4pSCb48. Decrease of pitch perception ambiguity in tone language processing

Xiao Perdereau*

*Corresponding author’s address: Center of Chinese studies of Burgundy University, Laboratoire Interdisciplinaire Carnot de Bourgogne, UMR 6303 CNRS-Université de Bourgogne, 9 Av. A. Savary, BP 47870, F-21078, Dijon, 21078, Burgundy, France, xiao.chen-perdereau@u-bourgogne.fr

Native tone language speakers were presented with speech materials in their language produced by non-native speakers. The speech materials are streams selected according to their acoustic characteristics. They were made out of monosyllabic words, disyllabic words and polysyllabic short sentences in spoken Mandarin. Participants were required to recognize the speeches in as short time as possible. Results revealed that the minimal time needed to identify the speech is longer for monosyllabic sequences, suggesting that ambiguities reside mostly in the lexical tone level. The ambiguity due to tone perception decreases when the segmented speech events increase. Although it contributes to word meaning, pitch perception is less important in a polysyllables group of words or sentence processing than in monosyllabic word identification. We will also present some applications of these findings.
INTRODUCTION

The speech facility in human is constantly updated to adapt itself to new language environment. When we meet a nonnative speaker, we are able to handle, some times unconsciously, previously unheard nonnative accents of our language. There is a need to understand, at all levels of speech and language processing, how the sources of the variation patterns are discovered and actively learned in a natural-language processing. These studies provide us guidelines for developing specific educational programs, they provide also useful information to the automatic speech system development.

The importance of tones in tonal languages is well known, because the tones contribute to word meaning. We know numerous examples of mis-understandings when the tones are not accurately produced. But in spite of their inaccurate tones, many nonnative speakers are well understood by native speakers. The contexts play a role to the understanding, especially when people are face to face (Perdereau 2011), thanks to shared cognition and shared emotion through each interlocutor's facial expressions, gestures and so on. Our question is, in a poor language environment, without specific context, without visual and other sensory cues, are the nonnative speaker's speeches well understood? If yes, how? Are foreign accents filtered or incorporated in the recognition process? Natural language is a complex system. Most of the nonnative speakers' speeches present many individual characteristics at a time. For example, in some vowels, consonants, tones, rhythms and melodies. It is not easy to separate the contributions of many factors to the listener's understanding. We present here an experimental study of a special case, in which the speaker had "bad tones", but all other characteristics in speaking skills in Mandarin were very close to a native speaker. This case provides us with a possibility to naturally isolate complex parameters, to study the situations in which speeches with wrong tones were recognized.

EXPERIMENT

The speaker in our experiment was a 23 years old man. He speaks many foreign languages (English, Chinese and Spanish). His mother tongue is French. His competences in Chinese reading and writing are at HSK level 4. He speaks Mandarin with high fluency. If we characterize the fluency by the articulatory rate, defined as the number of syllables without silent pause per unit time, our speaker has an articulatory rate of about 9 syllables per second, based on the speech materials used in the experiment. The rate for the native speakers involved in the same experiment is in the range from 3 to 9 syllables per second, based on the same speech material. Despite of a strong foreign accent due to incorrect tones, people in contact with him understand his Mandarin very well. According to himself, he knows that there are 4 tones in Mandarin, but he gave up, at the very beginning of his studies, the effort to pronounce them correctly. As a consequence, he does not pay any more any attention to the intonation in his Mandarin speaking.

The experiment comprised two parts, a speech production task and a speech recognition task. The speech production task was made by this nonnative speaker. The speech materials were frequently used typical phrases selected from a textbook (Matinrouge 2009). The speaker was required to produce selected sequences in the following manner.
1. Monosyllabic task: the syllables were produced one by one, without any lexical or syntactical connections between successive syllables.
2. Disyllabic task: the words were all disyllabic, they were produced one by one, without connections between the successive words. Each word contained a syllable appearing in the previous monosyllabic task.
3. Polysyllabic task: the phrases were produced one by one. Each phrase contained a word appearing in the disyllabic task.

The productions have been recorded using a digital sound recorder. Data have been stored and analyzed in a computer.

For the speech recognition task, our design was based on a natural environment by omitting visual cues as for example the facial expressions and gestures which contribute to the language understanding. We recruited 3 listeners, all of them were native Mandarin speakers. They were presented with audio materials recorded previously by the nonnative speaker. The listeners were instructed to utter responses upon listening by reproducing the heard sequences. They were required to utter the most probable tones in the syllables when they met ambiguities. The responses had to be given in a shortest time. Their reproductions were recorded and analyzed using the same recorder and computer as that were used for the nonnative speaker. For the tone description, we use in this paper, the alphabetic phonetic transcription "pinyin" to note the syllables. For example, the monosyllable ba(3), marks the
sound "ba" with the third Mandarin tone. The number 1 for high level constant tone, 2 for rising tone, 3 for low falling and rising and 4 for high falling tone (Xu, 2006).

RESULTS AND DISCUSSION

For the monosyllabic task, we measured the time (t1) between the end of the nonnative speaker's syllable and the beginning of the native speaker's syllable. We found that t1 varies from 480 msec to 1920 msec. Serious recognition errors have been found in the tone identifications. For example in a task where jǐ(3) occurred, none of the 3 listeners have identified it as it was. Two listeners identified it as jǐ(4), and the third listener identified it as jǐ(1).

For the disyllabic task, we measured the time (t2) between the end of the nonnative speaker's word and the beginning of the native speaker's word. We found that for most of the frequently used disyllabic words, the response time t2 was in the range from 250 msec to 650 msec, which is clearly shorter than the range of t1 measured in the monosyllabic task. The reproduced disyllabic words were compared to the original instructed ones, the discrepancy was much rarer than that in the monosyllabic task. The disyllabic task shows a higher accuracy and shorter recognition time. The ambiguities appearing in the monosyllabic task were significantly reduced. For example, the syllable jǐ(3) was combined to form a disyllabic word jǐ(3)ge(4), which means "how many", all three listeners have correctly identified the two syllables. For the polysyllabic task, the recognition accuracy was as high as the disyllabic words recognition.

In order to understand where the ambiguities in the tone recognition reside and how their roles are reduced in a word and in a phrase, we have selected some syllables among those which were mis-identified at the monosyllabic task. We analyzed, for example, the monosyllable jǐ(3) alone, in a word and in a short phrase. These sequences were en-framed and their rendition by the nonnative speaker were analyzed using Fourier transform. We compared these power spectra with those produced by native speakers. We found that their consonants and vowels characteristics were very similar, the main discrepancy resided in the following feature: a continuous frequency shift of the fundamental harmonic frequencies, which are typically associated with the Mandarin tones. This shift is relatively large and regularly directed toward lower or higher frequencies in a native speaker's speeches, while the shift was smaller and unstable in the nonnative speaker's speech. We observed that the time duration of the continuous frequency shifts were naturally reduced in a word and in a phrase, for the non-native speaker as well as for the native speaker. So that the time left for the tone to be perceived is naturally reduced. The continuous frequency shift becomes shorter and incomplete, to leave room for the successive syllable. It seems us that this reorganization in time is a small part of the rhythm, it is observable in a spectrogram for the disyllabic word jǐ(3)ge(4) as well as for a short phrase containing jǐ(3)ge(4), produced by the nonnative speaker and by the native speakers respectively. This result shows at which step the nonnative speaker's disyllabic words and polysyllabic phrases were better recognized. This study suggests that the recognition of a word and a phrase is not a superposition of successive recognition of syllables. It seems that the rhythm, one of the characteristics of grouping, starts to play a role once polysyllables enter into a speech event. This may explain why the wrong tones have less importance in a nonnative speakers speech recognition, in condition that his speech rhythm approaches that of native speakers.

CONCLUSION

Our studies of the speech production and recognition by native and nonnative Mandarin speakers are lying between sub word and word level. Results revealed that the ambiguities resided in the lexical tone level was decreased in a disyllabic word and in a polysyllabic phrase processing. An immediate application of these studies is in the educational programs. They show ways to design adaptive teaching materials for specific nonnative speakers as learners.

This study may also be of interest for the community working in research on sub word indexing and related automatic speech recognition. To account for variations in speech production, we should have many stored examples of each word (Livescu, Folser-Lussier and Metze, 2012). The artificial speech recognition systems would be robust against variability and shifts due to speaker characteristics, e.g., nonnative accents (Baker, 2009).

ACKNOWLEDGMENTS

We thank the Burgundy University and the Conseil Régional de Bourgogne for creating and supporting the Center of Chinese Studies. We also acknowledge the Burgundy University administration for supporting the presentation of the present research results in the International Conference in Acoustics 2013.
REFERENCES