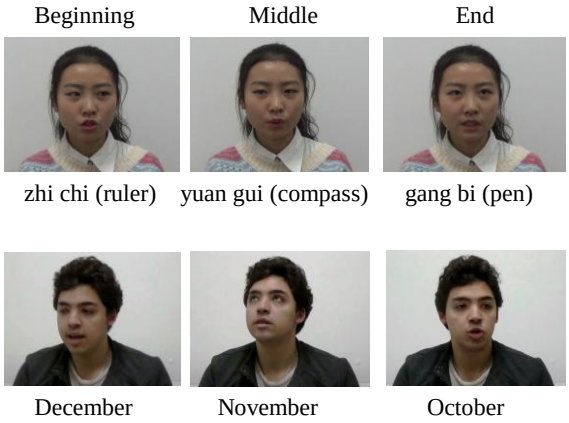

Figure 1: *Representative stills of a Chinese and English speaker producing a three-word utterance in the beginning, middle and final positions.*

2.2. Participants

Fifty-three participants (30 Chinese native speakers: 15 male and 15 female; 23 English native speakers: 13 male and 10 female) in total participated in a perception experiment on a voluntary basis. The Chinese participants (average age: 20) were university students and L2 learners of English (all of them had been studying English for more than ten years). The English participants (average age: 26) were L2 learners of Chinese (the majority of them had been studying Chinese for more than 3 years) and most of them were overseas university students in China. They all had normal or corrected-to-normal vision and good hearing. None of the participants had participated as speakers in the audiovisual recording.

2.3. Procedures

In the perception experiment, we measured two aspects of the participants’ responses: their reaction time (the speed they needed to make their classification decision) and their accuracy (the perception correct classification of utterances as being final or non-final). The participants were instructed to determine for each fragment whether it marked the end of a speaker’s utterance or not (final or non-final). Both Chinese and English participants were required to judge stimuli of their native and second languages in separate session.

Using a 3×3 Latin square design to avoid potential learning effects, the stimuli were presented to Chinese and English participants in three conditions (AO, VO, AV). Participants were told to press the corresponding buttons with labels “final” and “non-final” immediately after the display of each clip. Their reaction time and classification accuracy were measured simultaneously by *E-Prime 2.0*.

3. Results

The data of reaction time (RT) and classification (accuracy or ACC) were analysed with a repeated measures ANOVA having a 3×2×2×2×2 design with condition (three levels: audio-only, vision-only, audiovisual), length (two levels: short and long fragments) and finality (final and non-final fragments) as within-subjects factors, with language (two levels: Chinese and English) and nativeness (two levels: native Chinese

observers and native English observers) as between-subjects factors, and with the average RT and accuracy as dependent variables. *Mauchly's test* was used to check the homogeneity of variance, and the *Bonferroni* method was used for multiple pairwise comparisons for factors that had more than two levels.

3.1. Reaction Time (RT)

The ANOVA analysis revealed main effects of *language* (F (1, 102) = 4.72, $p < .05$, $\eta_p^2 = .04$), *condition* (F(2, 204) = 28.69, $p < .001$, $\eta_p^2 = .22$) and *finality* (F(1, 102) = 17.94, $p < .001$, $\eta_p^2 = .15$), while the main effects of *nativeness* and *length* were not significant.

For *language*, English stimuli received faster reaction time (M=1258.22, SD=460.88) than the Chinese ones (M=1516.53, SD=677.30). For *condition*, pairwise-comparisons using the *Bonferroni* method show that all conditions (AO, VO and AV) differed significantly from each other. The participants were fastest in the vision-only (VO) condition (M=1091.72, SD=482.60), and were slowest in the audio-only (AO) condition (M=1838.78, SD=1314.22), and the audiovisual (AV) condition (M=1231.63, SD=440.91) yielded results in between those extremes. For *finality*, participants responded faster to non-final fragments (M=1244.33, SD=521.04) than final ones (M=1530.42, SD=841.29) (see Table 1.).

The ANOVA also revealed significant two-way interactions between *finality* and *nativeness* (F(1, 102) = 4.16, $p < .05$,$\eta_p^2 = .04$), and between *condition* and *length* (F(2, 204) = 6.75, $p < .05$,$\eta_p^2 = .06$).

Table 1. ANOVA results of reaction time and accuracy.

Factor	Level	RT Mean (SD)	ACC Mean (SD)
Language	Chinese	1516.53 (677.30)	.80 (.08)
	English	1258.22 (460.88)	.84 (.09)
Condition	AO	1838.78 (1314.22)	.68 (.17)
	VO	1091.72 (482.60)	.87 (.13)
	AV	1231.63 (440.91)	.91 (.09)
Length	Short	1454.54 (567.58)	.76 (.13)
	Long	1320.21 (836.46)	.88 (.09)
Finality	Final	1530.42 (841.29)	.75 (.13)
	Non-final	1244.33 (521.04)	.89 (.09)

RT Results also suggested that observers of different languages performed differently when perceiving their native and second languages. It is interesting to find that English observers (Chinese stimuli: M=1191.11, SE=104.36; English stimuli: M=1268.58, SE=117.18) responded faster than Chinese observers (Chinese stimuli: M=1576.08, SE=102.61, English stimuli: M=1427.34, SE=117.18) in both languages (see Figure 2). Additionally, observers responded differently in their native and second languages. For Chinese observers, they responded faster in English stimuli than in Chinese stimuli, whereas for English observers, they responded slightly faster in Chinese stimuli than English stimuli. It seems that observers may think more in their native language then their second one due to the familiarity of the mother tongue.

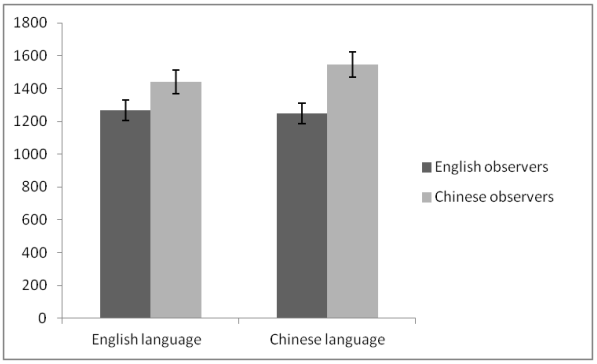

Figure 2: *Mean RT (ms) of two-way interaction between language and nativeness.*

3.2. Classification

The accuracy results were analysed with repeated measures ANOVA which had the exact same design as the one for the RT data. The analysis revealed main effects of *language* (F(1, 102) = 4.56, $p < .05$, $\eta_p^2 = .04$), *condition* (F(2, 204) = 104.44, $p < .001$, $\eta_p^2 = .51$), *length* (F(1, 102) = 71.33, $p < .001$, $\eta_p^2 = .41$) and *finality* (F(1, 102) = 71.21, $p < .001$, $\eta_p^2 = .41$). There was no significant effect of *nativeness.*

Regarding the effect of *language*, the percentage of accuracy of Chinese stimuli (M=.80, SD=.08) was lower than that of English stimuli (M=.84, SD=.09). Chinese stimuli tended to be more difficult to judge than English stimuli. Regarding the effect of *condition*, the audiovisual condition (M= .91, SD= .09) evoked the highest percentage of accuracy whereas the audio-only condition (M= .68, SD= .17) evoked the lowest percentage of accuracy, and the vision-only condition (M= .87, SD= .13) had scores in between the two conditions. Regarding the effect of *length*, it was easier to judge short fragments (M= .88, SD= .09) than long ones (M= .76, SD= .13). Regarding the effect of *finality*, the non-final fragments (M= .89, SD= .09) yielded higher accuracy than the final ones (M= .75, SD= .13).

The ANOVA also yielded three-way interaction between *condition, finality and language* (F(2, 204) = 39.37, $p < .001$, $\eta_p^2 = .28$), and between *condition, length* and *finality* (F(2, 204) = 7.03, $p < .01$,$\eta_p^2 = .06$).

Results of accuracy also indicated that observers of different linguistic background performed differently when perceiving their native and second languages. Figure 3 showed that Chinese observers performed better in judging their second language (M= .86, SE= .02) than in judging their native one (M= .80, SE= .02). English observers performed slightly better in judging their native language (M= .81, SE= .02) than in judging their second language (M= .80, SE= .02). This may suggest that it is easier to judge English finality than Chinese finality due to the language-specificity.

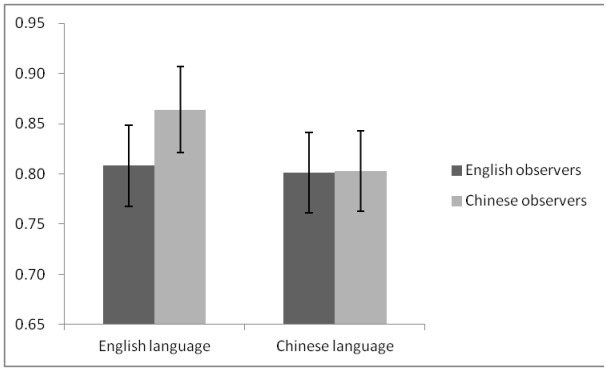

Figure 3: *Mean accuracy of two-way interaction between language and nativeness.*

4. Discussion and Conclusion

The current study was concerned with a comparative analysis of audiovisual cues to final-utterance boundaries in Chinese and English. A perception experiment was conducted in which listeners of both languages were presented with spontaneously elicited utterances of speakers of English and Chinese. The long and short fragments were cut from speakers' complete utterance that occurred in the final or non-final position. These fragments, which only contain one or two words without any lexical or syntactic cues to the utterance position, were present in three formats to listeners: audio-only (AO), vision-only (VO), audiovisual (AV). The participants were required to judge whether the perceived fragments occurred on the final or non-final position in an utterance. Both participants' reaction time (in milliseconds) and accuracy (in terms of proportion correct) were measured.

4.1. English vs Chinese

First, it is interesting to observe that we found remarkably comparable results for the Chinese and English data. In particular, it appears to be true for both languages that participants are better able to assess whether an utterance fragment was final or not, when they had access to both auditory and visual cues, rather than to only one such modality. And somewhat surprisingly, given a tradition of research which tended to be more often concerned with auditory cues, it turns out that participants for both languages, despite some differences in relative cue strength, make better use of visual than auditory features. This is intriguing as it has been argued that cultures can be markedly different in the use and function of eye gaze behavior. Also, it is remarkable that Chinese and English stimuli in the audio-only condition were about equally well classified, despite the fact that these languages exploit melodic means in a different way. As one can clearly see, the NPs in the English utterance are provided with quite different pitch patterns dependent on whether they occurred in final position (clear low boundary tone) or non-final ones (a pitch pattern which ends in the middle area of a speaker's pitch range). The Chinese NPs in comparable positions appear to differ in a more gradient manner. They all happen to have a low-ending contour in the different positions, but these phonologically similar patterns gradually decline as a function of utterance position. Even when the use of melody appears to be fundamentally different in the two languages, the observers of both languages are about equally good in judging finality, even when they have to assess the language which is not their native one. As a matter of fact, there turned out to be no significant difference in judgements (neither in terms of reaction time nor accuracy) between Chinese and English observers, suggesting that there is no in-group advantage in the way cues to finality are being perceived.

4.2. Accuracy versus reaction time

From a methodological perspective, it is interesting to note that the two dependent variables, accuracy and reaction time, are highly congruent. It appears to be true for almost all factors that a more accurate response correlates with a faster response, in line with previous claims ([33,34]). Apparently, an observer's confidence about his or her response is reflected simultaneously in a higher percentage correct and more rapid judgement. We can see that this holds for *language* (English is easier to judge than Chinese), *finality* (non-final fragments are easier than final ones), and *length* (longer fragments are easier than shorter ones, though this is only significant for accuracy). The effect of language is significant but has a very small effect size, though it does suggest that the audiovisual cues in the latter language are somewhat easier to assess. This is in line with the earlier conjecture that English may be more flexible than Chinese in how melodic means can be used for signaling finality, and that the pattern of eye gaze modulation could be more consistently used in the former language as well. The effects for the other factors are stronger: non-final fragments may be easier to classify, as they contain some marked features (deviant eye gaze, non-default intonation patterns) which clearly stand out. This result is consistent with what has previously been reported in a study on Dutch as well ([25]). The longer fragments are probably easier, simply because they potentially contain more cues, given their longer duration. The only slight difference between the measures appears to occur for the effect of condition: while both measures show that audio-only stimuli are more difficult, there is a slight difference for the other two modalities, given that vision-only data are more rapidly categorized than audiovisual ones, but the audiovisual stimuli get more accurate results than vision-only ones. The latter outcome would seem logical, again because audiovisual stimuli contain more cues than vision-only ones. The fact that the vision-only stimuli received faster responses could be due to the fact that observers are at the same time not distracted by the auditory cues, while the audiovisual stimuli may also lead to some cognitive overload.

4.3. Outlook

In the future, it would be useful to conduct a comparative study of a wider range of speakers, to see how much variation there exists between speakers of the same or a different language in how they exploit specific features for signaling finality, and to explore other sentence types as well, such as questions or commands, to see whether utterances with a different pragmatic meaning are processed differently in terms of their audiovisual cues for finality. We also conducted a perception experiment in which we used naturally elicited utterances as stimulus materials, given that we wanted to have a guarantee that these were representative of natural speaking behavior. The current study could naturally be extended by doing a more controlled set of analyses, for instance using stimuli of which audiovisual features that are carefully manipulated.

5. References

[1] S. Levinson, *Pragmatics*. Cambridge: Cambridge University Press, 1983.

[2] H. Koiso, Y. Horiucho, S. Tutiya, A. Icihikawa, and Y. Den, "An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogues," *Language and Speech*, vol. 41, no. 3-4, pp. 295-321, 1998.

[3] J. K. Burgoon, L. K. Guerrero, and K. Floyd, *Nonverbal Communication.* Pearson Education, Inc., 2010.

[4] J. R. de Pijper, and A. A. Sanderman, "On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues," *Journal of the Acoustical Society of America*, vol. 96, no. 4, pp. 2037–2047, 1994.

[5] P. Price, M. Ostendorf, S. Shattuck-Hufnagel, and S. Fong, "Theuse of prosody in syntactic disambiguation," *Journal of the Acoustical Society of America*, vol. 90, no. 6, pp. 2956–2970, 1991.

[6] M. G. J. Swerts, R. Collier, and J. Terken, "Prosodic predictors of discourse finality in spontaneous monologues," *Speech Communication*, vol. 15, no. 1-2, pp. 79-90, 1994.

[7] C. Wightman , S. Shattuck-Hufnagel, M. Ostendorf, and P. Price, "Segmental durations in the vicinity of prosodic phrase boundaries," *Journal of the Acoustical Society of America*, vol. 91, no. 3, pp. 1707–1717, 1992.

[8] C. Bartels, *The intonation of English statements and questions*. New York & London: Garland Publishing, Inc., 1999.

[9] A. Cruttenden, *Intonaiton* (2nd ed.). Beijing: Peking University Press, & London: Cambridge University Press, 2002.

[10] D. R. Ladd, "Phonological features of intonational peaks," *Language*, vol. 59, no. 4, pp. 721-759, 1983.

[11] D. R. Ladd, *Intonational phonology*. Cambridge: Cambridge University Press, 1996.

[12] J. D. O'Connor, and G. F. Arnold, *Intonation of Colloquial English*. (2nd Edition). London: Longman Group Ltd., 1973.

[13] J. B. Pierrehumbert, *The Phonology and Phonetics of English Intonation*. Ph.D Dissertation. MIT: Massachusetts Institute of Technology, 1980.

[14] J. B. Pierrehumbert, and J. Hirschberg, "The meaning of intonation contours in the interpretation of discourse," In P. R. Cohen, J. Morgen, and M. E. Pollack (Eds.), *Intentions in Communication*. MIT Press, Cambridge, pp. 271-276, 1990.

[15] J. C. Wells, *English Intonation – An Introduction*. Cambridge/New York: Cambridge University Press, 2006.

[16] N. Hedberg, J. M. Sosa, and L. Fadden, "Meanings and configurations of questions in English," *In Proceedings of the 2nd International Conference on Speech Prosody, Nara, Japan*, 2004, pp. 309-312.

[17] M. J. Doughty, "Consideration of three types of spontaneous eyeblink activity in normal humans: During reading and video display terminal use, in primary gaze, and while in conversation," *Optometry and Vision Science*, vol. 78, no. 10, pp. 712–725, 2001.

[18] P. Ekman, "About brows: Emotional and conversational signals," In M. v. Cranach, K. Foppa, W. Lepenies, and D. Ploog (Eds.), *Human ethology*. Cambridge: Cambridge University Press, pp. 169-248, 1979.

[19] E. Krahmer and M. Swerts, "More about brows: a cross-linguistic analysis-by-synthesis study," In C. Pelachaud and Z. S. Ruttkay (Eds.), *From brows to trust: Evaluating embodied conversational agents*. Kluwer Academic Publishers, pp.191-216, 2004.

[20] S. K. Maynard, "Interactional functions of a nonverbal sign: Head movement in Japanese dyadic casual conversation," *Journal of Pragmatics*, vol. 11, no. 5, pp. 589–606, 1987.

[21] M. Argyle and M. Cook, *Gaze and mutual gaze*. Cambridge: Cambridge University Press, 1976.

[22] J. K. Burgoon, L. K. Guerrero, and K. Floyd, *Nonverbal Communication.* Pearson Education, Inc., 2010.

[23] M. L. Knapp, J. A. Hall, and T. G. Horgan, *Nonverbal communication in human interaction.* (8th Edition). Wadsworth: Cengage Learning, 2013.

[24] M. Swerts and E. Krahmer, "Audiovisual prosody and feeling of knowing," *Journal of Memory and Language*, vol. 53, pp. 81-94, 2005.

[25] P. N. Barkhuysen, E. J. Krahmer, and M. G. J. Swerts, "The interplay between the auditory and visual modality for end-of-utterance detection," *The Journal of the Acoustical Society of America*, vol. 123, no. 1, pp. 354–365, 2008.

[26] A. T. Ho, "The acoustic variation of Mandarin tones," *Phonetica*, vol. 33, no. 5, pp. 353-367, 1976.

[27] A. T. Ho, "Intonation variation in a mandarin sentence for three expressions: interrogative, exclamatory and declarative," *Phonetica*, vol. 34, no. 6, pp. 446-457, 1977.

[28] O. J. Lee, *The prosody of questions in Beijing Mandarin.* Doctoral dissertation. The Ohio State University, 2005.

[29] J. Shen, "Beijinghua shengdiao de yinyu he yudiao (Pitch range of tone and intonation in Beijing dialect)," In T. Lin and L. J. Wang (Eds.), *BeijingYuyin Shiyanlu*. Beijing: Beijing University Press, pp. 73-130, 1985.

[30] X. N. S. Shen, "Tonal coarticulation in Mandatin," *Journal of Phonetics*, vol. 18, pp. 281-295, 1990.

[31] P. Shi, P, "SiZhong JuZi De YuDiao BianHua (Intonation variations in four types of Mandarin sentences)," *Yu Yan Jiao Xue Yu Yan Jiu (Language Teaching and Research)*, vol. 2, pp. 71-81, 1980.

[32] J. H. Yuan, L. S. Chih,and P. K. Greg, "Comparison of declarative and interrogative intonation in Chinese," *Proceedings of the Speech prosody 2002 conference*, Bel, B. and Marlien I. (eds). Aix-en-Provence: Laboratoire Parole et Language, pp. 711-714, 2002.

[33] A. Chen, "Reaction time as an indicator to discrete intonational contrasts in English," *Proceedings of Eurospeech 2003.* Geneva, 2003, pp. 97-100.

[34] K. Schneider, G. Dogil, and B. Möbius, "Reaction Time and Decision Difficulty in the Perception of Intonation," In *INTERSPEECH 2011*, Florence, Italy, 2011, pp. 2221-2224.