Perception of noise-vocoded sine-wave speech of Japanese pitch-accent words

Yasuaki Shinohara

View Online

Export Citation

ARTICLES YOU MAY BE INTERESTED IN

JASA
THE JOURNAL OF THE
ACOUSTICAL SOCIETY OF AMERICA

Special Issue:
Additive Manufacturing and Acoustics

Read Now!

Perception of noise-vocoded sine-wave speech of Japanese pitch-accent words

Yasuaki Shinohara[a]

Faculty of Commerce, Waseda University, 1-6-1 Nishiwaseda, Shinjuku-ku, Tokyo 169-8050, Japan

y.shinohara@waseda.jp

Abstract: The present study examined whether the identification accuracy of Japanese pitch-accent words increased after the sine-wave speech underwent noise vocoding, which eliminates the quasi-periodicity of the sine-wave speech. The results demonstrated that Japanese listeners were better at discriminating sine-wave speech than noise-vocoded sine-wave speech, with no significant difference in identification between the two conditions. They identify sine-wave pitch-accent words to some extent using acoustic cues other than the pitch accent. The noise vocoder used in the present study might not have been substantially effective for Japanese listeners to show a significant difference in the identification between the two conditions. © *2022 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).*

1. Introduction

Speech sounds convey numerous acoustic cues that people use to match with their phonological representations in their long-term memory (Dupoux *et al.*, 1997; Hallé *et al.*, 2004; Xu *et al.*, 2006). The acoustic cues in a speech sound are often redundant in recognizing words (Baer and Moore, 1993; Remez *et al.*, 1981; Shannon *et al.*, 1995; Warren *et al.*, 1995), and therefore degraded speech is intelligible to some extent (Remez and Thomas, 2013). Sine-wave speech (SWS) is a synthetic signal composed of three or four time-varying sinusoids replicating the first three or four formants of natural speech with its amplitude pattern (Remez *et al.*, 1981). People can identify vowels, words, and sentences fairly well when hearing the SWS (Hillenbrand *et al.*, 2011; Remez *et al.*, 1981; Remez *et al.*, 2011). However, as it does not convey the fundamental frequency (F0) (Chen and Fogerty, 2016; Feng *et al.*, 2012; Remez and Rubin, 1984, 1993; Rosen and Hui, 2015), SWS is less intelligible in tone languages [e.g., Mandarin (Feng *et al.*, 2012) and Cantonese (Rosen and Hui, 2015)].

Because the F0 information is not conveyed in SWS, tone-language speaking listeners might misidentify the tonal information based on the trajectory of the first formant (F1) (Feng *et al.*, 2012; Remez and Rubin, 1984, 1993). Rosen and Hui (2015) addressed this issue and found that Cantonese listeners increased their word recognition in sentences after the sine-wave speech underwent noise vocoding, which eliminates the quasi-periodicity of SWS. However, no study has investigated the effect of noise vocoding on sine-wave words contrasting with tone or pitch. The present study examined whether the identification and discrimination accuracies of Japanese pitch-accent words increased after the SWS was noise-vocoded.

Tokyo Japanese has lexical contrasts based on the bitonal [high (H) vs low (L)] pitch accent as the primary acoustic cue (Beckman and Pierrehumbert, 1986; Kitahara, 2001). For example, the word *hashi* [haɕi] with HL pitch (i.e., a relatively high-pitched first mora [ha] and relatively low-pitched second mora [ɕi]) means *chopsticks*, and the word *hashi* [haɕi] with LH pitch (i.e., a low-pitched first mora and a high-pitched second mora) means *bridge* or *edge*. Because the Japanese accent is marked by a drop of the pitch from high to low, it can make a lexical difference depending on the pitch of the following particle. The word and particle *hashi wo* [haɕi o] with LHL pitch (i.e., with the second mora [ɕi] accented) means *bridge*; *hashi wo* [haɕi o] with LHH pitch (i.e., unaccented) means *edge*. Thus, there are three-way pitch accent contrasts (i.e., the first mora accented, the second mora accented, and unaccented) in Japanese two-morae words.

In the present study, isolated two-morae words from Tokyo Japanese speakers were recorded. The naturally recorded Japanese words that lexically contrast in HL vs LH pitch were transformed into SWS, replicating the first three formants along with their amplitude patterns. Subsequently, SWS with no F0 information was noise-vocoded. These stimuli consisted of the main independent variable [i.e., natural recordings (NatRec), SWS, and noise-vocoded sine-wave speech (NzVocSWS)]. Japanese listeners' identification and discrimination accuracies were the dependent variables. Identification refers to the recognition of two-morae words contrasting with the HL vs LH pitch using the relevant

---

[a]Author to whom correspondence should be addressed.

acoustic cues. Discrimination refers to pointing out the odd auditory stimulus out of three. This does not have to include the word recognition process; thus, we tested whether the listeners could perceive the auditory difference using any acoustic cues.

Following the results of previous studies, hypotheses were established for identification and discrimination accuracies. With regard to the identification, Japanese listeners would increase their word recognition for NzVocSWS compared to that for the SWS. As the tone perception is poor for SWS (Feng *et al.*, 2012), the listeners would not be able to perceive the F0 of the original pitch-accent words in the SWS condition. However, noise vocoding would eliminate quasi-periodicity (Rosen and Hui, 2015), leading to higher identification accuracy of the pitch-accent words in the NzVocSWS condition. In contrast to the identification, noise vocoding would not increase the accuracy of SWS discrimination, although discrimination accuracy in both SWS and NzVocSWS would be lower than that in the NatRec condition. The listeners would discriminate the stimuli to some extent even in the SWS condition using any acoustic cue, and eliminating the misleading voice pitch contour would hardly affect the discrimination accuracy.

It should be noted that Japanese listeners use F0 as the primary and amplitude of mora as a secondary acoustic cue to recognize pitch-accent words. Specifically, they use the degree of the F0 fall to judge accented or unaccented words and the F0 peak location to distinguish the accented morae as the primary acoustic cue (Kitahara, 2001). They also use amplitude to identify the accented morae (Cutler and Otake, 1999). This suggests that because both sine-wave and noise-vocoded sine-wave words still include information on the amplitude envelope (Xu, 2016), Japanese listeners are able to identify and discriminate words using the secondary acoustic cue (i.e., amplitude) in both the SWS and NzVocSWS conditions to some extent.

2. Methods

2.1 Participants

Twenty Tokyo Japanese listeners (ten female and ten male) aged 19–25 years [median = 20, mean = 20.55, standard deviation (SD) = 1.67] participated in both the identification and discrimination tests at Waseda University (Tokyo, Japan). All the participants had lived in areas where the Tokyo dialect was spoken (see Akinaga and Kindaichi, 2014) for most of their lives. They had no history of speech or hearing impairments and had parents who were native Japanese speakers. They did not speak other languages in their daily life except at their university.

2.2 Stimuli and apparatus

2.2.1 NatRec

Twenty-two pitch-accent minimal pairs (44 words) listed in the Appendix were recorded from eight Tokyo Japanese speakers (four female and four male) in a sound-proof booth. Each word was randomly presented with Japanese *kanji* along with *hiragana* [e.g., 雨 (あめ)] using the ProRec 2.4 software (Huckvale, 2020) and recorded with 44,100 16-bit samples per second using a Rode NT2-A microphone, connected to a USB audio interface, Roland Rubix24. The minimal pairs with contrasting pitch accents in Tokyo Japanese were selected from a Japanese pitch-accent dictionary (Akinaga and Kindaichi, 2014). Of the 22 minimal pairs, two pairs were used for practice sessions of each task, and the remaining 20 pairs were used for each of the identification and discrimination tests. Words with a close vowel /i/ or /u/ placed between voiceless obstruents were intentionally excluded from the recording list to avoid the devoicing phenomenon (Kitahara, 2001; Kitahara and Amano, 2001). The duration between the minimal-pair words produced by each speaker was normalized, and the intensity across all tokens was normalized by the root-mean-squared method in Praat (Boersma and Weenink, 2022).

2.2.2 SWS

The naturally recorded stimuli (i.e., 44 pitch-accent words from each of the eight Japanese speakers) were transformed into sine-wave speech using a Praat script written by Darwin (2005) with the default setting. The first three formant frequencies (i.e., F1, F2, and F3) and their amplitudes were automatically tracked every 10 ms with the respective gender setting. Each of the independent sinusoids replicating the F1, F2, and F3 was then created, and they were combined to construct the sine-wave version of the original stimuli.

2.2.3 NzVocSWS

Each of the sine-wave words was noise-vocoded in Praat (Boersma and Weenink, 2022). The Praat script written by Winn (2021) was used to pass the sine-wave tokens through a noise vocoder following the settings used in Rosen and Hui (2015). Spectral analysis was conducted by a bank of 33 filters with a frequency range of 70 Hz to 10 kHz. A low-pass filter was then set at 32 Hz. Figure 1 displays the waveforms and narrowband spectrograms of the minimal-pair words, [ame] with the HL pitch (meaning *rain*) and the LH pitch (meaning *candy*), in the NatRec, SWS, and NzVocSWS conditions.
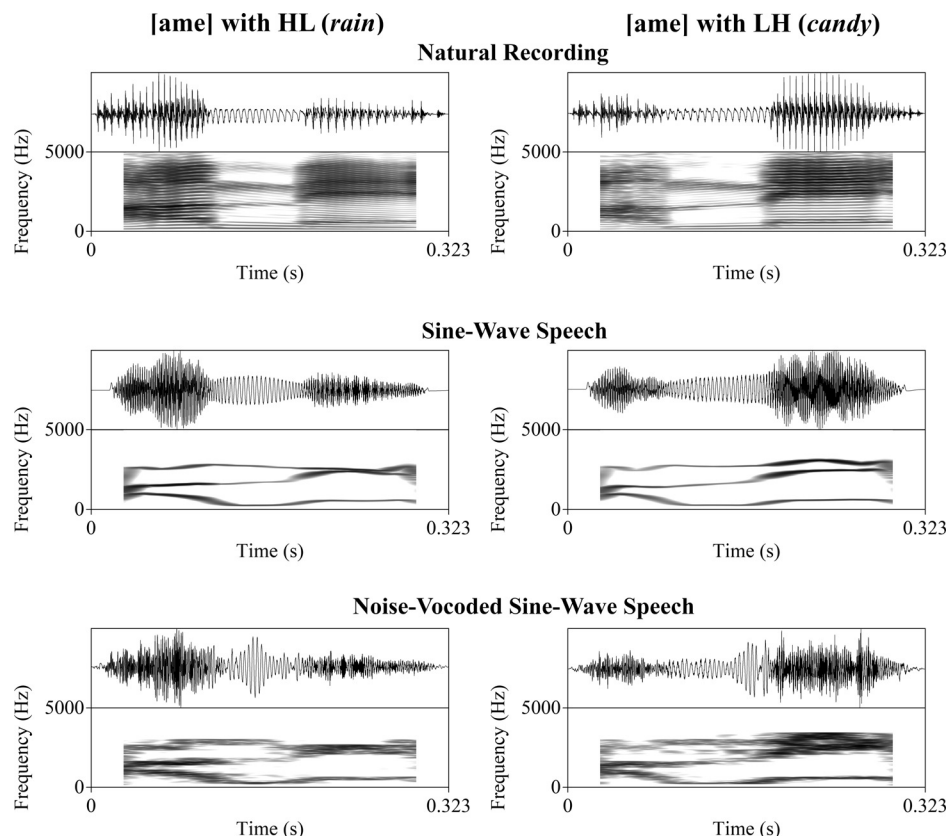
Fig. 1. Waveforms and narrowband spectrograms of the Japanese word [ame] with the HL pitch (meaning *rain*) and the LH pitch (meaning *candy*) in three conditions (NatRec, SWS, and NzVocSWS).

2.3 Procedure

2.3.1 Identification

In the identification test, one stimulus (e.g., [ame] with HL) was presented through a pair of headphones [Sennheiser (Wedemark, Germany) HD280 Pro Mk2] in a sound-proof booth, and a minimal pair [e.g., 雨 (あめ) and 飴 (あめ)] was displayed on the screen using Praat (Boersma and Weenink, 2022). Participants were instructed to click on the word that they thought they had heard. They heard each token only once with no option of hitting a replay button or feedback during the test.

The participants had a practice session with three tokens before each of the three testing conditions (NatRec, SWS, and NzVocSWS). The order of the testing conditions was counterbalanced between the participants. Half of them took the test in the order NatRec, SWS, and NzVocSWS, whereas the other half took it in the order NatRec, NzVocSWS, and SWS. Due to the large number of stimuli, two of the 20 minimal pairs produced by all eight speakers were presented to each participant. Each token was randomly played three times, resulting in 96 tokens (4 words × 8 speakers × 3 repetitions) for each condition. In total, 288 tokens were identified by each participant, and the task took approximately 16 min to complete, including time for breaks.

2.3.2 Discrimination

After the identification test, participants took a discrimination test. In the discrimination task, three stimuli from a pitch-accent minimal pair (e.g., [ame] with HL, HL, and LH) were presented through a pair of headphones (Sennheiser HD280 Pro MK2). The inter-stimulus interval (ISI) was 300 ms. Stimulus numbers (1, 2, 3) were then displayed on the screen, and participants clicked on the stimulus that sounded different from the other two. They heard each stimulus only once with no replay button, and they received no feedback regarding their responses.

Before starting the task in each of the three conditions (NatRec, SWS, and NzVocSWS), they had a practice session with three trials from the two minimal pairs that were not included in the test. The order of the conditions in the discrimination task was counterbalanced between participants, the same as the identification test (NatRec-SWS-NzVocSWS or NatRec-NzVocSWS-SWS). There were three odd stimulus positions for two minimal pairs produced by eight speakers (3 odd stimulus positions × 4 words × 8 speakers), resulting in 96 tokens presented to each participant in

each condition. In total, each participant had 288 tokens, and the task took approximately 22 min to complete, including time for breaks.

3. Results

3.1 Identification

Figure 2 displays the identification and discrimination accuracies of Japanese pitch-accent words in the three testing conditions (NatRec, SWS, and NzVocSWS). Regarding the identification accuracy, a logistic mixed-effects model was used for the correct/incorrect binomial responses. The fixed factor was the condition (NatRec, SWS, NzVocSWS), and crossed random effects for speaker and minimal pair were included. Participant was not included as a random factor due to the singular fit. The results demonstrated that the identification accuracy in the NatRec condition was significantly higher than that in the degraded-speech (SWS and NzVocSWS) conditions, $\beta = 1.22$, standard error (SE) = 0.06, $z = 21.53$, $p < 0.001$. However, there was no significant difference in the accuracy between the SWS and NzVocSWS conditions, $\beta = -0.03$, SE = 0.03, $z = -1.05$, $p > 0.05$.

3.2 Discrimination

For the discrimination accuracy analysis, a logistic mixed-effects model was used for the correct/incorrect binomial responses. The fixed factor was the condition (NatRec, SWS, NzVocSWS), and crossed random effects for speaker, minimal pair, and stimulus pair (e.g., *LH-HL-HL*, *HL-LH-LH*) were included. As Fig. 2 shows, the discrimination accuracy in the NatRec condition was significantly higher than in the degraded-speech (SWS and NzVocSWS) conditions, $\beta = 0.99$, SE = 0.05, $z = 18.39$, $p < 0.001$, and the accuracy in the SWS condition was significantly higher than in the NzVocSWS condition (SWS vs NzVocSWS), $\beta = 0.10$, SE = 0.04, $z = 2.61$, $p < 0.01$.

4. Discussion

The present study hypothesized that Japanese listeners' identification accuracy of pitch-accent words in the NzVocSWS condition would be higher than that in the SWS condition. The pitch-accent words in SWS do not convey F0 information; thus, identifying pitch-accent words would be difficult for them. However, noise vocoding would eliminate the quasi-periodicity of the sine-wave words and make the stimuli sound more speech-like (Rosen and Hui, 2015). Regarding the discrimination test, where the word recognition process is not involved, although the accuracy would be lower in the SWS and NzVocSWS than that in the NatRec condition, eliminating the misleading voice pitch contour would not significantly increase the discrimination accuracy in the NzVocSWS condition. Japanese listeners would be able to use any acoustic cues to perceive the auditory difference in both the SWS and NzVocSWS conditions.

The results did not support the hypotheses. There was no significant difference in identification between the SWS and NzVocSWS conditions. This indicates that noise vocoding did not adequately eliminate the quasi-periodicity and did not significantly increase Japanese listeners' word recognition. They might have used the amplitude envelope to some extent for identification in both the SWS and NzVocSWS conditions. Supporting this interpretation, the discrimination accuracy in the SWS condition was higher than that in the NzVocSWS condition. The listeners' auditory discrimination in the NzVocSWS condition might have been affected by the quasi-periodicity, and the additional noise might have made it more difficult for them to perceive the auditory difference between stimuli.
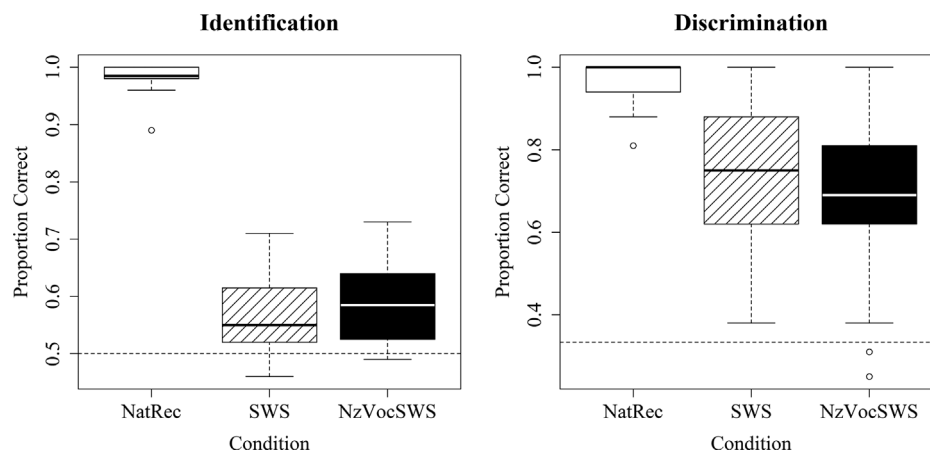

Fig. 2. Boxplots of the identification (left) and discrimination (right) accuracies of Tokyo Japanese pitch-accent minimal-pair words in the three testing conditions (NatRec, SWS, and NzVocSWS). Dashed horizontal lines represent the chance levels for each test (50.0% for identification, 33.33% for discrimination).

These results are not consistent with those of a previous study. Rosen and Hui (2015) demonstrated that speech intelligibility in the Cantonese language increased after the SWS was noise-vocoded. A possible reason for this inconsistency could be the methodological difference. Rosen and Hui (2015) did not examine the perception of tone contrast but tested the word recognition in sentences between the two conditions. The participants repeated the sentences that they heard under two conditions, meaning there was no chance to accidentally mark a correct answer. In the present study, however, participants clicked on a word that they thought they had heard with only two options; the chance level was 50%. Japanese listeners use the amplitude of each mora to identify the pitch-accent patterns (Cutler and Otake, 1999), and the amplitude envelope is still present in the SWS condition (Xu, 2016). Their identification accuracy in the SWS condition was thus already higher than the 50% chance level, and the effect of eliminating the misleading voice pitch contour may not have been sufficient to demonstrate a significant increase in the NzVocSWS condition.

One possible solution to show a significant difference in identification between the two conditions is to distort the SWS stimuli more. Rosen and Hui (2015) explained that noise vocoding makes the sine-wave tokens more speech-like. This suggests that the current distortion was not substantially effective to allow the listeners to show a significant difference between the SWS and NzVocSWS conditions. The present study used a conventional noise vocoder having the same number of analysis and synthesis filters. Instead, another way of noise vocoding (e.g., using a peak-picking vocoder with a lower number for synthesis filter; see Winn, 2021) may be necessary to adequately eliminate the effect of the SWS quasi-periodicity and to substantially increase the Japanese listeners' identification accuracy in the NzVocSWS condition. A future study should be conducted to identify the factors affecting word recognition and auditory discrimination in degraded speech.

Acknowledgments

APPENDIX

A list of Tokyo Japanese pitch-accent minimal pairs is shown in Table 1.

Table 1. The Tokyo Japanese pitch-accent minimal pairs used in the identification and discrimination tests. HL means that the first mora has a relatively high pitch and the second mora a relatively low pitch; LH means that the first mora has a relatively low pitch and the second mora a relatively high pitch.

Minimal pairs		Minimal pairs	
HL	LH	HL	LH
雨 (あめ)	飴 (あめ)	二時 (にじ)	虹 (にじ)
[ame] *rain*	[ame] *candy*	[niʑi] *two o'clock*	[niʑi] *rainbow*
腿 (もも)	桃 (もも)	朝 (あさ)	麻 (あさ)
[momo] *thigh*	[momo] *peach*	[asa] *morning*	[asa] *hemp*
降る (ふる)	振る (ふる)	組む (くむ)	汲む (くむ)
[huɾu] *fall*	[huɾu] *shake*	[kumu] *team up*	[kumu] *scoop out*
飼う (かう)	買う (かう)	春 (はる)	張る (はる)
[kau] *keep (a pet)*	[kau] *buy*	[haɾu] *spring (season)*	[haɾu] *stretch*
神 (かみ)	紙 (かみ)	病む (やむ)	止む (やむ)
[kami] *God*	[kami] *paper*	[jamu] *fall ill*	[jamu] *stop (raining)*
白 (しろ)	城 (しろ)	練る (ねる)	寝る (ねる)
[ɕiɾo] *white*	[ɕiɾo] *castle*	[neɾu] *knead*	[neɾu] *sleep*
夜 (よる)	寄る (よる)	隅 (すみ)	墨 (すみ)
[joɾu] *night*	[joɾu] *drop by*	[sumi] *corner*	[sumi] *ink*
狩り (かり)	借り (かり)	膿む (うむ)	産む (うむ)
[kaɾi] *hunting*	[kaɾi] *debt*	[umu] *fester*	[umu] *give birth*
赤 (あか)	垢 (あか)	切る (きる)	着る (きる)
[aka] *red*	[aka] *dirt*	[kiɾu] *cut*	[kiɾu] *wear*
汁 (しる)	知る (しる)	維持 (いじ)	意地 (いじ)
[ɕiɾu] *soup*	[ɕiɾu] *know*	[iʑi] *maintenance*	[iʑi] *pride*
琴 (こと)	事 (こと)	経る (へる)	減る (へる)
[koto] *Japanese harp*[a]	[koto] *thing*[a]	[heɾu] *pass*[a]	[heɾu] *decrease*[a]

[a]Minimal-pair words used in the practice sessions.

References and links

Akinaga, K., and Kindaichi, H. (**2014**). *Shin Meikai Nihongo Akusento Jiten*, 2nd ed. (Sanseido, Tokyo).

Baer, T., and Moore, B. C. J. (**1993**). "Effects of spectral smearing on the intelligibility of sentences in noise," J. Acoust. Soc. Am. **94**(3), 1229–1241.

Beckman, M. E., and Pierrehumbert, J. B. (**1986**). "Intonational structure in Japanese and English," Phonol. Yearb. **3**, 255–309.

Boersma, P., and Weenink, D. (**2022**). "Praat: Doing phonetics by computer (version 6.2.09) [computer program]," http://www.praat.org/ (Last viewed 4/12/2022).

Chen, F., and Fogerty, D. (**2016**). "Factors affecting the intelligibility of sine-wave speech," in *Proceedings of Interspeech 2016*, San Francisco, CA, September 8–12, pp. 1692–1695.

Cutler, A., and Otake, T. (**1999**). "Pitch accent in spoken-word recognition in Japanese," J. Acoust. Soc. Am. **105**(3), 1877–1888.

Darwin, C. (**2005**). "SWS," http://www.lifesci.sussex.ac.uk/home/Chris_Darwin/Praatscripts/SWS (Last viewed 4/12/2022).

Dupoux, E., Pallier, C., Sebastian, N., and Mehler, J. (**1997**). "A destressing 'deafness' in French?," J. Mem. Lang **36**(3), 406–421.

Feng, Y.-M., Xu, L., Zhou, N., Yang, G., and Yin, S.-K. (**2012**). "Sine-wave speech recognition in a tonal language," J. Acoust. Soc. Am. **131**(2), EL133–EL138.

Hallé, P. A., Chang, Y.-C., and Best, C. T. (**2004**). "Identification and discrimination of Mandarin Chinese tones by Mandarin Chinese vs. French listeners," J. Phon. **32**(3), 395–421.

Hillenbrand, J. M., Clark, M. J., and Baer, C. A. (**2011**). "Perception of sinewave vowels," J. Acoust. Soc. Am. **129**(6), 3991–4000.

Huckvale, M. (**2020**). "ProRec: Speech prompt and record system (version 2.4) [computer program]," https://www.phon.ucl.ac.uk/resource/prorec/ (Last viewed 4/12/2022).

Kitahara, M. (**2001**). "Category structure and function of pitch accent in Tokyo Japanese," Ph.D. dissertation, Indiana University at Bloomington, Bloomington, IN.

Kitahara, M., and Amano, S. (**2001**). "Perception of pitch accent categories in Tokyo Japanese," Gengo Kenkyu **120**, 1–34.

Remez, R. E., Dubowski, K. R., Broder, R. S., Davids, M. L., Grossman, Y. S., Moskalenko, M., Pardo, J. S., and Hasbun, S. M. (**2011**). "Auditory-phonetic projection and lexical structure in the recognition of sine-wave words," J. Exp. Psychol. Hum. **37**(3), 968–977.

Remez, R. E., and Rubin, P. E. (**1984**). "On the perception of intonation from sinusoidal sentences," Percept. Psychophys. **35**(5), 429–440.

Remez, R. E., and Rubin, P. E. (**1993**). "On the intonation of sinusoidal sentences: Contour and pitch height," J. Acoust. Soc. Am. **94**(4), 1983–1988.

Remez, R. E., Rubin, P. E., Pisoni, D. B., and Carrell, T. D. (**1981**). "Speech perception without traditional speech cues," Science **212**(4497), 947–950.

Remez, R. E., and Thomas, E. F. (**2013**). "Early recognition of speech," WIREs. Cogn. Sci. **4**(2), 213–223.

Rosen, S., and Hui, S. N. C. (**2015**). "Sine-wave and noise-vocoded sine-wave speech in a tone language: Acoustic details matter," J. Acoust. Soc. Am. **138**(6), 3698–3702.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**(5234), 303–304.

Warren, R. M., Riener, K. R., Bashford, J. A., and Brubaker, B. S. (**1995**). "Spectral redundancy: Intelligibility of sentences heard through narrow spectral slits," Percept. Psychophys. **57**(2), 175–182.

Winn, M. (**2021**). "Vocoder [Praat script] (No. 45)," http://www.mattwinn.com/praat/vocode_all_selected_v45.txt (Last viewed 4/12/2022).

Xu, L. (**2016**). "Temporal envelopes in sine-wave speech recognition," in *Proceedings of Interspeech 2016*, San Francisco, CA, September 8–12, pp. 1682–1686.

Xu, Y., Gandour, J. T., and Francis, A. L. (**2006**). "Effects of language experience and stimulus complexity on the categorical perception of pitch direction," J. Acoust. Soc. Am. **120**(2), 1063–1074.