9th International Conference on Speech Prosody 2018
13-16 June 2018, Poznań, Poland

Exploratory study in ethnophonetics: Comparison of cross-cultural perceptions of Japanese cake seller voices among Japanese, Chinese and American English listeners

Donna Erickson[1] • *Toshiyuki Sadanobu*[2] • *Chunyue Zhu*[3] • *Kerrie Obert*[4] • *Hayato Daikuhara*[5]
[1]Haskins Laboratories, U.S.A., Kanazawa Medical University, Japan
[2]Kyoto University, Japan
[3]Kobe University, Japan
[4]The Ohio State University, U.S.A.
[5]Renmin University of China, China

Ericksondonna2000@gmail.com, sadanobu.toshiyuki.3x@kyoto-u.ac.jp, shu_s_y@koala.kobe-u.ac.jp,kerriebobert@gmail.com, daikuhayato@yahoo.co.jp

Abstract

This study examines how ethnophonetic sounds are perceived in three different language/cultural groups. Specifically, Japanese, Chinese and American listeners were asked to listen to samples of voices of Japanese cake-selling street voices, and to rate which voice was the "best". The results indicate Japanese listeners are quite sensitive to what voice is best as a seller of fashionable Western cakes, and that this voice is different from sellers in less fashionable stores. The non-Japanese listeners rated the experienced Japanese cake-street seller voice considerably lower than did the Japanese listeners; moreover, Chinese and American listeners' differed on which street-seller voice they preferred. Tentative analysis suggests that Chinese listeners preferred a street selling voice with a higher F0, one that sounds like the *moe* anime voice, while American listeners preferred the voice with a more dynamic range of F0. Japanese listeners, on the other hand, preferred the voice that sounded "more elegant"—one with a touch of twang and some breathiness, a voice quality that is often perceived as being nasal (*hana ni kakatta koe*). An interesting question to be explored in the future is why the same voice is interpreted differently in different cultures.

.

Index Terms: ethonophonetics, Japanese cake-street sellers F0, twang, cross-cultural perceptions

1. Introduction

In daily life various ways of phonation of speech sounds are encountered, some of which may sound strange to listeners belonging to another language or culture. Peoples' impression of a sound can vary according to their native language. One source of the differences in impressions may be a semantic one. Gumperz [1], for example, argues that cross-cultural trouble between British cargo handlers and the Indian and Pakistani staff cafeteria workers was based on the semantic difference between British English and Indian English of falling intonation. Another example is from Erickson and Maekawa [2], who pointed out that an American speaker can express the attitude of admiration in terms of an intonation consisting of low pitch followed by a rising pitch, whereas a similar intonation conveys the attitudinal meaning of doubt in Japanese society.

In addition to semantic factors of the sound, can the sound itself also cause differences in impressions across languages? One approach for addressing this is through the study of *ethnophonetics*, the study of choreographed movements of articulation and voice as they relate to certain cultural settings to convey phonetic information within the framework of a particular society. The term is a newly-coined term introduced by [3,4]. It is analogous with the term *ethnosyntax* described by Enfield [5:3-4], who broadly defines *ethnosyntax* as "the study of connections between the cultural knowledge, attitudes, and practices of speakers, and the morphosyntactic resources they employ in speech.... This field of research asks not just how culture and grammar may be connected, but also how they may be interconstitutive, though overlap and interplay between people's cultural practices and preoccupations and the grammatical structures they habitually employ. This may make reference to the semantics of grammar, or to ways in which the use of productivity of grammatical resources are constrained or licensed by culture."

Some examples of ethnophonetic expressions in Japanese society include the adolescent, innocent archetype voice of the *moe* characters in anime; Yoku Hata's comedic rendition of switching from an ordinary male voice to that of a high-pitched female; the classic Kyogen voice heard in a popular comedy; and the voices of those heard selling goods in a noisy market. Examples of a number of these ethnosounds, including cake sellers' voices, can be found at http://www.speech-data.jp/sadanobu_book/ethnophonetics/.

In this paper, we examine cake-seller speech sounds of three young Japanese female speakers, as produced in fashionable cake shops. The "Cake Street Seller's Voice" is heard in some Japanese young girls, especially selling something fashionable, such as western-style cakes. In Japan as in many countries, but not usually in the United States, sales people call out in a loud voice for shoppers to come buy their products. Sometimes they stand on the street in front of the shop, sometimes they stand behind their counter inside their shop; we refer to both these voices as "street seller voices." The cake-seller voice can never be heard at a drug store or electronic store, or even *wagashi* (i.e. Japanese cake) shops, since *wagashi* is not fashionable, unlike western cake which for Japanese people, is.

Previous work reported on perception of the cake sellers voices by Japanese listeners [3,4] and Chinese listeners [6]. In this study, we examine the effect of culture/language on the

10.21437/SpeechProsody.2018-80

perception of these voices; specifically, the differences in perception by Japanese, Chinese and American listeners of the cake sellers' voices. We suggest that these differences can be explained by the multi-facetedness of sound; specifically, that different cultures perhaps cue into different acoustic features of the sounds.

2. Methods

Acoustic recordings were made of three females using cake seller voices (MM, MT, and MY). MM and MY actually had experience selling cakes, while MT had been working for 8 months as a supermarket cashier. At the time of the MRI experiment (2013 Nov.22 for MM and 2014 Jan.6 for MY), MM had been selling cakes for a famous cake shop at a department store for three years, and MY for two and a half years. MM said she was not given any instructions for working at the cake shop other than very general ones like "use a bright and cheerful voice", but, in addition, she said she imitated her senior's way of speaking for selling cakes. Also, she continues to sell cakes at the same department store, even after graduating from university.

The sentences recorded were *irasshaimase, douzo gorankudasaimase* ('Welcome! Please enjoy looking!') and were spoken in both a cake street selling voice and a "usual" voice (therefore six voices in total: MM_S, MM_U, MT_S, MT_U, MY_S, and MY_U). The recordings were made with an Optoacoustics Optimic 1140 microphone and Marantz PMD671 recorder at ATR-BAIC, Kyoto, Japan.

To elicit perceptions about the cake seller voices, naïve subjects (Japanese, Chinese and Americans) were asked to listen to the 6 voices (3 cake seller voices and 3 "usual" voices) and then answer two questions: (1) which voice is the "best" (i.e., "most fit" voice) for a cake seller? Please rate on a scale of 1 (lowest) to 5 (highest), and (2) what are your opinions/impressions about these voices as cake-selling voices?

The Japanese listeners were all undergraduate students in the Kansai area. For the evaluation question, N=110; for the impression question, N=58~105. The Chinese listeners were undergraduate students studying Japanese at a university in Beijing, China (Beginning: 7, Intermediate: 15, and Advanced: 16). For the evaluation question, N=38; for the impression question, N=28-35.

The American listeners were 20 students at a university in northern California, none of whom had studied Japanese. The average age was 29, with an age range of 13 to 51; most of them were monolingual, but 3 were bilingual (Spanish, Arabic, Hmong). Before giving the questionnaires, a brief powerpoint introduction with illustrations of cake sellers was given in order to familiarize the students with Japanese cake sellers.

For the acoustic analysis, Wavesurfer software (Ver. 8.5.8) was used to measure peak, minimum and range of F0 and also to examine some spectral characteristics of the street seller voices.

3. Results

3.2. Perception results

Table 1 shows the mean ratings of the six speech samples by Japanese, Chinese and American listeners. All language groups share the same tendency to give higher rates to cake-street selling voices (i.e. MM_S, MT_S, MY_S) than to their "usual" speaking voices (i.e. MM_U, MT_U, MY_U). However, we see an interesting difference across the three language groups of listeners. Japanese listeners gave MM_S the highest score, whereas Chinese and American listeners clearly preferred the voices of MT_S and MY_S, respectively, as cake-sellers. The mean rating of MM_S by Japanese listeners is 3.7, that by Chinese and American listeners are only 2.3 and 2.1, respectively.

Table 1: *Mean ratings of each speech sound by Japanese, Chinese and American listeners*

	Japanese	Chinese	Americans
MM_S	**3.7**	2.3	2.1
MT_S	3.4	3	**3.3**
MY_S	3.4	**3.8**	3.2
MM_U	2.3	1.8	2.9
MT_U	2.7	2.3	2
MY_U	1.9	2.3	2.6

What the listeners were paying attention to for making their evaluations can be glimpsed from looking at the impressions of the listeners about the voices. Figure 1 shows the comments made by the three groups of listeners for the street voice of MM_S.

For MM_S, 88% of Japanese listeners felt she had a quality street seller voice (including judgements of "typical street seller", "mastered manual of waiting on customers", and "good impression"), with an additional 10% giving positive vocal attributes, such as decent, pleasant, relaxed, polite, and, interesting; 2%, judged her voice as being too classy for a cake-seller. MM's street seller voice had only positive evaluations by Japanese listeners, but they evaluated MT_S and MY_S negatively sometimes, rating each with a score of 3.4. MY_S, who had some experience as a cake seller, had a voice evaluated about half of the time as positive (45%: relaxed, pleasant, cute, gentle, weak/delicate) and half of the time as negative (55%: inexperienced, non-professional, not-classy, unnatural-sounding voice, no motivation, even a "scary" voice).

MT, who was actually a supermarket cashier, was judged (24%) by Japanese listeners to have a street seller voice appropriate for a less-quality store, like a market, electronic shop, or drug store or somewhat negative qualities (33%) (inexperienced (11%), or like a machine (22%)). Interestingly, MT was also felt to have an energetic (22%) or perspicuous/astute (13%) voice (which apparently is not considered by Japanese culture to be an aspect of a quality street seller's voice), as well as 8%, a pleasant voice. It may be that MT uttered her street seller voice with no awareness of the difference between cake selling and selling of pharmaceuticals or electronic equipment.

Whereas Japanese listeners tended to evaluate MM_S positively, 81% of Chinese listeners' impressions of her voice were negative: "too fast" (25%), "too controlled, no feelings, like a robot" (25%), "tired, uncheerful, unhappy" (11%), "indifferent" (11%), "strange" (9%). American listeners also evaluated MM_S negatively: boring, tired (35%), short, abrupt (15%), shy, scared (15%), annoying (5%); however, 10% judged her voice positively or neutrally (10% flowy like a song, 10% question, 10% straight-forward).

In addition, both Chinese and American listeners tended to evaluate MT_S and MY_S more positively than MM_S.

Chinese listeners heard MT_S as 61% positive (energetic, cheerful), active (40%), pretty, good (12%), kind and

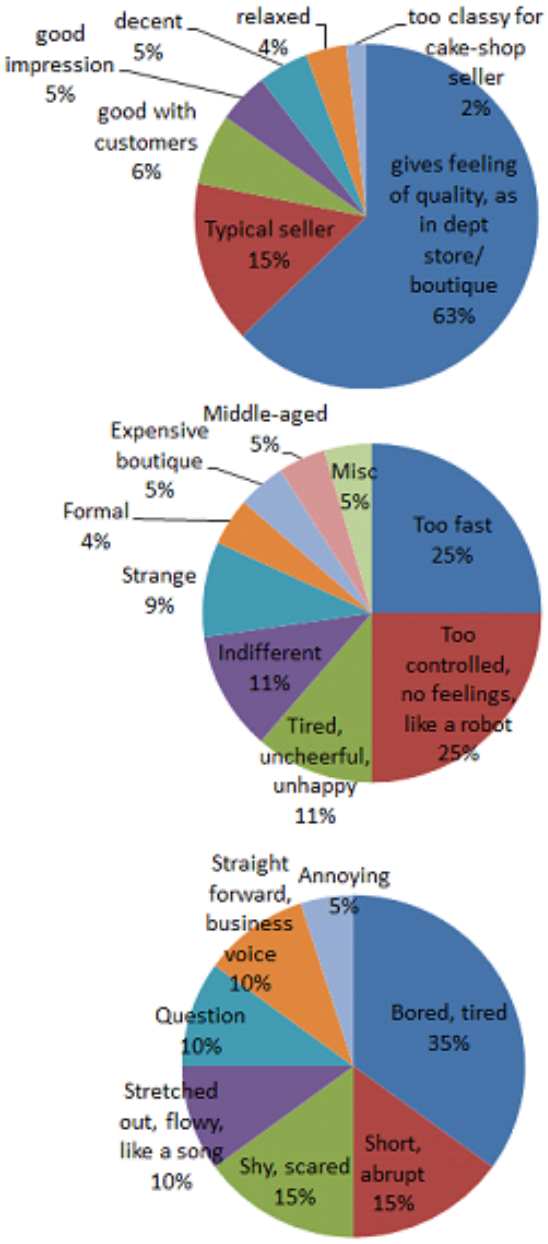

Figure 1: *Japanese (top), Chinese (middle) and American (bottom) listeners' impressions of the MM_S cake-street seller voice.*

motivated (9%). American listeners heard MT_S about half of the time (45%) as positively enthusiastic, excited, upbeat, good cake seller, enjoys job, confident, inviting; however, 50% was negative (sounds like a child, cold, arguing, scolding, annoying, urgent, demanding). Nevertheless, Americans gave MT_S a slightly higher rating (3.3) than the Chinese did (3.0).

For MY_S, Chinese heard the voice 33% as good, pretty, soft, gentle, like the *moe*-voice of Japanese anime, and gave the highest mean rating (3.8) to her; Americans heard the voice as 20% friendly, pleasant, enthusiastic and rated her as a slightly less good cake seller than MY_S (3.2).

As for the non-cake selling voices, all groups of listeners mostly had negative impressions about these voices. For Japanese listeners, MM_U, MY_U and MT_U were heard as blunt, no emotion like a machine, no motivation for speaking (36%, 18%, and 31%, respectively) or inexperienced, unnatural, unpleasant, unenergetic, bad impression, preoccupied (40%, 66%, and 41%, respectively). In addition, MY's voice was heard as weak (15%), while MT's was heard as angry or scary (17%) or relaxed (11%). Similar negative comments were made about MT's and MY's usual speaking voice by Chinese and American listeners.

However, MM's usual voice was heard by some Japanese as friendly (7%), or possibly the voice of someone working in a department store or boutique (7%) or less fashionable market or bookstore (10%). For the Chinese listeners, MM's usual voice was never heard positively, only negatively (strange, 33%; tired, uncheerful, 16%, etc.). For American listeners, 60% agreed that MM's usual speaking voice was tired (30%) or not cheerful, annoyed (30%); but, 40% gave positive comments (clear, pleasant 30%, upbeat, inviting 10%).

3.2. Acoustic results

According to the F0 analysis, shown in Table 2, MY_S had the highest overall F0 of all the speakers, while MT_S had the greatest range of F0.

Table 2: *Highest and lowest values and their ranges for F0 (Hz) of the six voices.*

Voice / F0	MM_S	MM_U	MY_S	MY_U	MT_S	MT_U
Max	335	296	**457**	347	390	262
Min	202	228	340	275	205	156
Range	145	65	117	72	**185**	106

Spectral differences in the final /e/ of the street sellers' voices for the word, *gorankudasaimase* are shown in the following figures: Figure 2 compares MM_S with MY_S, and Figure 3, compares MM_S with MT_S. The measurements were made at a point in the signal where the F0 was nearly the same for each speaker. Notice that MM_S always has a boost-up of energy around the 2.5 to 3.5 kHz region—which is characteristic of a "twang" voice [7,8]. Notice that this boost up of energy is not seen for either MY_S or MT_S.

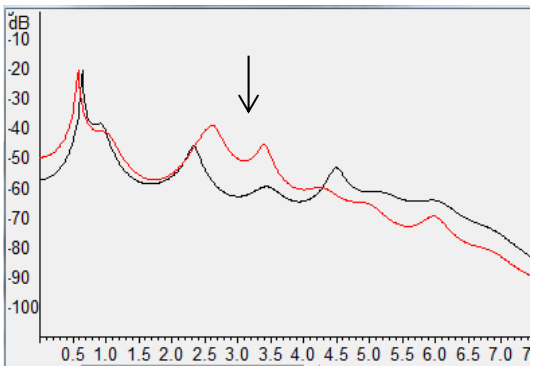

Figure 2: *MM street seller voice: red, MT street seller voice black . Final /e/, 290.34 Hz*

Additionally, MM_S shows more of a drop-off of energy after 4 kHz than do the other two speakers, which may reflect her comparatively more breathy quality. Especially, for speaker, MY, we see no drop off of energy, but rather an increase of energy, which may account for some perceptions of an *angry* voice by some Japanese listeners.

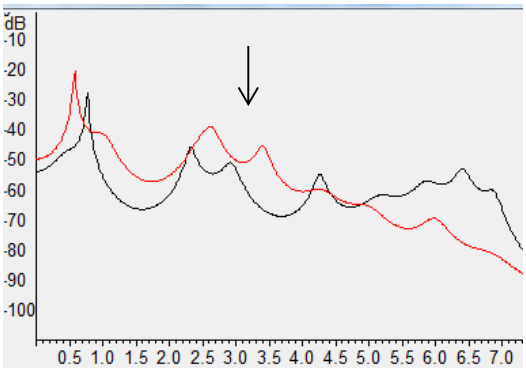

Figure 3: *MM street seller voice: red, MY street seller voice black., Final /e/, 290.34 Hz and 190.83 Hz, respectively*

4. Discussion

The results show that culture affects how ethnophonetic sounds are perceived, in this case, cake-selling street voices. Japanese listeners are very sensitive to what voice is best as a seller of fashionable Western cakes, and indicate sensitivities to differences between non-fashionable street selling voices (e.g., MY_S, MT_S) and fashionable ones (MM_S).

Three interesting points are (1) non-Japanese listeners rate the cake-street seller voice considerably lower than Japanese listeners do, (2) non-Asian listeners (Americans) differ from other Asian listeners (i.e., Chinese) in rating the cake-street seller voice—MT_S is rated higher by Americans (3.3) than Chinese (3.0), while Chinese prefer MY_S (3.8) compared with Americans rating of MY_S (3.2), and (3) non-Japanese listener rate the usual (speaking) voices differently from Japanese liseners.

In terms of acoustic differences among the three street selling voices, MY_S had the highest overall F0 of the three speakers, MT_S had the greatest range of F0, and MM_S had the most twang.

Interesting questions arise as to why did each language/culture group prefer a different street-seller voice? Was there something in the acoustic signal that they tuned into?

Our suggestions at this point are the following. Chinese listeners preferred the high F0 of MY_S, associating this with the sweet, high F0, *moe*-type voice [9]. In fact, many of the Chinese listeners wrote “moe” for impressions of MY_S street selling voice. Is this possibly because of the current popularity of anime in China, and the knowledge that the *moe* voice has a high F0?

American listeners, on the other hand, preferred the dynamic pitch range of MT_S, perhaps because it is more similar to American female voices, which are reported to have wide pitch ranges (e.g., [10]). Roughly half of the American listeners wrote “enthusiastic, excited, upbeat” for their impressions of MT’s street selling voice.

As for Japanese listeners, they strongly preferred the cake-street selling voice of MM. This is the only voice that had the twang-type boost-up of energy in the upper regions. This boost-up of energy is often referred to as the “singer’s formant” [11] for singers, or “actor’s formant” (e.g., [12]), and is what allows the sound to be heard over loud music of the orchestra, or at the back of a large auditorium, or in the case of street sellers, in a very noisy street environment. The interesting thing is twang phonation not only allows a voice to be heard in a noisy or loud environment, it also does not involve any extra vocal effort of the vocal folds [12].

In addition, MM’s cake-street selling voice had a slight breathiness to it, which may have softened the twang a bit to give it a more “elegant” tone characteristic of elegant (*jouhin*) Japanese ladies. The *jouhin* voice, also said to be used by women salesclerks in fashionable boutiques, tends to also be referred to as ”hanagoe” (nasal voice) or “hana ni kakatta koe”(nasal twang voice) https://matome.naver.jp/odai/2133726441194925701. The term implies that the voice is nasal; however, in actuality, there may be no airflow from the nasal passage ways [13]. Also, we note that moderate breathiness in the voice, also found with MM’s voice, is a voice characteristic of heroes in Japan [14, 15].

Currently, a number of studies have examined different types of Japanese female voices, e.g., *tsun* vs. *moe* voice which is distinguished, among other things, by F0 height and loudness [9]; sweet voice, which is not related to F0 but to voice quality [16]. Our current and ongoing work with cake-seller voices adds to this body of literature, suggesting that different cultures cue into different acoustic cues, which results in cultural-specific semantic impressions of phonations.

In addition, in Japan, as well as in other cultures, there are different types of "street sellers" (computer/ electronics store, fish market, etc), each one of which may be characterized by a slightly different type of voice. In future work we hope to examine cross-cultural comparison of a variety of street seller voices. Also, it will be interesting to examine possible links between good cake selling voices and selling more cakes, as well as better understanding how gender as well as cultural/language experience affects listeners’ perceptions of which voice is considered to be the best cake selling voice.

5. Conclusions

This ethnophonetic study reveals that the perceptual abilities of Japanese listeners are highly tuned to subtle voice quality differences. We advocate that these differences are specific to the Japanese culture. Our finding suggests that ethnophonetic research is a very important aspect of appreciating different cultures, and these findings in particular, lend insight into the Japanese culture.

One also wonders about the applications for this--could this research, for instance, be helpful in training quality cake-selling voices, or teaching Japanese culture and language to second language learners? Is Japan unique with regard to ethnophonetic expressions? Or can these be heard around the world? We suggest that each culture has a large number of ethnophonetic expressions, and we hope that this study will encourage other researchers to examine similar types of expressions in their culture/society.

6. Acknowledgments

This study was partially supported by the Ministry of Education, Culture, Sports, Science and Technology, Grant-in-Aid for Scientific Research (A), 15H02605, Japan.

7. References

[1] J. Gumperz, *Discourse Strategies*. Cambridge: Cambridge University Press, 1982

[2] D. Erickson and K. Maekawa, K. "Perception of American English emotion by Japanese listeners," *Acoustical Society of Japan, Spring Meeting*, pp. 333-334, 2001.

[3] T. Sadanobu, Z. Chunyue, D. Erickson and K. Obert, "Japanese 'street seller's voice'," *The 5th Joint Meeting of the Acoustical Society of America and Acoustical Society of Japan, Honolulu, Hawaii*, p. 3400, 2016.

[4] T. Sadanobu, C. Zhu, D. Erickson, K. Obert, "Japanese 'street seller's voice'," *Proc. Mtgs. Acoust.* **29**, 060003, doi: 10.1121/2.0000404, 2016.

[5] N. J. Enfield, "Ethnosyntax: Introduction," In N. J. Enfield (ed.), Ethnosyntax: Explorations in Grammar and Culture, 3-30, Oxford: Oxford University Press, 2002.

[6] T. Sadanobu, C. Zhu, D. Erickson, K. Obert, "Phonation by Japanese cake-sellers and its impressions: A contrastive viewpoint between Japanese and Chinese listeners," *Fall Meeting of Phonetic Society of Japan*, pp. 61-66, 2017.

[7] J. Estill, T. Baer, K. Honda, and K. Harris, "Supralaryngeal activity in a study of six voice qualities," *Proc. Stockholm Music Acoustics Conference*, pp. 157-174, 1983.

[8] K. Honda, H. Hirai, J. Estill and Y. Tohkura, "Contribution of vocal tract shape to voice quality: MRI data and articulatory modeling," In O. Fujimura and M. Hirano (eds.), Vocal fold physiology, Voice Quality Control, San Diego: Singular Publishing Group, pp. 23-38, 1995.

[9] S. Kawahara (2016) "The prosodic features of "tsun" and "moe" voices," *Journal of the Phonetic Society of Japan*, vol. 20:2, pp. 102-110, 2016.

[10] I. Mennen, F. Schaeffler and G. Docherty, Cross-language differences in fundamental frequency range: A comparison of English and German. *J. Acoustic Soc. Am.* Vol. 131.3, pp. 2249-60, 2012.

[11] J. Sundberg, *The Science of the Singing Voice*, DeKalb, Illinois : Northern Illinois University Press, 1987

[12] T. R. Gates, A. Forrest, and K. Obert, *The Owner's Manual to the Voice: A Guide for Singers and Other Professional Voice Users*, NY, NY: Oxford University Press, 2013.

[13] D.Erickson, K. Obert, R. Hayashi, C. Zhu, K. Perta, K., T. Sadanobu, K. Sakakibara, K. "Is 'hanagoe' really nasal ? – Acoustic and MRI data analysis of Japanese cake seller's voices," *Spring Meeting Acoustical Society of Japan*, 2018

[14] E. Z. Murano and M. Teshigawara, "Articulatory correlates of voice qualities of god guys and bad guys in Japanese anime: an MRI study," *Interspeech 2004 - ICSLP, 8th International Conference on Spoken Language Processing, Jeju Island, Korea, October 4-8*, 2004

[15] M. Teshigawara, *Voices in Japanese Animation: A Phonetic Study of Vocal Stereotypes of Heroes and Villains in Japanese Culture*. PhD thesis. Nagoya University, 2003.

[16] R. L. Starr, Sweet voice: The role of voice quality in a Japanese feminine style, Language in Society, vol. 44, pp, 1–34, 2015. doi:10.1017/S0047404514000724, 20015.