The phonetics of Japanese maid voice I: A preliminary study*

Shigeto Kawahara
Rutgers University

ABSTRACT. Maids working at meido kissa ("maid cafés") in Akihabara use specific speech styles. This paper reports a preliminary acoustic analysis of the characteristics of maid voice style. The analysis of two professional maids shows that the differences between normal voice and maid voice manifest in various acoustic dimensions, including rate of F0 change, voice quality, intensity, and formant patterns. The two maids shared some, but not all, strategies to create their maid voice. The study reported here is preliminary, and it is hoped that it will stimulate interest in future investigation on this topic. The pedagogical value of this sort of research is also discussed at the end.

Keywords: job-specific speech style, acoustic analysis, vocal attractiveness, maids, Akihabara

1. Introduction

Meido kissa (or "maid cafés") developed starting in March 2001, mainly in Akihabara (Takatora 2012). Maids working in meido kissa prototypically use specific phrases; for example, they say *okaerinasaimase, goshujinsama* "Welcome home, my master" to male customers, and use *ojoosama* "young lady (honorific)" to refer to female customers. Casual observation in Akihabara tells us that these maids use a distinctive tone of voice. (NB: there are various kinds of meido kissa, and not all maids change their voice at work.) This maid voice is closely tied to the culture of "*moe*", which has been developing in Akihabara and elsewhere, as well as on the internet (see the Wikipedia article on *moe* for an extensive discussion on this concept: http://en.wikipedia.org/wiki/Moe_(slang)).

The question that this project attempts to address is what these maids do exactly to create their maid voice. This paper reports an exploratory acoustic experiment that attempts to address this question. Since virtually nothing is known about the acoustic characteristics of maid voice, this study is necessarily a preliminary and exploratory one. My hope is that this study will spur interest in this and related areas. I also discuss the possible pedagogical value of this sort of research.

Before proceeding to the main discussion, one point needs to be made clear. The current project does not simply come out of curiosity, but instead (or in addition) is motivated by some larger related questions: (i) what is *moe* (from a (psycho)linguistic perspective)? (ii) even more generally, what kinds of voice do listeners find attractive (assuming that *moe* voice is attractive)? and (iii) once we find acoustic characteristics of attractiveness, how can we deploy these results for speech synthesis technologies? See Babel et al. (2011) and references cited therein for related discussion on vocal attractiveness.

2. Method

Two maids working at Félicie in Akihabara were recorded (Maid R and Maid S). They were both professional maids working at meido kissa at the time of recording. They all had experience working as a maid for more than a few years. The task was to read phrases and sentences in their normal voice (*ji-goe*) and their maid voice (*meido-goe*).

2.1. Stimuli

The recording session started with a warm-up phase, consisting of typical maid phrases. These warm-up phrases were read only in maid voice. To measure the differences in intonational contours between normal voice and maid voice, four sets of SOV sentences were prepared. Both subject and object nouns were 4 mora long, and accented on the second syllable, with an LHLL contour (Pierrehumbert and Beckman 1988). The verbs were also accented. For all the words, obstruents were avoided as much as possible to prevent the perturbation of F0. For example, one test sentence is *Mori'mura-ga Ama'ria-o aware'nda* 'Morimura felt sorry for Amalia'.

The maids repeated each sentence 4 times in normal voice and then 4 times in maid voice. They repeated this procedure twice (i.e. 4 repetitions with normal voice => 4 repetitions with maid voice => 4 repetitions with normal voice => 4 repetitions with maid voice). This repetition structure was deployed to assure that normal voice does not simply serve as a practice for maid voice. Each sentence was thus recorded 8 times for each type of voice.

To measure other vocalic properties (such as formant values and intensity), the stimuli also included the five vowels in Japanese [a, i, u, e, o]. The five vowels were ordered in three different orders: [a, i, u, e, o], [i, e, a, o, u], and [u, o, a, i, e]. For each order, they first pronounced the five vowels 10 times in normal voice and then 10 times in maid voice. They repeated this same procedure (10 repetitions in normal voice and 10 repetitions in maid voice) twice. Therefore, each vowel was recorded a total of 60 times (=3 vowel orders * 10 repetitions * 2 repetition orders) for each type of voice (normal and maid).

2.2. Recording setting

Their speech was recorded using a DR-40 recorder (TASCAM) with a 44.1k sampling frequency and a 16 bit quantization level. Recording took place in a quiet room at Félicie.

2.3. Analysis

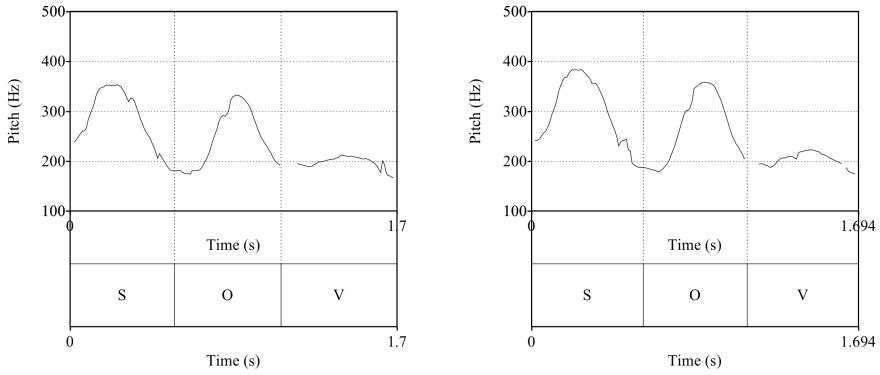
Since virtually nothing is known about the acoustics of maid voice, the analysis was determined post hoc, and proceeded as we found interesting patterns. All the acoustic analyses were done using Praat (Boersma and Weenink 1999-2012).

3. Results

3.1. Sentential intonation

One of the most noticeable differences between normal voice and maid voice manifests itself in the intonational contour. Figure 1 exemplifies this difference, using a pair of intonational contours from

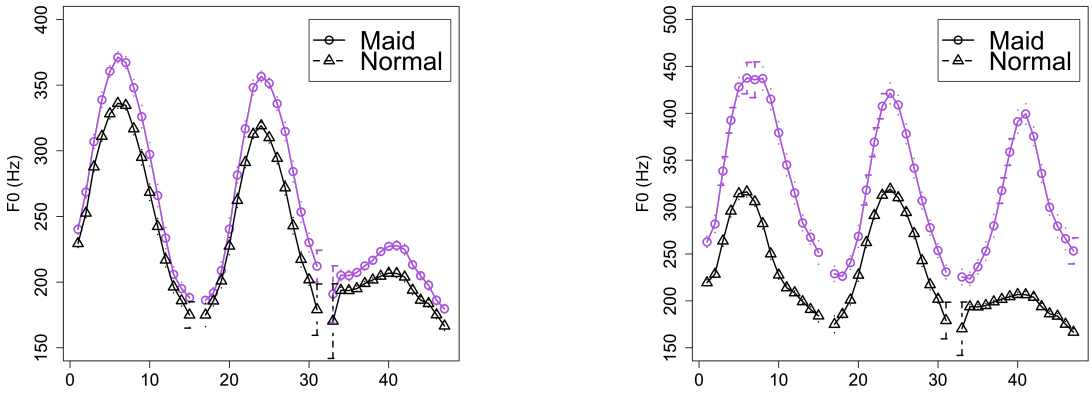
Maid R. We observe that the F0 is generally higher in maid voice (the right panel), which is especially visible in the first H(igh) peak of the subject noun (which is almost as high as 400Hz). We also observe that while an initial LH rise is barely visible on the verb in normal voice, it is more clearly manifested in maid voice.



(a) normal voice (b) maid voice

Figure 1. An illustrative pair of intonational contours in normal and maid voice. Maid R.

To quantitatively assess this difference between normal voice and maid voice, each syntactic phrase (subject, object, verb) was divided into 15 equally-timed windows, and the average F0 was calculated within each window. Figure 2 shows the normalized intonational contours obtained this way, averaging over 4 sentences and 8 repetitions.



(a) Maid R (b) Maid S

Figure 2. Normalized intonational contours. Maid R (left) and Maid S (right).

We observe that both of the maids generally show higher pitch in maid voice (the line with circles) than in normal voice (the line with triangles); however, there are some differences between the two maids. For Maid R (the left panel), L(ow) targets seem to be barely raised in maid voice, whereas for Maid S (the right panel), both L and H targets are raised in maid voice.

To assess these observations statistically, linear mixed models were run in which the

condition (normal vs. maid) and syntactic positions were fixed factors. For Maid R, the difference between normal voice and maid voice is not significant for L-tones (t = -1.81, *n.s.*), but is significant for H-tones (t = -16.23, p<.001). The difference is also significant for the size of LH-rises (H-tones minus L-tones) (t = -2.97, p<.01). These results show that Maid R raises her H targets in maid voice, but there is no evidence that she raises her L targets. This pattern is compatible with Fujisaki's (1983) model of intonation, in which the intonational baseline stays more or less constant.

For Maid S, the difference between normal voice and maid voice is significant for L-tones (t = -6.71, p<.001), for H-tones (t = -26.91, p<.001), and for the LH-rises (H-tones minus L-tones) (t = -16.69, p<.01). These results show that Maid S raises both L and H targets. However, the fact that the LH-rises are larger in maid voice than in normal voice shows that it is not the case that her pitch range simply shifts upwards—the H targets are raised more than the L targets[1].

3.2. Speech rate and rate of F0 change

A question arising from the previous analysis is whether the maids speak slower in maid voice, as maid voice involves larger F0 movement (and larger F0 movement should take more time, all else being equal). To address this question, Table 1 shows the averaged duration of the subject nouns and object nouns in ms with 95% confidence intervals (the verbs are excluded from this analysis because they are not controlled in terms of mora counts). From the results in Table 1, it does not seem to be the case that the maids consistently speak faster in maid voice.

Maid R	SUBJECT	OBJECT
Normal voice	56.8 (1.0)	55.1 (1.7)
Maid voice	56.4 (0.8)	55.2 (1.4)
Maid S	SUBJECT	OBJECT
Normal voice	56.2 (1.07)	55.8 (1.3)
Maid voice	53.6 (1.0)	55.9 (1.2)

Table 1. Averaged duration of each interval in ms with 95% confidence intervals in parentheses.

The results in Table 1, together with the results of sentential intonation, suggest that F0 change rate should be greater in maid voice. To test this expectation, F0 changes per ms were calculated for both rise and fall for both subject and object nouns. The results appear in Table 2.

Maid R	SUBJECT-Rise	SUBJECT-Fall	OBJECT-Rise	OBJECT-Fall
Normal	5.1 (0.17)	- 4.0 (0.14)	6.2 (0.41)	- 6.4 (0.54)
Maid	6.3 (0.17)	- 4.8 (0.14)	6.9 (0.35)	- 7.6 (0.37)
Maid S	SUBJECT-Rise	SUBJECT-Fall	OBJECT-Rise	OBJECT-Fall
Normal	5.9 (0.35)	- 3.3 (0.19)	4.3 (0.54)	- 2.8 (0.58)
Maid	9.9 (0.79)	- 8.9 (0.84)	9.6 (0.66)	- 8.9 (0.47)

Table 2. Averaged F0 change per ms with 95% confidence intervals in parentheses.

Table 2 reveals that for both of the maids, the difference between normal voice and maid voice

manifests itself in a greater F0 change per ms. In other words, the maids use more dynamic F0 movement in maid voice.

3.3. Some observations about [a, i, u, e, o] reading: Maid R

Next, I move on to some observations about the pronunciation of the five vowels, starting with Maid R. First, Figure 3 illustrates the intonational contours of the five vowels in three different orders. As observed in Section 3.1, her L targets are about the same between maid voice and normal voice, but H targets are raised in maid voice, again compatible with the Fujisaki model of intonation.

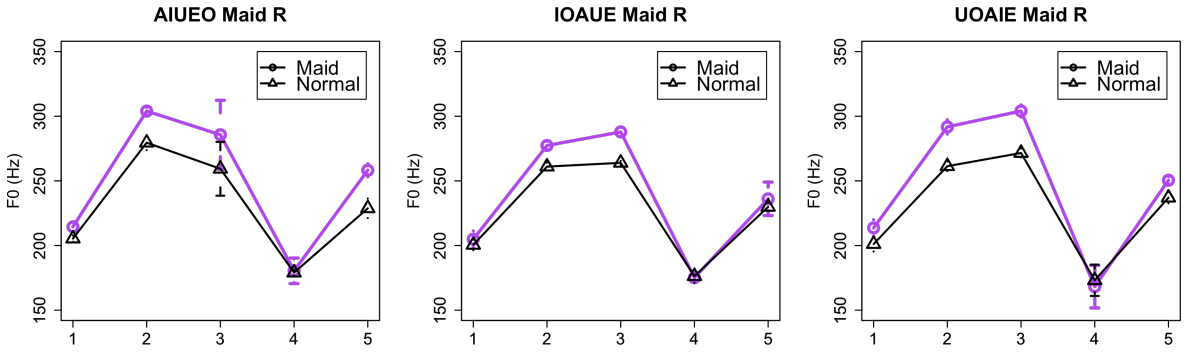
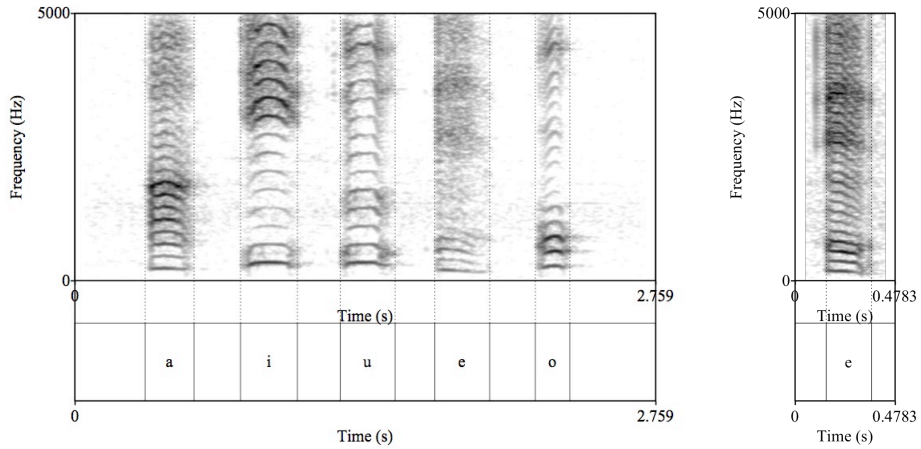


Figure 3. Intonational contours of the five vowels. Maid R.

Another impressionistic characteristic of her maid voice is breathy voice found in the penultimate vowel. Figure 4(a) shows a narrow-band spectrogram of [a, i, u, e, o] (window length=0.05s). The penultimate vowel [e] is breathy, as evidenced by the lack of clear harmonic structures in high frequency ranges[2]. Figure 4(b) shows the narrowband spectrogram of her [e] in normal voice when reading [a, i, u, e, o], which does show clear harmonic structures in high frequency ranges; i.e. penultimate breathiness does not appear in normal voice.



(a) maid voice (b) normal voice

Figure 4. (a) A narrow-band spectrogram of [a, i, u, e, o] in the maid voice of Maid R. Window length=0.05s. (b) A narrow-band spectrogram of [e] in the normal voice of Maid R.

This breathiness does not seem to be a property of [e] *per se*, but a property of penultimate vowels, as penultimate vowels were breathy in the other two orders of vowel readings[3]. Figure 5 illustrates her reading of [u, o, a, i, e], which shows that [i] is breathy. Some attempts to quantify this impressionistic breathiness (e.g. spectral tilt and harmonic-to-noise ratio: Gordon and Ladefoged 2001) have not successfully characterized the penultimate breathiness, however. Quantitatively assessing the breathiness in penultimate syllables is thus left as a topic for future research.

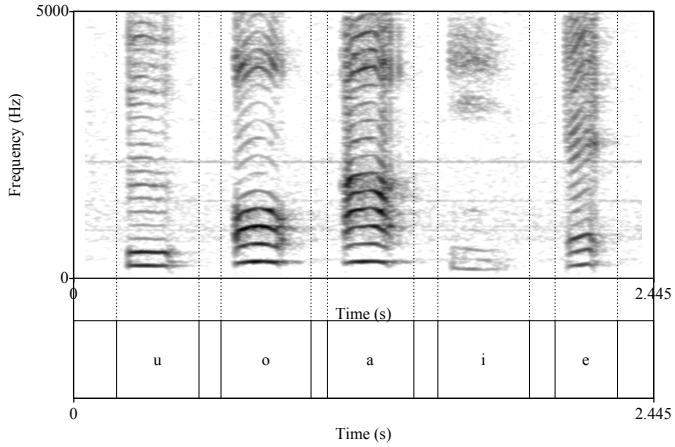


Figure 5. A narrow-band spectrogram of [u, o, a, i, e]. Maid R.

Finally, maximum intensity for each of the five vowels was measured, the results of which are shown in Figure 6. There is an occasional tendency for maid voice to be louder than normal voice, but the tendency was not so consistent (cf. Maid S below).

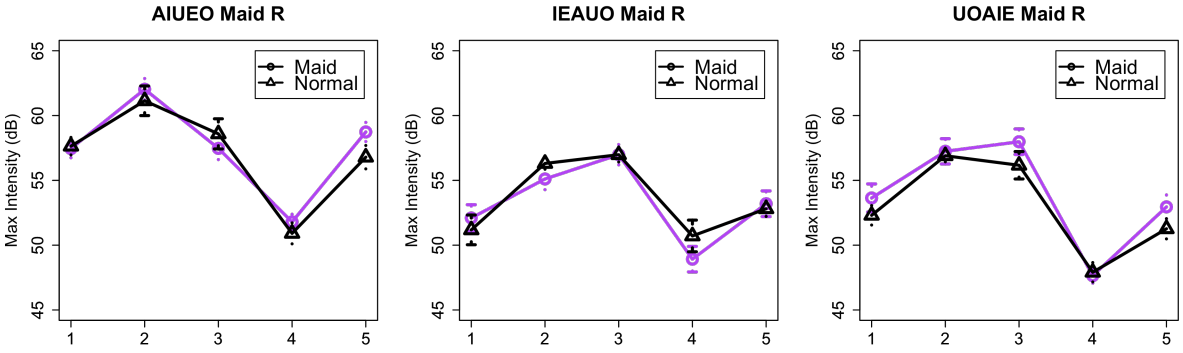


Figure 6. Averaged maximum intensity contours. Maid R.

3.4. Some observations about [a, i, u, e, o] reading: Maid S

Next, we turn to Maid S. Figure 7 illustrates the intonational contours of the five vowels of Maid S. As was the case for the intonational contours we observed in Section 3.1, Maid S raises both the L and H targets. We also observe that the size of the differences between L and H are much larger in maid voice than in normal voice.

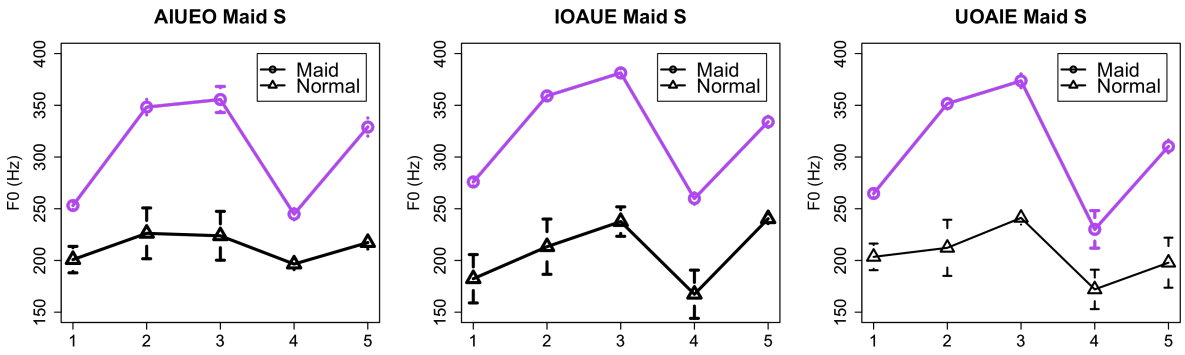


Figure 7. Intonational contours of the five vowels. Maid S.

Unlike Maid R, Maid S does not show any sign of penultimate breathiness, at least impressionistically (and as observed from the inspection of her spectrograms, which are not shown here due to space limitations). However, she showed clearer differences in intensity. This difference in intensity is illustrated in Figure 8. Maid S speaks much louder in maid voice than in normal voice.

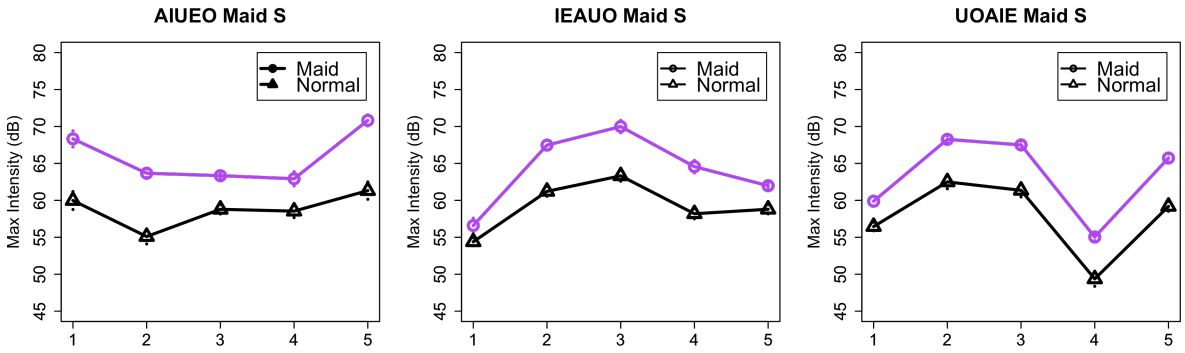


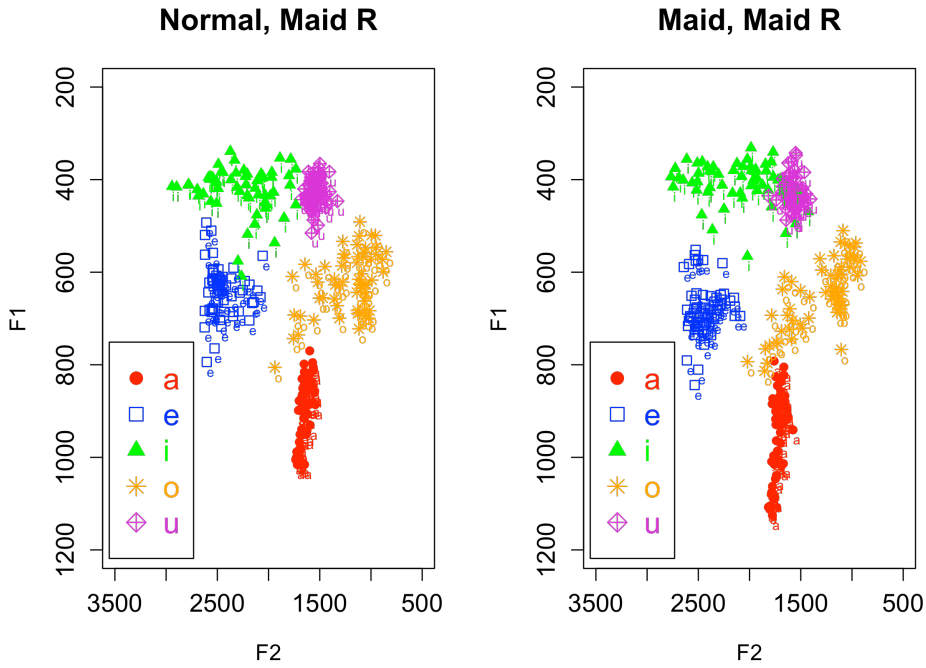
Figure 8. Averaged maximum intensity contours. Maid S.

3.5. Differences in vowel spaces

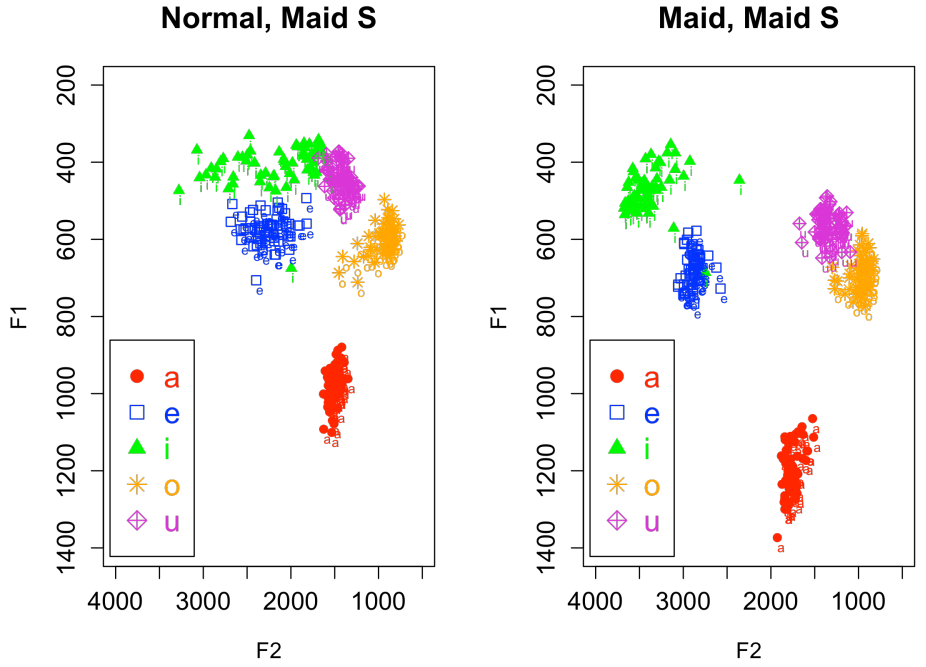
Finally, the maids' formant values were extracted based on their pronunciations of the five vowels. Using Praat, after some trials and errors to extract correct formant values without mistracking, 5 formants were extracted within 6000Hz for Maid R; 5 formants were extracted within 6000Hz for normal voice and within 7000Hz for maid voice for Maid S (different settings had to be used between the two conditions for Maid S, because her F0 was very different between normal voice and maid voice). Figure 9 illustrates the results.

First, for Maid R, [i] seems characteristically less acute in maid voice. F2 average is 2264Hz for normal voice, whereas it is 2110Hz for maid voice. My speculation is that this lowering of F2 is perhaps due to less spreading/compression of the lips, which reduces the images of sharpness and gives a soft impression: see Shinohara and Kawahara (to appear) for the discussion of these images pursued by maids in Akihabara. Next, [a] is lower in maid voice: F1 average is 900Hz for normal

voice and 959Hz for maid voice.



(a) Maid R



(b) Maid S

Figure 9. The vowel spaces of Maid R and Maid S.

For Maid S, differences between normal voice and maid voice are more apparent. First, [i] is much more acute in maid voice. F2 average is 2170Hz for normal voice, and 3340Hz for maid voice.

Likewise, [e] is also much more acute in maid voice. F2 average is 2261Hz for normal voice, and 2888Hz for maid voice. Second, [a] is lower in maid voice. F1 average is 978Hz in normal voice and 1199Hz in maid voice. Finally, [u] is lower and grave in maid voice. F1 average is 443Hz vs. 565Hz, whereas F2 average is 1425Hz vs. 1338Hz.

To summarize, both maids show larger opening of the mouth for [a]. Maid R shows less acute [i] whereas Maid S shows more acute [i, e]. My impressionistic (admittedly unsubstantiated) speculation is that Maid R gives a less sharp impression (assuming that [i] with lip compression gives sharper impression), whereas Maid S is shooting for a more lively character, expressed by way of more hyperarticulated vowels with much higher intensity.

4. Conclusion

4.1. Summary

To summarize, maids do (or can) change their voice. Two different maids use different strategies, although they do have some in common. (1) and (2) provide a summary of strategies for each of the maids. I do not wish to imply that these are exhaustive lists.

(1) Maid R:
- Raising of F0 H targets.
- More dynamic F0 movement per ms.
- Penultimate breathiness.
- Less acute [i] and lower [a].

(2) Maid S:
- Raising of F0 L and H targets (H-raising more extensive).
- More dynamic F0 movement per ms.
- More acute [e, i], lower and more grave [u], and lower [a].

4.2. What's coming next?

A few more maids have been recorded from a different meido kissa—however, the analysis of these new maids cannot be reported here due to space limitation. I have also recorded maid voice of two professional voice actresses (*seiyuu*) as well, whose analysis is in progress. Although no extensive discussion on these data is possible due to space limitation, impressionistically speaking, maid voice by these speakers seems to show higher F0 as well as more dynamic F0 movement, like the two maids discussed in this paper. A more extensive study of these speakers may allow us to identify defining acoustic features that are common to all maid voices.

Future investigations should also test other acoustic properties (such as higher formants, jitter, simmer, etc) to further investigate the nature of maid voice. Impressionistically speaking, maid voice is often characterized by nasalization throughout utterances. An analysis using bandwidth (Johnson 2003) did not successfully characterize the nasalization of the two maids analyzed in this project. A more direct articulatory experiment using a nasometer would address this question. Finally, perceptual tests addressing how the acoustic manipulations made by the maids impact the perceived attraction are also warranted.

4.3. A final remark

As a final remark, this project was at first partly motivated by my personal curiosity, but I found that this project has significant pedagogical value as well. Since teaching phonetics involves mathematics and physics, it can be challenging, especially in undergraduate education. In my personal experience at Rutgers University, some students give up before mastering the basics, and never get to see how much they can do once they learn how to do acoustic analyses. I have presented this project as a special guest lecture at International Christian University, and it turned out that the presentation was a very effective means to introduce phonetic concepts/analyses to undergraduate students. Indeed, to my pleasant surprise, some students there have started their own acoustic phonetics study group.

****

Notes

* Acknowledgements: I would like to thank Minako Maezato for arranging the recording sessions, Takatora-san for sharing his knowledge of the Akiba culture, Michinao Matsui for discussing the phonetic analyses, and my lab assistants at Rutgers for helping with the acoustic analysis. Thanks to Aaron Braver, Jeremy Perkins, my lab assistants, especially Jess Trombetta, and anonymous reviewers for comments on earlier versions of this paper. This talk was presented at ICU and at the Spring meeting of the Phonological Society of Japan, at which I received useful comments. This research is supported by a Research Council Grant from Rutgers University.

[1] As Haruo Kubozono pointed out (p.c.), it may be that Maid S is attempting to keep the L targets the same between normal voice and maid voice (as the Fujisaki model predicts), just like Maid R; however, Maid S's Hs are too high for her L tones to go back to her true baseline.

[2] When I gave this talk at International Christian University, many students agreed that the penultimate vowels were the cutest among the five vowels of this maid.

[3] Impressionistically, however, [o] may be less clearly breathy than front vowels. A quantitative analysis measuring this breathiness is on-going.

References

Babel, Molly; Joseph King; Grant McGuire; and Teresa Miller. 2011. Acoustic determiners of vocal attractiveness go well beyond apparent talker size. Vancouver, BC: University of British Columbia and Santa Cruz, CA: University of California, Santa Cruz, MS.

Boersma, Paul, and David Weenink. 1999-2012. Praat: doing phonetics by computer. Software.

Fujisaki, Hiroya. 1983. Dynamic characteristics of voice fundamental frequency in speech and singing. *The production of speech*, ed. by Peter F. MacNeilage, 39–47. New York: Springer-Verlag.

Gordon, Matthew, and Peter Ladefoged. 2001. Phonation types: A cross-linguistic overview. *Journal of Phonetics* 29.383-406.

Johnson, Keith. 2003. *Acoustic Phonetics, 2nd Edition.* Oxford: Blackwell.

Shinohara, Kazuko, and Shigeto Kawahara. to appear. The sound symbolic nature of Japanese maid names. *Proceedings of Japan Cognitive Linguistic Association*.

Pierrehumbert, Janet, and Mary Beckman. 1988. *Japanese tone structure*. Cambridge: MIT Press.

Takatora. 2012. Meido kissa deeta bukku, vol 0. Doojinshi.