



One-step models in pitch perception: Experimental evidence from Japanese

Takeshi Kishiyama¹, Chuyu Huang², Yuki Hirose¹

¹Graduate School of Arts and Sciences, The University of Tokyo, Japan

²Faculty of Foreign Studies, Nagoya Gakuin University, Japan

kishiyama.t@gmail.com, huang@ngu.ac.jp, hirose@boz.c.u-tokyo.ac.jp

Abstract

Several psycholinguistic and computational models have examined the perception of *illusory vowels*, where listeners of a language insert an epenthetic vowel to repair illegal consonant clusters, perceiving VCCV as VCVCV. This study investigated whether these top-down effects can be extended to pitch patterns and induce *illusory pitches*, where a pitch was perceived on the epenthetic vowel. Tokyo and Kinki Japanese are two dialects in Japan with the same phonotactics, but Tokyo and Kinki Japanese regard LLH (low low high) and LHH as illegal tonal patterns, respectively. We used an index representing linguistic exposure to the Tokyo pitch pattern and had an AXB discrimination task to investigate whether the pitch patterns influence the perception. We found that Tokyo dialect listeners with the high index, who have long exposure to the Tokyo pitch pattern, perceived “H” pitch between L and H, whereas the subjects with the low index did not show any preference. These results indicated that pitch patterns were also used in the perception of illusory pitches and were reproduced in a simulation study.

Index Terms: pitch perception, speech perception, phonotactics, context effects, perceptual epenthesis

1. Introduction

When a speaker of a language hears speech sounds violating its phonotactics, the perception tends to be affected to repair the illegal input. For example, native speakers of Japanese perceive [ekto] as /ekuto/, inserting the *illusory vowel* because the /kt/ sequence violates its phonotactics (Fig. 1a). This phenomenon has been studied in many languages [1, 2, 3, 4], and the question in this study is whether we can extend this observation to other linguistic patterns, namely, pitches. More concretely, do Japanese speakers, who have patterns not only in phonemes but also in pitches, perceive the *illusory pitches* (Fig. 1b)?

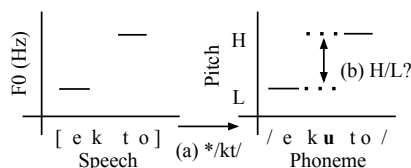


Figure 1: The left side shows physical features, and the right side shows the perceptual categories such as phonemes and their pitches, denoted by H (High) and L (Low).

Many dialects of Japanese are pitch-accented languages, and some pitch patterns are exclusive to one or the other dialect [5]. For instance, the pitch pattern for a three-mora word with a low (L) on the first mora and a high (H) on the third mora must be LHH in Tokyo and LLH in Kinki [6]. Accent information including the presence of falling pitch (HL) and its position is

lexically assigned by the word in both Tokyo dialect and Kinki dialect. Tokyo dialect assigns H to morae preceding HL except for the leftmost one, whereas Kinki dialect specifies the pitch of preceding morae with either H or L. Consequently, a trisyllabic word with three-mora starting with L can be LHL or LHH in Tokyo dialect, while it can be LHL or LLH in Kinki dialect.

If this difference of contrastive patterns in each dialect affects perception, the two groups may perceive differently towards trisyllabic words with illusory vowels without F0 values preceded by L and followed by H (Fig. 1b). Tokyo dialect speakers may perceive LHH even though the second mora is voiceless, whereas Kinki dialect speakers may perceive LLH. This study examines the question through experiments, reporting a novel illusion in pitch perception. Then, it provides computational simulations and extends the traditional models of phoneme perception to pitch perception.

This section is organized as follows. First, we review the previous studies on illusory vowels, listing the phenomena and models. Next, illusory pitches in morae without F0 are reviewed as research of the imaginary pitch, pointing out that top-down influences have not been considered yet. Finally, we introduce the dialectal differences necessary to examine the top-down effect of pitch patterns on perception and move on to an overview of the discrimination task and computational simulation.

1.1. Illusory vowels and one-step models

In vowel illusion, listeners perceive vowels to fit the speech sound to the phonotactics of their native language [1, 2, 3, 4]. Below we introduce two experiments that examined (i) the role of phonotactics in vowel illusion and (ii) the role of acoustic information in the selection of illusory vowels. We also introduce the dominant model and a probabilistic representation of the patterns to explain the results of these experiments.

First, experiments with native speakers of French and Japanese revealed the effect of phonotactics on prelexical vowel illusion [1]. French has the consonant sequence /bz/, but Japanese does not, and the speakers of each language were instructed to discriminate the sounds [ebuzo] and [ebzo] in speeded AXB tasks. French native speakers did not perceive a third vowel in [ebzo], whereas Japanese native speakers did, decreasing the discrimination accuracy between [ebuzo] and [ebzo]. These results support that phonotactics affect prelexical perception.

Second, experiments with native speakers of Brazilian Portuguese and Japanese showed the influence of acoustic cues on the selection of epenthetic vowels [2]. Both languages do not have the consonant cluster /bz/, with /i/ and /u/ being the default epenthetic vowels, respectively. In the experiments, the vowels between /b/-/z/ in [ebuzo] and [ebuzo] were removed, leaving the coarticulation cues of the vowel in /b/. The coarticulation cues affected speakers of both languages, supporting the influence of acoustic information.

Psycholinguistic models that explain phoneme activation can be classified into the two-step models [3] and the one-step models [2]. The former converts acoustic signal into phonemes, dropping acoustic cues, whereas the latter inserts phonemes checking the auditory cues. The influence of coarticulation can only be explained by the one-step models, which can refer to acoustic information of consonants at the time of insertion. The models can express the phonotactics as transition probabilities, which has been reported to influence the illusion in experiments [7] and discussed in simulations [8, 9, 10, 11].

1.2. Pitch perception in the devoiced environments

We focus on pitch perception in segments without F0, where vowel illusion occurs. Pitch perception in mora without F0 has already been studied in a different context, pitch perception of devoiced vowels in Japanese [12, 13]. In the Tokyo dialect of Japanese, high vowels surrounded by voiceless consonants tend to get devoiced. For example, the first mora of /kusa/ (grass) is realized as a devoiced vowel, yielding a CVC or sometimes even a CC [14]. Because these realizations are acoustically closer to CCs, where vowel illusion takes place, we will review pitch perception in devoiced environments to get a basis for pitch perception on illusory vowels.

When the first mora is devoiced in a two-mora word, the F0 contour of the next mora contributes to the perception of the entire pitch pattern [12, 13]. A previous study controlled the shape of the F0 contour of the mora that follows devoiced mora and investigated how native Japanese listeners recognize that two-mora word [12]. The results showed that the pitch pattern was perceived as HL when the second mora's F0 fell rapidly even though F0 values in the first mora were not available due to devoicing. Thus, the acoustic information of F0 contributes to perception in the case of devoicing.

The pitch perception in the previous study was bottom-up, referring to the acoustic information of F0, without assuming top-down processing reported in the phoneme illusion. However, if the illusion also occurs in pitch perception, it might suggest that this can be regarded as another example of the same top-down processing.

1.3. Dialect differences and present study

While bottom-up pitch perception using F0 is examined [12, 13], top-down processing in prelexical pitch perception is open for discussion. If there are dialects with different pitch patterns within the same language, it would be possible to conduct between-subjects perceptual experiments that control the pitch patterns. We will introduce the regional differences in pitch patterns, the predictions in this study, and two indices used to make the predictions.

Tokyo and Kinki dialects in Japanese differ in pitch patterns, and Tokyo speakers' exposure to pitch patterns is more limited than that of Kinki speakers. As long as 3-mora-long words are concerned, the Tokyo dialect has LHH and HLL, while LLH and HHL are absent. In contrast, the Kinki dialect has LLH, HHL, and HLL, while LHH is absent [6]. However, the speakers of Kinki dialect may also have heard the LHH pattern because the Tokyo dialect is prevalent in Japan. Therefore, Kinki dialect speakers' exposure could be gradual depending on their exposure to Tokyo Japanese.

If the pitch patterns affect pitch perception, the perception of [ekto] shown in Figure 1 should depend on exposure to Tokyo and Kinki dialects. In other words, we would expect to see an increase in the perception of LHH for Tokyo-dialect speak-

ers and LLH for Kinki-dialect speakers. This effect, however, would be gradual because the exposure to a dialect depends on the individuals. The prediction can be tested by having two indices representing gradual language experience and perceptual tendencies, which we will introduce as *Tokyo-residence ratio* and *Tokyo-pattern preference*, respectively.

First, we will calculate the Tokyo-residence ratio, an index standardized by age for the difference in residential history between Tokyo and the Kinki region, to approximate language experience. For example, if a 40-year-old subject has lived in Tokyo and the Kinki region for 25 and 15 years, respectively, the ratio is $(25 - 15)/40$, 0.25. We will first examine whether this ratio is linguistically valid enough to reproduce the following robust difference in vowel devoicing between the Tokyo and Kinki dialects [15, 16, 17]. The devoicing described in Section 1.2 tends to occur in the Tokyo dialect, while there are individual differences in the Kinki dialect [15, 16, 17]. Although the Kinki dialect speakers are less likely to devoice, younger generations of Kinki dialect speakers tend to devoice more frequently [15, 16]. We will test the validity of the Tokyo-residence ratio by seeing if this index can explain this tendency of the devoicing in the production experiment (Section 2.1).

Second, we assume that the Tokyo-pattern preference indicates a perceptual tendency, calculating it based on the results of the discrimination experiment. The left pane in Figure 2 shows some hypothetical data for explanation, and discrimination accuracies of Subject 1 and 2 between LH and LHH are lower than that of LH and LLH, indicating the subjects perceived [ekto] as LHH (Tokyo pattern). In contrast, the discrimination accuracy of LLH is lower, indicating Subject 3 perceived [ekto] as LLH. Thus, we can calculate the preference for the Tokyo pattern by subtracting the accuracy of the Tokyo dialect (LHH) from that of Kinki dialect (LLH). As shown in Figure 2 right, we would expect to see an upward trend if the Tokyo-residence ratio can explain the Tokyo-pattern preference (Section 2.2).

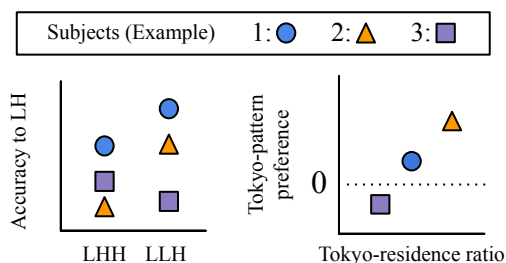


Figure 2: The left figure shows discrimination accuracy between LH and LHH/LLH from three example subjects. The right figure shows the prediction in this study.

2. Production and perception experiments

2.1. Validation of the index

This production experiment tested whether the Tokyo-residence ratio, the measure of language experience, is linguistically valid. We recruited subjects from the Tokyo and Kinki regions and aggregated the devoicing ratio for each subject. If the Tokyo-residence ratio represents the linguistic experience, it should increase the devoicing rate. To test this prediction, we made a list of words with the devoicing environment read aloud, and the devoicing ratio was calculated [16].

2.1.1. Materials and methods

We recruited the participants from Tokyo and Kinki region using an online outsourcing service¹, restricting age to 18 years and older. We had 62 subjects in the experiment, including 33 Tokyo speakers (mean age 38, $SD = 10.9$) and 29 Kinki speakers (mean age 37, $SD = 11.9$).

The word list had 8 target and 16 filler items. The target words were kutōten, akutenkō, supōtsu, esuperant, pusanmēbutsu, epuson, otsukai, tsukebarai. The underlined devoicing parts tend to be devoiced [16]. In addition, we controlled the location of the devoicing environment: the first mora or the second mora. We avoided using words with an accent nucleus (H in HL) because it could suppress devoicing [16]. The 16 filler items had /i/ or non-high vowels instead of /u/, and environment that inhibits voicing.

On an online experiment application², subjects read aloud the randomized 24 words three times. They were allowed to recheck the recorded voice and re-record it. To determine the presence or absence of devoicing in the recorded speech, the first author annotated the spectra of the target records based on the previous study [16]. On the platform, we did not control the recording equipment since the spectrogram was clear enough to annotate in the preliminary study.

2.1.2. Results and discussion

We calculated the devoicing ratio by checking the spectrogram [16]. A linear mixed-effects model [18] with the devoicing ratio as the dependent variable and Tokyo-residence ratio as the fixed effect reveals that the effect of the Tokyo-residence ratio was significant ($p < .001$). As a complementary result, using /kut/ as the baseline, the overall devoicing ratio increased for /sup/ and /tsuk/ ($p < .001$). The first mora of the devoicing environment was less likely to be devoiced ($p < .005$).

The results indicate that the Tokyo-residence ratio could reproduce the linguistic tendency of the devoicing ratio. Furthermore, the results are consistent with the previous study, which found that a subsequent fricative or affricate followed by a closure sound is more likely to be devoiced [16]. Therefore, we would assume that the Tokyo-residence ratio could represent linguistic exposure. Next, we will test whether the experience of the language reflected by the ratio affects the pitch illusion in a perceptual experiment.

2.2. Dialect differences in illusory pitches

The perceptual experiment used an AXB discrimination paradigm to test whether pitch patterns induce pitch illusion. In this experiment, different speakers recorded items for each of A, X, and B as in Dupoux et al. (1999) and Dupoux et al. (2011), which allows subjects to determine whether the pitch pattern of stimulus X is closer to A or B at a more abstract level than the acoustic information. If the pitch pattern induces illusory pitches, then the Tokyo-residence ratio should explain the Tokyo-pattern preference, which indicates the illusion tendencies to Tokyo dialect pitch.

2.2.1. Materials and methods

The experiment list had 16 target and 40 filler trials and each trial had one distractor and two references. The distractors were either the Kinki pattern (LLH/HHL) or Tokyo pattern

¹<https://crowdworks.jp/>

²<https://www.cognition.run/>

(LHH/HLL), whereas the references were from either LH or HL with illegal CCs. The mean duration of the distractors was 646 ms ($SD = 72.8$) and that of the references was 617 ms ($SD = 92.6$). The references were placed either AX or XB as correct answers. For example, an AXB could be HL, HL, and HHL if AX were the references and B was the distractor. To avoid effects from specific phonemes, we prepared two types of phonemes (esto/etsko). Thus, the target list consists of 16 trials (4 pitch patterns, 2 correct positions, and 2 types of phonemes). The filler list had 40 items unrelated to pitch pattern, and both targets and fillers were from pseudo-words. Stimuli were recorded by three trained male phoneticians who spoke Japanese as L1 (Speaker 1 and Speaker 3) and L2 (Speaker 2, JLPT-N1 level).

The same 62 subjects in the production task participated in the experiment. After three practice tasks unrelated to the pitch difference, we presented them with randomized 56 trials in the online experiment platform. Before the audio stimuli were presented, a “+” was presented at the screen center for 1000 ms, followed by an AXB sequence with 200 ms intervals. After the audio presentation, they were instructed to match the X stimulus to one of the two categories, A or B as quickly as possible. To avoid learning during the experiment, we did not give them correct/incorrect responses in the main tasks. They could start the subsequent trial by pressing the space key.

2.2.2. Results and discussion

Figure 3 shows the Tokyo-residence ratio on the x-axis and Tokyo-pattern preference on the y-axis. A linear mixed-effects model was made with Tokyo-pattern preference as the dependent variable, Tokyo-residence ratio as the fixed effects, and the items as the random effects. Random effects of participant were not included because the model did not converge. The model revealed that the effect of the Tokyo-residence ratio was significant ($p < .05$).

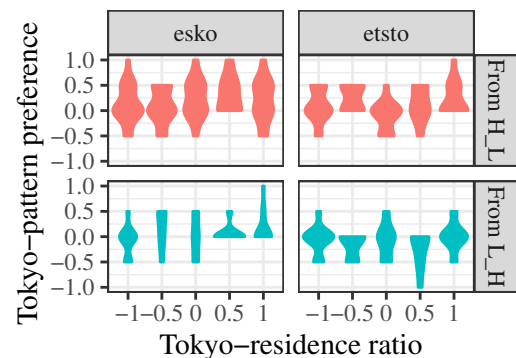


Figure 3: The probability density of the Tokyo-pattern preference as a function of Tokyo-residence ratio. The top left panel shows responses to /esko/ in the HL pitch.

The results showed that experience with pitch patterns contributes to the pitch illusion, suggesting that knowledge of pattern also plays a top-down role in pitch perception, as in the one-step model that explains the mechanism of phoneme perception. The Tokyo-pattern preference approached 1 for Tokyo-dialect speakers, while it was around 0 for Kinki-dialect speakers, not negative 1. These asymmetric results might come from the exposure of Kinki-dialect speakers to the pitch patterns of the Tokyo dialect.

3. Simulation study

The AXB discrimination task suggested that the experience of the pitch pattern induces an illusion. We will verify whether a computational model that implements a top-down process can reproduce the phenomenon in simulations and propose a baseline model that can reproduce the phenomenon.

3.1. Materials and methods

Since it is known that hidden Markov models (HMMs) are weak in modeling duration [19], this study used hidden semi-Markov models (HSMMs) with explicit duration, extending the HMMs that reproduced the vowel illusion in a previous study [11]. We created the duration models by extracting F0 every 20 ms from the training data described below and then defined the probability mass function based on the mean and standard deviation of the durations for each HH, H, LL, and L³.

To make the models represent the knowledge of pitch patterns, we defined transition probabilities (t_{mat}) and initial probability (p_i) for Tokyo and Kinki dialect speakers based on the present/absent in the dialect. Both probabilities were weighted by the Tokyo-residence ratio. The symbols used in the language models were H, HH, L, and LL. For the model without top-down processing, a uniform distribution was assumed for the probabilities. As a result, we created models with four conditions that differ with and without assuming transition and initial probabilities. The acoustic models of the HSMMs were trained based on F0 values from 24 speech data, including the three-mora distractors and the two-mora fillers. Note that the training data did not include references (LH/HL), which would induce illusions. This is because if LHH or LLH were trained with LH, the insertion could be interpreted as learning the mapping from LH to LHH/LLH. The acoustic features were extracted using the `parselmouth` package [20, 21].

We simulated the “individual differences” including the Tokyo-residence ratio, language models, pitch expression, and acoustic models. The Tokyo-residence ratio was set to one of the five levels $[-1, -0.5, 0, 0.5, 1]$ and the language models (t_{mat} and p_i) were averaged based on the Tokyo-residence ratio, with a small noise between -0.1 and 0.1 on probabilities. Plus, two types of pitch expressions were created: a relative value pattern (semitone) and an absolute value pattern (Hz). Based on the expression, the acoustic models were made with 50% of the data from the training data at random.

After creating the models for each condition, we gave the models the 12 references (HL/LH) that caused the illusion in the perceptual experiment. The references consisted of the four data presented at X and the 8 data presented at A or B in the perception experiment. Each output of the models was labeled as 1 if it belonged to the Tokyo pattern and -1 if it belonged to the Kinki pattern. The results were averaged for each model to obtain the Tokyo-pattern preference and tested the reproducibility of the perception experiment.

3.2. Results and discussion

Figure 4 shows the Tokyo-residence ratio on the x-axis and the Tokyo-pattern preference, an aggregate of inference results, on the y-axis for each model. The models with the transition probabilities are shown on the left panels, and those with the initial

probabilities are on the top. For example, the upper left shows models that use both initial probability and transition probability. Each color represents *Success rate*, the average of the cases where the inference results of the model fit the possible pitch patterns.

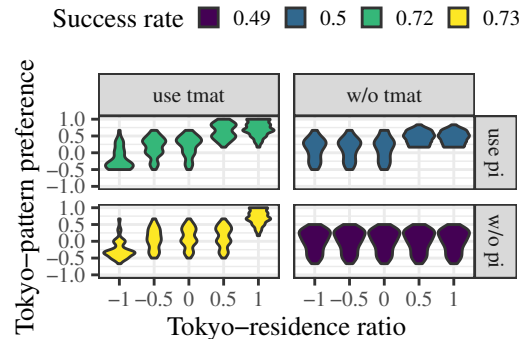


Figure 4: *The probability density of the Tokyo-pattern preference as a function of Tokyo-residence ratio. The top left panel shows responses from models with the transition and initial probabilities.*

As the bottom right of Figure 4 shows, the model without top-down processing failed to reproduce the main effect of the Tokyo-residence ratio, whereas the other models did. Furthermore, the model with only initial probability, shown in the upper right corner, resulted in a low Success rate, whereas the Success rate for the model with transition probability was high regardless of whether initial probability was used or not.

The results suggested that models using transition probabilities are closer to human perception, indicating that human pitch perception cannot be reproduced without top-down processing. Furthermore, the Success rate increased when the transition probability was used, which would indicate that the pattern between pitches contributes more to pitch perception than the initial probability.

4. General discussion

In a perceptual experiment, subjects with different pitch transition probabilities were given speech sounds, and the results indicate that the pitch pattern of the native language affects pitch perception. In the simulation, we controlled transition probabilities, and the results would support that those top-down models are necessary to reproduce human behavioral data. Thus, this study argues that, like the vowel illusion, the pitch illusion also occurs in a top-down process.

Although this study proposed a baseline model that can explain the perceptual experiment, other models that store the entire pitch pattern could explain this phenomenon. In addition, making artificial neural networks and interpret the weights could be an option to further investigate the effects of other continuous contexts to be reproduced.

5. Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers 20H01254 and 21K20027. The authors are grateful to Kei Furukawa for recording the stimuli, and Masaki Sone for recruiting the participants.

³ The data and scripts are available at the first author’s repository <https://github.com/kishiyamat/interspeech-2022-replication> and `hsmmlearn` package from <https://github.com/jvkersch/hsmmlearn> was used to make models.

6. References

- [1] E. Dupoux, K. Kakehi, Y. Hirose, C. Pallier, and J. Mehler, “Epenthetic vowels in Japanese: A perceptual illusion?” *Journal of Experimental Psychology: Human Perception and Performance*, vol. 25, pp. 1568–1578, 1999.
- [2] E. Dupoux, E. Parlato, S. Frota, Y. Hirose, and S. Peperkamp, “Where do illusory vowels come from?” *Journal of Memory and Language*, vol. 64, no. 3, pp. 199–210, 2011.
- [3] I. Berent, D. Steriade, T. Lennertz, and V. Vaknin, “What we know about what we have never heard: Evidence from perceptual illusions,” *Cognition*, vol. 104, no. 3, pp. 591–630, 2007.
- [4] B. Kabak and W. J. Idsardi, “Perceptual distortions in the adaptation of English consonant clusters: Syllable structure or consonantal contact constraints?” *Language and speech*, vol. 50, no. 1, pp. 23–52, 2007.
- [5] H. Kubozono, “Japanese phonetics and phonology,” *Iwanami Shoten*, 1999.
- [6] S. Haraguchi, “1. accent,” *The Handbook of Japanese Linguistics. Oxford: Blackwell*, 1999.
- [7] A. Kilpatrick, S. Kawahara, R. Bundgaard-Nielsen, B. Baker, and J. Fletcher, “Japanese perceptual epenthesis is modulated by transitional probability,” *Language and Speech*, pp. 1–21, 2020.
- [8] A. Guevara-Rukoz, “Decoding perceptual vowel epenthesis: Experiments & modelling,” Ph.D. dissertation, Ecole Normale Supérieure (ENS), 2018.
- [9] T. Schatz and N. H. Feldman, “Neural network vs. HMM speech recognition systems as models of human cross-linguistic phonetic perception,” in *Proceedings of the Conference on Cognitive Computational Neuroscience*, 2018.
- [10] J. Gong, M. Cooke, and M. L. G. Lecumberri, “A quantitative model of first language influence in second language consonant learning,” *Speech Communication*, vol. 69, pp. 17–30, 2015.
- [11] T. Kishiyama, “The influence of parallel processing on illusory vowels,” *Proc. Interspeech 2021*, pp. 1708–1712, 2021.
- [12] M. Sugito and H. Hirose, “Production and perception of accented devoiced vowels in Japanese,” *Annual Bulletin of Research Institute of Logopedics and Phoniatrics*, vol. 22, pp. 19–36, 1988.
- [13] K. Maekawa, “Production and perception of the accent in the consecutively devoiced syllables in Tokyo Japanese,” in *First International Conference on Spoken Language Processing*, 1990.
- [14] J. A. Shaw and S. Kawahara, “The lingual articulation of devoiced/u/in Tokyo Japanese,” *Journal of Phonetics*, vol. 66, pp. 100–119, 2018.
- [15] M. Fujimoto and S. Kiritani, “Tookyoo hoogen-to Oosaka hoogen-niokeru boin-no museika-no hikaki,” *Journal of the Phonetic Society of Japan*, vol. 7, no. 1, pp. 58–69, 2003.
- [16] H. Byun, “High vowel devoicing in the Keihanshin area of Japan,” *Journal of the Phonetic Society of Japan*, vol. 15, no. 2, pp. 23–37, 2011.
- [17] ———, “High vowel devoicing in Japanese : As an indicator of standardization of dialect,” Ph.D. dissertation, The University of Tokyo, Nov 2012.
- [18] D. Bates, M. Mächler, B. Bolker, and S. Walker, “Fitting linear mixed-effects models using lme4,” *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.
- [19] M. Ostendorf, V. V. Digalakis, and O. A. Kimball, “From HMM’s to segment models: A unified view of stochastic modeling for speech recognition,” *IEEE Transactions on speech and audio processing*, vol. 4, no. 5, pp. 360–378, 1996.
- [20] Y. Jadoul, B. Thompson, and B. de Boer, “Introducing Parselmouth: A Python interface to Praat,” *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.
- [21] P. Boersma and D. Weenink, “Praat: doing phonetics by computer [Computer program],” Version 6.1.38, retrieved 2 January 2021 <http://www.praat.org/>, 2021.