The data in this article present position-dependent variation of focus prosody within phone number strings in Tokyo Japanese. Four acoustic parameters (duration, mean intensity, maximum pitch, and time-normalized pitch contours) are reported to illustrate focus prosody of Japanese phone numbers, separately for broad focus and corrective focus. The data also include four attached files: 1) time-normalized pitch contours for all speakers (Appendix A), 2) aggregated data of duration, mean intensity, and maximum pitch for on-focus effects (Appendix B), 3) a Python script automatically generating target stimuli (Appendix C), and 4) target stimuli used for each focus type (Appendix D). The data set can be used for several research projects including speech recognition, focus study, speaker variation in marking prosodic focus, and prosody modeling in Tokyo Japanese. Detailed discussion of data interpretation can be found in the article entitled “Prosodic focus of telephone numbers in Tokyo Japanese” (Lee et al.).

© 2019 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
1.1. Bipodic template in Tokyo Japanese

Fig. 1 illustrates how phone number strings are realized in Tokyo Japanese, conforming to its particular prosodic structure, known as a bipodic template (Amino and Osanai [2]). The line in Fig. 1 fits a time-normalized pitch contour averaged by 500 ten-digit phone number strings produced for broad focus in the format of (NNN)-(NNN)-(NNNN). In the bipodic template, every two digits join together, and an accentual peak occurs in the second digit. Thus, three-digit strings contain one accentual peak in the second digit, and the four-digit string has two accentual peaks—in the second and fourth digits.

1.2. Pitch contours for broad focus and corrective focus

Fig. 2 shows that focused digits were produced differently depending on their position within the bipodic template. This figure indicates a mismatch in focus prosody between the first and second positions within the bipodic template. The position with an accentual peak is more favorable to focus marking than the position without it. Each line in Fig. 2 is a time-normalized pitch contour averaged by five speakers, in which the red solid line refers to a digit string produced with corrective focus and the blue dotted line shows a digit string under broad focus.

1.3. Focus effects in the bipodic template

Fig. 3 demonstrates on-focus effects by duration, mean intensity, and maximum pitch in each of the first and second focus positions within the bipodic template. This figure shows that prosodic effects of focus vary by position within the bipodic template. The points represent the mean values of duration,
mean intensity, and max pitch in different focus conditions (broad, corrective), separated by focus positions (first, second) within the bipodic template.

1.4. Focus effects by digit position

Fig. 4 illustrates z-scores of duration, mean intensity, and maximum pitch in all digit positions within the phone numbers strings, separately for focus type. The on-focus effect is evident in all positions, even though certain positions are more subject to a robust on-focus effect than the other positions due to the bipodic template.

2. Experimental design, materials, and methods

2.1. Participants

Two male and three females (mean age: 28.4 years), who were naïve to the purpose of the study, participated in the production experiment. They were recruited from the National Institute for
Japanese Language and Linguistics in Tokyo and reported history of neither hearing nor speech disorders. All of the speakers were compensated 1,000 yen for their participation after the experiment.

2.2. Speech materials

We generated 100 10-digit phone number strings using a Python script (see Appendix C for details). The 100 phone number strings were set up so that each digit (0-9) occurs equally ten times in each position within a digit string and each pair of adjacent digits (e.g., 0-1, 1-2) occurs equally often in each pair of positions within a digit string. We embedded the target stimuli into a question-answer pair to elicit both broad focus and corrective focus (see Appendix D for the entire digit strings). Regarding broad focus, we used a simple yes/no question (1a) where no narrow focus was induced on any particular element of a sentence. For corrective focus, however, a particular question was asked whether a phone number string is correct and the speaker answered by correcting only one incorrect digit in the phone number string, as demonstrated in (1b). In the question-answer pairs, only the answer parts were used to extract acoustic values in the data set.

Table 1 shows the pronunciation for each digit from 0 to 9 in Tokyo Japanese. Since some digits allow two different pronunciations, the participants were asked to use the forms with an asterisk for consistency in recordings.

2.3. Recording procedures

Speech materials were recorded using a built-in microphone on a Mac laptop, in a sound-attenuated booth at the National Institute for Japanese Language and Linguistics. Participants sat comfortably before a laptop computer at a distance of about 0.5 m from the microphone. Recordings were conducted at 44.1 kHz sampling frequency and 16-bit resolution, saved directly to the computer as WAV files for acoustic analysis. Speech materials were presented visually to the participants at the center on the computer screen using PowerPoint slides.
There were two recording sessions in the experiment. The first session included stimuli for broad focus, followed by the subsequent session for corrective focus. Prior to the two sessions, a practice session in which they read three 10-digit strings that were irrelevant to the target stimuli was put to let them familiarize with the recording procedure. In the first session, they read sentences for broad focus aloud after listening to a pre-recorded prompt question, as in (1a). After a five-minute break, they read sentences for corrective focus that contain the same sequences as the ones for broad focus but were

Fig. 4. Distributions of duration, maximum pitch, and mean intensity by focus position and focus type. “1” is for the first position of a phone-number string, and “10” is for the last position.
served in a different context in such a way to correct one digit in a question, as in (1b). The participants were instructed to produce the target stimuli in a natural way, conforming to the design of each session. The experiment yielded a total of 1,000 digit strings (100 digit strings x 5 speakers x 2 focus types).

2.4. Acoustic measurements

Using a Praat script called ProsodyPro (Xu [3]), we manually marked boundaries for every digit in every digit string. From each labeled interval, we obtained the following three acoustic cues automatically generated from the script: duration in milliseconds, mean intensity in decibels, and maximum pitch in hertz. The script also computed time-normalized pitch contours in hertz at ten equidistant points per each digit in each digit string. The output of this time-normalization for a digit string (Appendix A) is 100 pitch values (= 10 equidistant x 10 digits). Because the broad-focus recordings always preceded the corrective-focus ones in the experiment, the values of duration, mean intensity, maximum pitch, and time-normalized pitch contours were converted to z-scores independently by each speaker and by each digit string in order to counterbalance the order effect and more importantly to normalize inter-speaker variations inherent in acoustic cues for focus marking (Appendix A2 and B2).

Acknowledgement

Preliminary data of this research were presented at the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) [4] and were also included in the first author’s dissertation [5].

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.dib.2019.104139.

References

[1] Y. Lee, S. Nambu, S. Cho, Focus prosody of telephone numbers in Tokyo Japanese, JASA 143 (2018) EL340–EL346.
[2] K. Amino, T. Osanai, Native vs. non-native accent identification using Japanese spoken telephone numbers, Speech Commun. 56 (2014) 70–81. https://doi.org/10.1016/j.specom.2013.07.010.
[3] Y. Xu, ProsodyPro — a Tool for large-scale systematic prosody analysis, in: Proceedings of Tools and Resources for the Analysis of Speech Prosody, 2013, pp. 7—10.

[4] Y. Lee, B. Wang, S. Chen, M. Adda-Decker, A. Amelot, S. Nambu, M. Liberman, A crosslinguistic study of prosodic focus, in: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2015), 2015, pp. 4754—4758.

[5] Y. Lee, Prosodic focus within and across languages, Ph.D. thesis, University of Pennsylvania, Philadelphia, 2015.