REALISATION OF THE PROSODIC STRUCTURE OF SPOKEN TELEPHONE NUMBERS BY NATIVE AND NON-NATIVE SPEAKERS OF JAPANESE

Kanae Amino[a,b] *& Takashi Osanai*[a]

[a]National Research Institute of Police Science, Japan;
[b]Japan Society for the Promotion of Science, Japan
amino@nrips.go.jp; osanai@nrips.go.jp

ABSTRACT

This paper reports the different realisations of the prosodic structure of Japanese telephone numbers in native and non-native Japanese speech. In Japanese, spoken telephone numbers have a structured prosody called bipodic template; and their accentuation is determined to occur every two digits.

Pitch contours of the spoken telephone numbers were analysed and compared among speakers of Japanese natives and Chinese and Korean learners of Japanese. The results revealed that only the native speakers followed the bipodic template structure, and non-natives showed different prosodic structures depending on their first languages. We further analysed the differences of the F0 patterns between native and non-native speakers by using cosine similarity measure. Japanese speakers' utterances had very similar F0 contours to typical Japanese, whereas those of non-natives showed less similarity and larger variance.

Keywords: bipodic template, F0 contour, L2 Japanese, foreign accent identification, forensics

1. INTRODUCTION

1.1. Foreign accent identification

Investigation of non-native accents is important for second language acquisition research as well as speech technologies such as speech recognition. Most of the second language (L2) learners wish to acquire "native-like" accents of the target language; and a speech recognition system must be able to cope with foreign users.

Investigation of foreign accents can also contribute to forensic speech science. Identifying a speaker's first language (L1) by using L2 speech, and consequently his/her nationality, is often an urgent matter in forensic situations. In order to build a foreign accent identification (FAI) system, we should be conversant with linguistic and acoustic characteristics of L1-L2 transfer.

Generally speaking, FAI systems are classified into two types: systems using segmental information [1, 2] and supra-segmental, or prosodic, information [5]. The latter method is believed to be more robust against noise. In forensics, speech data available for investigation often contain noise or other unwanted signals; therefore for forensic purposes, FAI based on prosodic information may be more preferable.

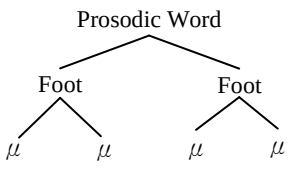
1.2. Prosody of Japanese telephone numbers

One of the linguistic forms where language-dependent prosodic structures are highlighted is spoken telephone numbers (STN) [3]. Japanese is no exception.

The Japanese language employs two-level pitch accent and mora-timed rhythm [11]. Each mora in a word is inherently associated with a specific pitch, either high or low. A word's inherent pitch pattern is then integrated into the intonation pattern of the phrase or sentence that contains that word. Furthermore, Japanese has bimoraic metrical feet. The bimoraic foot unit plays an important role in Japanese phonology in accounting for various phenomena concerning accentuation [10].

In Japanese, the linguistic forms including reduplicated mimetics, clipped words, and STN adhere to the prosodic structure based on bimoraic foot unit. This structure is called bipodic template (BT) [8-10]. As shown in Figure 1, BT consists of two bimoraic feet.

Figure 1: Structure of bipodic template; mu indicates a mora.



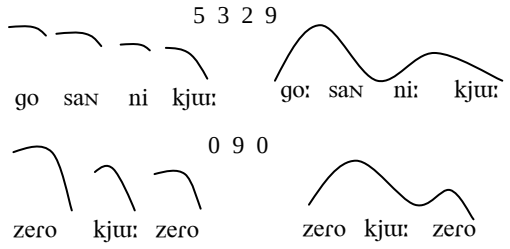
Digits in Japanese STN are read in isolation, not to be grouped as in English and other European

languages (e.g. reading "11" as "double one" or "62" as "sixty-two"). Japanese digits are either one- or two-moraic (Table 1); one-moraic digits are elongated and pronounced as two-moraic, and every two digits are phonologically grouped together to compose one BT. Accentuation occurs for every BT; accordingly, one accentual peak appears every two digits, i.e., four morae. The difference between the digit sequences and STN are depicted in Figure 2. As shown in the figure, 3-digit numbers are read as 2-1 combinations. We read the first two digits in one BT, and remaining digit in another BT [8].

Table 1: Inherent pitch-accent patterns for the digits in Tokyo Japanese [6]; H and L stand for high and low pitch accents, respectively; in STN, words with * are more ordinarily used forms for 0, 4, and 7.

Digit	Segmentals	Accent Pattern
0	/zeɾo/* or /ɾeː/	HL, HL
1	/i͡tɕi/	LH
2	/ni/	H
3	/saɴ/	*non-accented*
4	/joɴ/* or /ɕi/	HL, H
5	/go/	H
6	/ɾokɯ/	LH
7	/nana/* or /ɕi͡tɕi/	HL, LH
8	/ha͡tɕi/	LH
9	/kjɯː/	HL

Figure 2: Schematic image of F0 patterns of Japanese digit sequences (left) and STN (right); the second F0 peak is smaller than the first one due to downstep.



1.3. Objective of study

If the peculiar prosodic structure of Japanese STN is realised only by native speakers; we will be able to use it for native vs. non-native judgments. In addition, if the realisation of the said prosodic structure in STN is dependent on the speakers' L1, then we can exploit them in future for FAI. In order to confirm these two points, prosodic analyses of native and non-native speech were conducted.

2. EXPERIMENT

2.1. Speech materials

Six telephone numbers shown in Table 2 were recorded twice from 18 Japanese (9 female and 9 male), 15 Chinese (9 female and 6 male), and 10 Korean (8 female and 2 male) speakers. Information on the speakers is summarised and shown in Table 3. We recorded some balanced-bilinguals, but their data were omitted from the analyses this time.

Recordings were conducted in an anechoic room. Speech materials were recorded using PCM recorder (Marantz, PMD671) through a condenser microphone (SONY, ECM-23F5) and a telephone (Nitsuko, 2002K), simultaneously. The data were digitised at sampling frequency of 44.1 kHz with 16 bit quantisation. The speech data submitted for the analysis were the telephone numbers recorded through the telephone, downsampled at 8 kHz.

Table 2: List of telephone numbers used for analyses in this study.

Num-bering	ID	Numbers	# Syllables (Morae)
3-3-4	N1	053 574 0182	15 (20)
	N2	097 993 0312	14 (20)
3-4-4	N3	090 0978 8135	18 (22)
	N4	080 2912 6830	18 (22)
2-4-4	N5	03 3736 2319	14 (20)
	N6	06 6715 1362	17 (20)

Table 3: Information on the speakers.

L1	Gender (*N*)	Age Range (*Mean*)	Dialects (*N*)	MeanYOR[a]	Mean AOA[b]	Mean YL[c]
Japanese	Female (9)	19-48 (26)	Kanto (4), Kinki (2), Kyushu (1), Tohoku (2)	-	-	-
	Male (9)	20-26 (23)	Kanto (4), Kinki (1), Kyushu (2), Tohoku (1)	-	-	-
Chinese	Female (9)	23-46 (29)	Jin (2), Mandarin (4), Min (2), Wu (1)	3	25	8
	Male (6)	21-27 (25)	Mandarin (3), Wu (3)	2	23	3
Korean	Female (8)	19-29 (23)	Busan (4), Jinju (1) Seoul (3)	1	21	5
	Male (2)	20, 37[d]	Busan (1), Jinju (1)	1.5, 14[d]	18, 23[d]	2, 20[d]

[a]Years of Residence, [b]Age on Arrival, [c]Years of Learning Japanese, [d] we did not give average age, YOR, AOA or YL for Korean male speakers, as there were only two of them.

2.2. Analysis of F0 contours

Pitch contours for the whole telephone number were extracted in the following method. First, the fundamental frequency (F0) was calculated every 10 ms using an auto-correlation method by Praat [4]. In order to make a good comparison across different speakers and genders, F0 in Hertz (f [Hz]) was then converted to F0 in semitones (F [st]) by using the following formula:

$$F = 12 \cdot \log_2\left(\frac{f}{f_{ave}}\right), \quad (1)$$

where f_{ave} [Hz] is the average F0 of an analysed utterance.

After that, by using Manipulation-Stylise commands of Praat, we normalised the temporal property of the F0 contours by decreasing the number of the analysis points and taking two representative points per syllable. Thus we obtained F0 vectors for STN utterances whose lengths are 2 times the number of the syllables. We compared the F0 vectors among three speaker groups, by drawing F0 contours and by calculating the cosine similarity to the average Japanese contour.

2.3. Calculation of cosine similarity

In order to show quantitative differences of the F0 contours, comparison using cosine similarity was conducted. Cosine similarity S ($\boldsymbol{x}$, $\boldsymbol{y}$) of two parameter vectors $\boldsymbol{x} = \{x_1, x_2, ..., x_n\}$ and $\boldsymbol{y} = \{y_1, y_2, ..., y_n\}$ can be derived by calculating the cosine of the angle between $\boldsymbol{x}$ and $\boldsymbol{y}$ [7]. In this study, the length of the vectors n equals two times the number of syllables. The similarity S ($\boldsymbol{x}$, $\boldsymbol{y}$) is obtained through the following formula; and the resulting value ranges from -1 meaning exactly opposite, to 1 meaning exactly the same.

$$S(\boldsymbol{x}, \boldsymbol{y}) \stackrel{def}{=} \cos\theta = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{\| \boldsymbol{x} \| \| \boldsymbol{y} \|}. \quad (2)$$

We first calculated the average F0 contour of the native Japanese speakers, and then calculated cosine similarities between that and each of all utterances. When the vowel in a syllable was devoiced or got creaky and F0 could not be measured, we omitted those analysis points and decreased the number of parameter dimensions.

3. RESULTS

3.1. F0 contours

Figure 3 and 4 show the F0 contours for two different telephone numbers, N1 and N4, uttered by three groups of speakers. In both figures, the contours of Japanese natives displayed an F0 peak per two digits, reflecting the BT structure. Also for 3-digit area codes, native speakers applied BT structure, and read them as 2-1-digit combinations. On the other hand, the contours of Chinese and Korean speakers deviated from typical BT prosody. The contours for Chinese speakers exhibited more peaks than there supposed to be; and those for Korean speakers had less fluctuation than native speakers' contours.

In both Chinese and Korean contours, F0 range was smaller than native speakers'. Some speakers (mostly young returnees) produced similar F0 contours to Japanese natives, but not always and not for all numbers. In addition, no consistent tendency as to the mother dialects was observed in both Chinese and Korean speech.

3.2. Cosine similarity

The results of the evaluation using cosine similarity are presented in Figure 5. We find that cosine similarity of native Japanese speakers converge around 0.9 regardless of the telephone numbers, whereas that of the non-native speakers varied widely. The average cosine similarities and standard deviations for Japanese, Chinese, and Korean speakers were 0.90 (*S.D.* = 0.05), 0.62 (*S.D.* = 0.19), and 0.68 (*S.D.* = 0.16), respectively. Slight correlation between cosine similarity and length of residence in Japan was found with both Chinese (r = .30) and Korean (r = .54) speakers.

4. DISCUSSION AND CONCLUSIONS

We analysed F0 patterns for STN of Japanese, by visually comparing the F0 contours and by calculating the cosine similarities. The contours of Japanese native speakers showed a structure that reflects BT, whereas those of non-native speakers did not.

In Chinese speakers' contours, we can see more fluctuation than Japanese native speakers; and they seem to be synchronised with syllabic boundaries.

Figure 3: F0 contours for N1 (053 574 0182) uttered by native speakers of Japanese (top), Chinese (middle), and Korean (bottom).

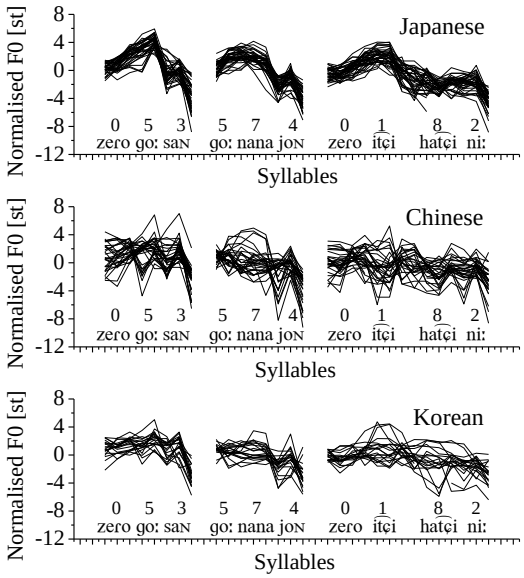


Figure 4: F0 contours for N4 (080 2912 6830) uttered by native speakers of Japanese (top), Chinese (middle), and Korean (bottom).

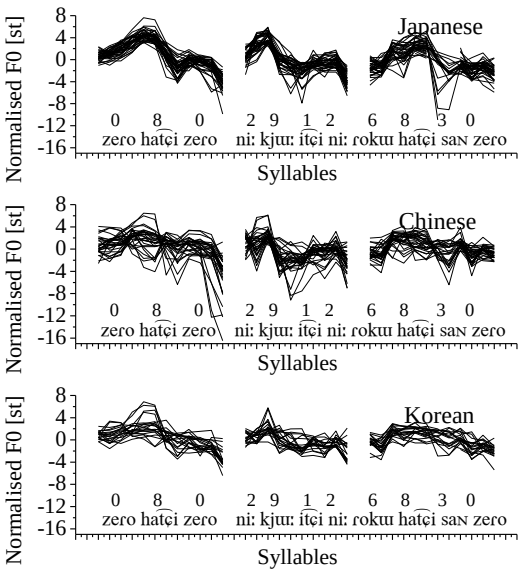
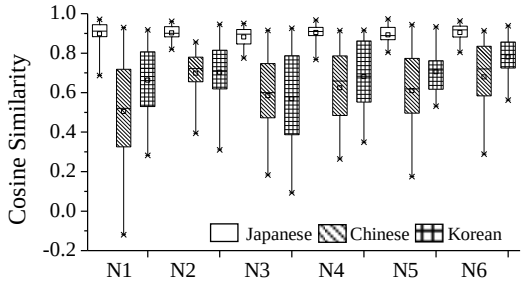


Figure 5: Average cosine similarities of each speaker group for each telephone number.



This can be considered as L1-L2 transfer of the tonal system of Chinese, i.e., the tones are associated with each syllable in Chinese. Korean speakers' utterances had very flat F0 contours. We can also see that the F0 range was much smaller in Korean utterances compared to native speakers. In order to find the reasons for the different prosodic realisations in non-native utterances, we have to examine the prosody of STN of their L1, too.

For forensic purposes, the results of this study suggest that we can exploit F0 patterns of STN as an index to judge whether a speaker is a native speaker of Japanese or not. In order to identify a speaker's L1, we will need further analyses on the prosodic structure of STN, in terms of which acoustic parameters we should use and whether and how we should normalise the temporal information. Moreover, investigations on how to determine the threshold of cosine similarity or other distance/similarity measures for native vs. non-native judgments are also necessary.

5. ACKNOWLEDGEMENTS

This work was supported by a Grant-in-Aid for JSPS Fellows (22 · 3118) and Grants-in-Aid for Scientific Research (C21510185 and B21300060).

6. REFERENCES

[1] Arslan, L.M., Hansen, J. 1996. Language accent classification in American English. *Speech Comm.* 18, 353-367.

[2] Arslan, L.M., Hansen, J. 1997. Frequency characteristics of foreign accented speech. *Proc. ICASSP* Munich, 1123-1126.

[3] Baumann, S., Trouvain, J. 2001. On the prosody of German telephone numbers. *Proc. Eurospeech* Aalborg, 557-560.

[4] Boersma, P., Weenink, D. Praat: Doing Phonetics by Computer. *http://www.praat.org/*

[5] Hansen, J., Arslan, L.M. 1995. Foreign accent classification using source generator based prosodic features. *Proc. ICASSP* Detroit, 836-839.

[6] Kindaichi, H., Akinaga, K. 2001. *Dictionary of Japanese Accents* (2nd ed.). Sanseido: Tokyo.

[7] Manning, C., Schuetze, H. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge: The MIT Press.

[8] Nasu, A. 2001. Prosodic structure of reduplicated mimetics in Japanese. *J. Osaka Univ. For. Studies* 25, 115-125.

[9] Poser, W. 1990. Evidence for foot structure in Japanese. *Language* 66, 78-105.

[10] Tsujimura, N. 1996. *An Introduction to Japanese Linguistics*. Oxford: Blackwell.

[11] Vance, T. 2008. *The Sounds of Japanese*. Cambridge: Cambridge Univ. Press.