To appear in the Journal of the Phonetic Society of Japan, 2016

The prosodic features of the "moe" and "tsun" voices*

Shigeto Kawahara
Keio University
The Institute of Cultural and Linguistic Studies

Abstract

As a case study of the general theme of this special issue—"different phonetic realizations of different social characters", this paper explores prosodic features of two types of prototypical maid voices, which have been emerging in recent Japanese culture: "moe" and "tsun". Two professional voice actresses read simple Japanese SOV sentences in three types of voices: moe, tsun, and normal. Acoustic analyses show that the moe voice is characterized by higher f0 and louder voice than the normal voice, whereas the tsun voice is characterized by lower f0 and quieter voice. The current study also finds that the speakers manipulate H-tone targets more extensively than L-tone targets to differentiate different speech styles, which is compatible with some previous studies and models of intonation. In terms of its research value, the current findings may not be ground-breaking; however, an additional value of this research lies in the fact that this sort of material makes phonetic analyses more accessible to the general public as well as to students in undergraduate education. To that end, some sample sounds are made available at `http://bit.ly/1WCu5DA`.

1 Introduction

This research note investigates the prosodic features of "tsun" and "moe" voices, two prototypical voice types observed in recent Japanese anime characters. Although these two concepts have existed for a while now, they are admittedly very difficult to define precisely. Both terms refer to types of people in the recent Japanese anime characters and related

*Thanks to Kazuko Shinohara for her collaboration on this general project, and to Akiko Takemura for her help with the annotation in Praat. I am also grateful to Toshio Matsuura and two anonymous reviewers for comments on a previous version of the paper, as well as Helen Stickney for her careful proofreading. This work is partially supported by JSPS KAKENHI Grant #26770147 and #26284059. All remaining errors are mine.

cultural domains. "Moe" can be characterized by a set of adjectives like "pretty, cute, small, accessible, friendly, and cheerful", whereas "tsun" would be characterized as "beautiful, tall, thin, shy, and inaccessible". This paper is not a place to settle this debate; readers who are interested in this debate are encouraged to refer to their relevant wikipedia pages (`Moe (slang)` and `tsundere`, and pages referenced therein).[1] The precise definitions of these concepts are not crucial in this paper—what is crucial is the fact that there is this distinction.

The psychological reality of this distinction has been tested in previous experiments on sound symbolism. It has been shown that maids working at Akihabara—and also more "generic" Japanese speakers—usually associate obstruents with names for tsun girls, and sonorants with names for moe girls (Kawahara et al., 2015; Shinohara & Kawahara, 2013). These results show that there is a sense in which this distinction between tsun and moe has been established in the contemporary Japanese culture, and most likely, in the Japanese language as well.

These two types of girls are often distinguished in terms of their voice characteristics in anime (see also Starr 2015, Teshigawara 2003a,b for analyses of speech styles in Japanese anime). One salient feature is their intonational differences—but how do these intonational differences between different voice types manifest themselves? This question is the main topic of this paper. We can conceive of two (or three) hypotheses about how the different types of voices can differ from one another in terms of intonation. Consider Figure 1, which illustrates different ways in which different speech styles can affect tonal low-high-low (LHL) patterns.

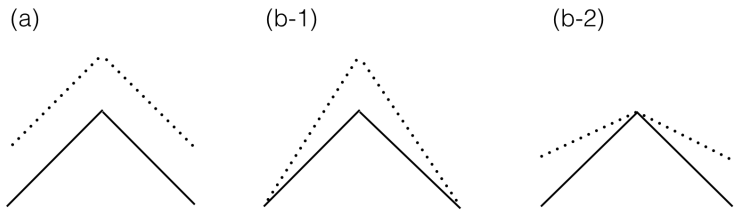

Figure 1: Schematic illustration of how different speech styles may result in different pitch contours, given a LHL tonal sequence. (a)=change in pitch range/register; (b-1)=change in high tonal targets; (b-2) change in low tonal targets.

One hypothesis is that voice differences vary in terms of pitch register, which would realize itself as shifts in pitch range, as in Figure 1(a). For example, moe voice can be shifted, with respect to the normal voice, higher overall. The alternative hypothesis is that the differences

[1]Tsun is often used in a larger phrase "tsun-dere", but it can be used in isolation. In fact, tsun can be understood as one phase of tsundere, either diarchonically or synchronically (Togashi, 2009). Moe voice is perhaps very similar, if not identical, to what Starr (2015) refers to as "sweet voice".

may be more localized; instead of shifting the whole pitch range, speakers may choose some particular tonal targets to express different kinds of voice, for example, primarily high tonal targets (Figure 1(b-1)) or low tonal targets (Figure 1(b-2)).

Kawahara (2013) found evidence for the pattern schematized in Figure 1(b-1). Kawahara (2013) explored whether two maids working at a maid cafe in Akihabara make a difference between their own voice and maid voice in terms of intonational differences. The results demonstrate that maid voice is characterized by H(igh) tone targets, but not so much in the L(ow) tone targets.

However, Kawahara (2013) did not explore different types of maids, a distinction which was discovered to be crucial in their naming conventions in later studies (Kawahara et al., 2015; Shinohara & Kawahara, 2013). This paper therefore explores the intonational differences between these two types of maid voices, in terms of how they differ from the normal voice. The analysis is based on the speech of professional voice actresses, who were asked to produce two types of maid voices, tsun and moe.

2 Method

The data reported in this paper is a part of a larger pool of data based on the pronunciation of two professional voice actresses. In general, they were asked to act as two types of Japanese maids, tsun and moe. They also produced most of the materials in their normal voice.

2.1 Stimuli

To explore the nature of intonational differences, several simple SOV Japanese sentences were prepared, which are listed in (1). An effort was made to avoid obstruents as much as possible, as they can perturb f0 contours. Both the subject and object nouns were four-mora long, and accented on the second (antepenultimate) mora. Hence they had a LHLL f0 contour. /ga/ and /o/ are nominative and accusative case particles. Three of the four verbs were five-mora long, and one of them was four-mora long. All verbs were accented on the antepenultimate mora.

(1) Stimulus sentences (accent shown with apostrophe)

a. Nono'mura-ga neri'ume-o aema'shita. "Nonomura mixed kneaded plums."
b. Nino'miya-ga oma'wari-o uraya'nda. "Ninomiya envied a police officer."
c. Nira'uri-ga ao'mori-o era'nda. "A leek-seller chose Aomori."
d. Mori'mura-ga ama'ria-o aware'nda. "Morimura felt sorry for Amalia."

2.2 Speakers

The speakers were two professional female voice actresses. Speaker 1 had a long professional experience (her seventh year at the time of recording), whereas Speaker 2 had slightly less than two years of experience. They were paid for their time. These two speakers had no trouble understanding the concept of tsun and moe.

2.3 Recording

The recording took place in a sound-attenuated recording room at the National Institute for Japanese Language and Linguistics (NINJAL), Tachikawa, Tokyo, by courtesy of Prof. Hanae Koiso. A portable recorder (TASCAM DR-40) was placed on a table in front of the speakers. Unlike "usual" speakers that we record in phonetic experiments, the voice actresses preferred to stand up during their recording.

2.4 Procedure

The recording for this experiment was embedded within a larger recording session, which lasted for two hours for each speaker. The total compensation was 20,000 yen for each speaker. The speakers were first asked to produce basic phrases that are associated with the maid or anime culture in Japanese, both in tsun and moe voices; e.g. *okaerina-saimase goshujin-sama* "welcome home, Master" and *goshujin-sama sorosoro omezame-no ojikandesu* "Master, it's time to wake up" (some speech samples are made available at `http://bit.ly/1WCu5DA`). The stimuli were presented in standard Japanese orthography, with a mixture of *kanji* and *hiragana*. The voice actresses examined all of the stimulus lists carefully before the recording, during which time they also confirmed the reading of the *kanji* characters as well as the accent placement of each stimulus word.

The recording was blocked in a Latin Square format: in the first session, the speakers were asked to read in the order of normal voice, tsun voice and moe voice. The second Latin Square order was tsun-moe-normal. The third was moe-normal-tsun. This Latin Squire design was deployed to control for potential order effects. Each sentence was repeated three times within each block. Therefore, each voice type was represented by 36 tokens (3 Latin Square orders $\times$ 3 repetitions $\times$ 4 sentences).

2.5 Analysis

Using Praat (Boersma, 2001), for each sentence, three annotation intervals were assigned, each corresponding to the subject noun, the object noun and the verb, as shown in Figure

2.

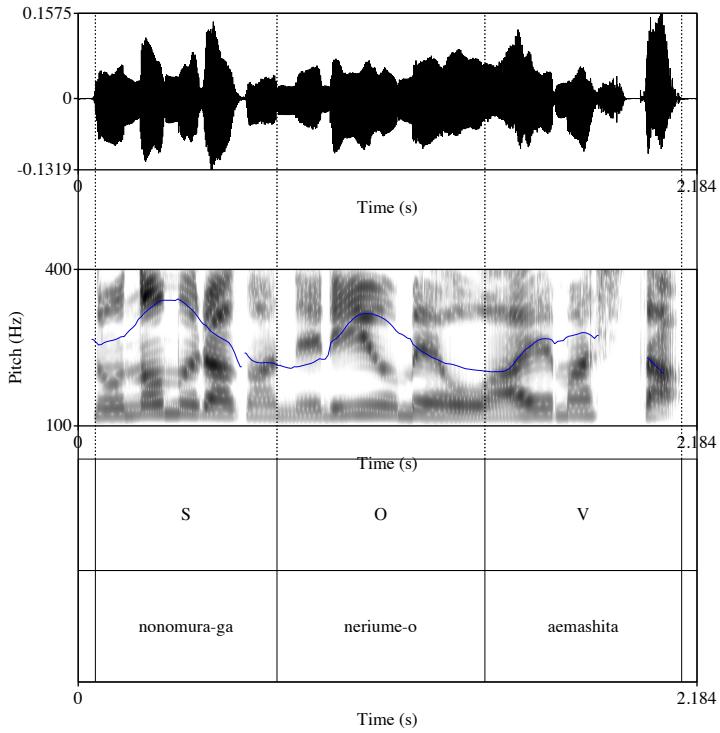

Figure 2: A sample illustrative annotation. S=Subject, O=Object, V=Verb. The second annotation tier showing actual Japanese words was added here for illustration. The intonational contour is superimposed on the spectrogram with a blue line.

To normalize the f0 contour of each sentence, each interval was divided into 15 equally-timed windows, and the mean f0 value was calculated within each window.[2] The last window of the verb phrase often failed to produce f0 values due to sentence-final creakiness (Kawahara & Shinya, 2008), and hence it was discarded from the following analysis.

Since Kawahara (2013) found that maids working in Akihabara made the difference between the maid voice and their own voice in terms of the magnitude of LH rise and HL fall in LHL tonal sequences, these measures were also calculated in the current study. First, within each of the SOV intervals, the maximum f0 (=H) was extracted; the minimum f0 before the H (=L1 in L1HL2) and the minimum f0 after H (=L2 in L1HL2)) were also extracted. Rise was calculated as the difference between the maximum and the first minimum (=H-L1); the fall was calculated as the difference between the maximum and the second minimum (H-L2). All of these processes were automated using Praat's scripting function.

As an additional analysis, maximum[3] intensity was analyzed for each type of sentence,

[2]This analytical technique was deployed in Déprez et al. (2013) and Kawahara (2013).

[3]A similar analysis was conducted with average intensity, which yielded very similar results.

because the general auditory impression was that the moe voice was louder than the normal voice, and that the tsun voice was quieter than the normal voice. Since there is a sense in which loudness or intensity is one acoustic dimension that defines the general prosody of a speech (e.g. Fry 1955; Plag et al. 2011), the intensity analysis was expected to shed light on potential prosodic differences between the different voice patterns.

2.6 Statistical analyses and predictions

R (R Development Core Team, 1993–2016) was used to conduct statistical analyses as well as to create figures of the results. Two statistical tests were deployed to compare the three hypotheses illustrated in Figure 1. First, a general linear model (`glm`) was fit to L-tone and H-tone targets, respectively, to examine whether and how the three different speech styles affect these tones. Second, a Browne-Forsythe test, which compares variability across conditions, was run to compare L-tones and H-tones. Figure 1(a) predicts no variability differences between L-tones and H-tones across different speech styles; Figure 1(b-1) predicts H-tone to show more variability across different speech styles; Figure 1(b-2) predicts L-tones to show more variability across different speech styles.

3 Results: Intonation

Figure 3 illustrates the intonational contour of the three voice types of Speaker 1: normal (+), tsun (△) and moe (○). The first observation is that with respect to the normal voice, the tsun voice is lower, whereas the moe voice is higher.

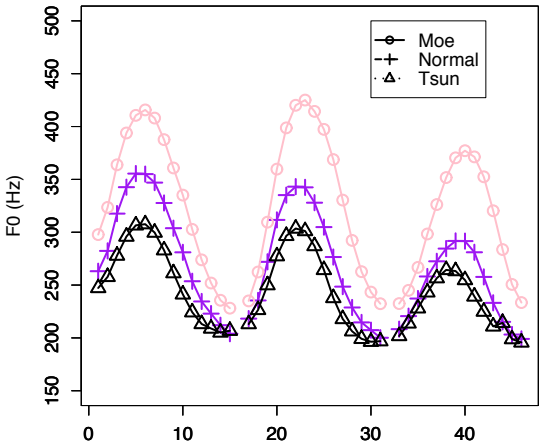

Figure 3: The normalized intonation contours of Speaker 1.

Another observation is that the differences between the three intonational contours manifest themselves more when the f0 is high (i.e. H-tone targets) than when it is low (i.e. L-tone targets). A `glm` model with the normal voice as a baseline reveals that the tsun voice and normal voice do not differ in the L tones ($t(318) = -0.94, n.s.$), but the moe voice is slightly higher than the normal voice ($t(318) = 5.62, p < .001$). Another `glm` model, targeting H tones, shows that the tsun voice is lower with respect to the normal voice ($t(318) = -7.65, p < .001$), and the moe voice is higher than the normal voice ($t(318) = 15.55, p < .001$). A Browne-Forsythe test, which compares degrees of variability, shows that there is more variation in the H-tone targets than in the L-tone targets ($F(1, 640) = 108.29, p < .001$), suggesting that speakers manipulate H-tone targets more to express different speech styles; i.e., it supports a model in Figure 1(b-1).

Can it be that H-tones inherently involve more natural variability than L-tones? To address this alternative, another Browne-Forsythe test was run using just the data from the object nouns of the normal voice—which is supposedly the "most neutral" environment (neither sentence-initial nor sentence-final)—to compare the natural variability between H-tones and L-tones. This test reveals no significant differences ($F(1, 72) = 1.07, n.s.$). Another alternative would be to say that H-tones show more variability due to syntactic differences than do L-tones. To address this alternative, another Browne-Forsythe test is run using only the data from object nouns including all the three speech styles. The result is significant ($F(1, 212) = 74.4, p < .001$). Therefore, the higher variability in H-tones seem to come about at least partly due to speech style differences.

The observation that a speaker expresses the speech style differences more clearly on

the H-tone targets than on the L-tone targets is compatible with the finding by Kawahara (2013). It also predicts that the magnitude of rises and falls should be bigger for the moe voice than for the normal voice, and bigger for the normal voice than for the tsun voice (see Figure 1(b-1)). The patterns in Figure 1(b-2) predict the opposite. Figure 1(a) predicts no differences across the speech styles. To test these predictions, the magnitude of rises and falls of Speaker 1 is shown in Figure 4 and 5, respectively. The error bars represent 95% confidence intervals.

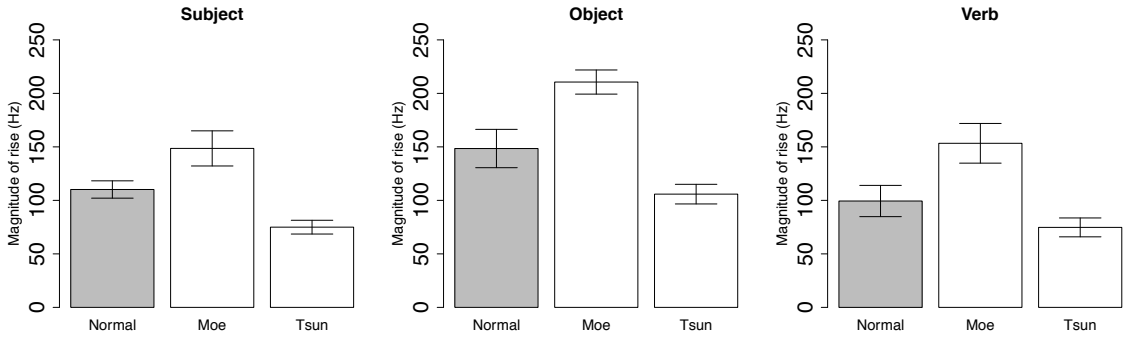

Figure 4: The magnitude of rise in three syntactic positions for Speaker 1. The error bars represent 95% confidence intervals.

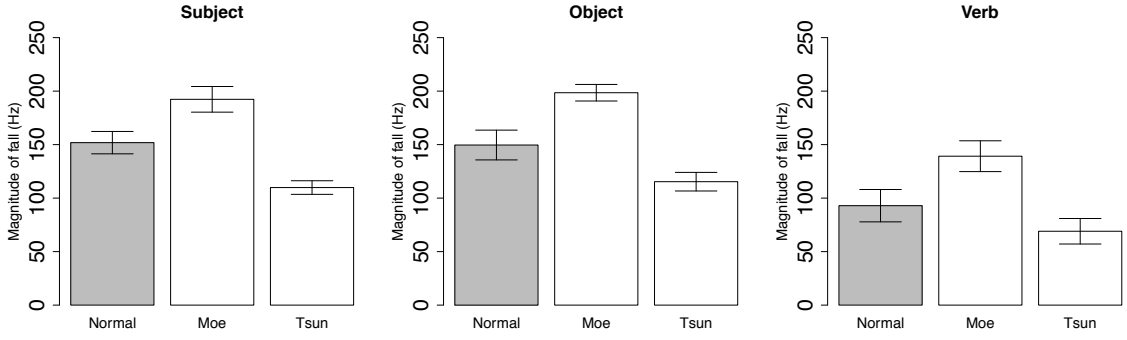

Figure 5: The magnitude of fall in three syntactic positions for Speaker 1. The error bars represent 95% confidence intervals.

Across all three syntactic positions (subject, object, verb), we observe that the moe voice shows the largest magnitude of rise and fall, the tsun voice shows the smallest, and the normal voice lies in-between. A large number of repetitive multiple comparisons (3 types of voice $\times$ 3 syntactic position $\times$ 2 rise and fall = 18 comparison) was avoided, as it would increase the probability of a Type 1 error (Myers & Well, 2003); however, the fact that the

error bars do not overlap seems to suggest that all differences between the three voice types are reliable throughout. These results corroborate the observation we made above that the speaker manipulates H-tones more than L-tones, rather than changing her whole pitch range.

Let us now move on to the next speaker. Figure 6 shows the intonational contours of the tsun, moe and normal voice of Speaker 2. As was the case for Speaker 1, the moe voice is generally higher than the normal voice, and the tsun voice is lower than the moe voice.

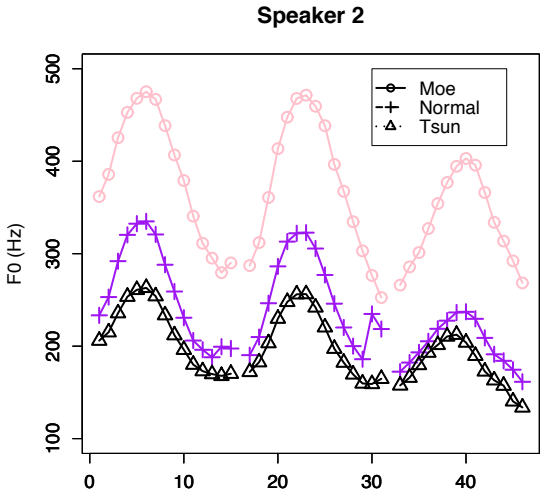

Figure 6: The normalized intonation contours of Speaker 2.

Unlike Speaker 1, we observe clear separations between the three voice types in the L-tones (normal vs. tsun: $t = -4.38, p < .001$; normal vs. moe: $t = 15.31, p < .001$). Nevertheless, we observe larger separations in the patterning of H-tones (normal vs. tsun: $-9.08, p < .001$; normal vs. moe: $t = 24.0, p < .001$.). A Browne-Forsythe test shows that H-tones show more variability than L-tones ($F(1, 640) = 129.41, p < .001$), indicating that this speaker, too, manipulates H-tone targets more than L-tone targets to express different speech styles.[4] We may be able to consider this pattern to be a combination of Figure 1(a) and Figure 1(b-1).

The magnitude of the rises and falls of Speaker 2 is shown in Figures 7 and 8. The fact that the moe voice shows larger rises and falls compared to the normal voice supports the previous observation that H-tone is raised more than L-tone. Also, similarly, the fact that

[4]No inherent differences are found in the variability of H-tones and L-tones in the object nouns of the normal voice ($F(1, 72) = 0.52, n.s.$). A Browne-Forsythe test using only the object noun data across the three speech styles reveals a significant effect ($F(1, 216) = 37.5, p < .001$), showing that the variability difference does not (solely) come from amenability to syntactic differences.

the tsun voice shows smaller rises and falls with respect to the normal voice indicates that the L-tone is not lowered as much as H-tones.

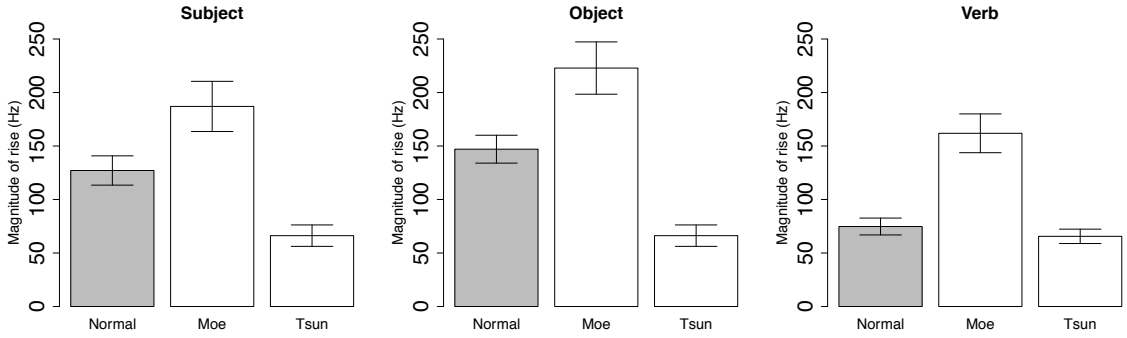

Figure 7: The magnitude of rise in three syntactic positions for Speaker 2.

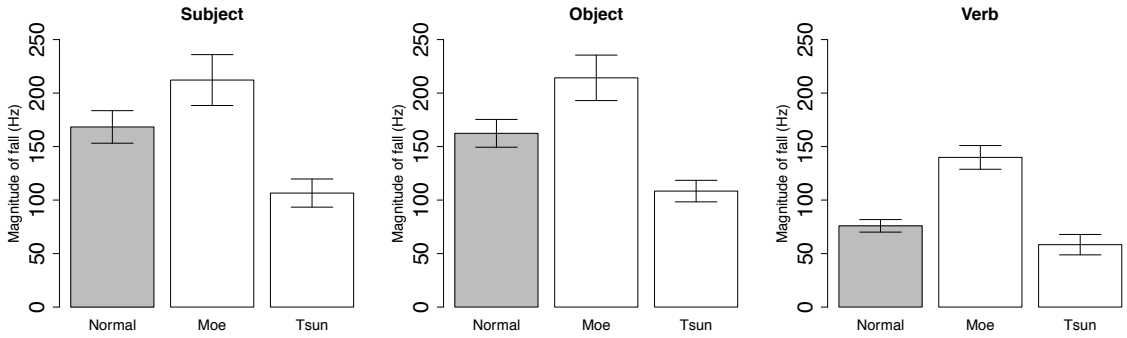

Figure 8: The magnitude of fall in three syntactic positions for Speaker 2.

4 Results: Intensity

Figures 9 and 10 show the average intensity of three voice types for Speakers 1 and 2. The results confirm the impressionistic observation stated in the method section—the moe voice is loud and the tsun voice is quiet. A `glm` analysis shows that for Speaker 1, the moe voice is louder than the normal voice ($t = 8.76, p < .001$), and the tsun voice is quieter than the normal voice ($t = -7.24, p < .001$). The same patterns holds for Speaker 2, although the difference between tsun voice and the normal voice is less clear. Statistically speaking, however, the moe voice is louder than the normal voice ($t = 8.42, p < .001$), and the tsun voice is indeed quieter than the normal voice ($t = -2.24, p < .05$).

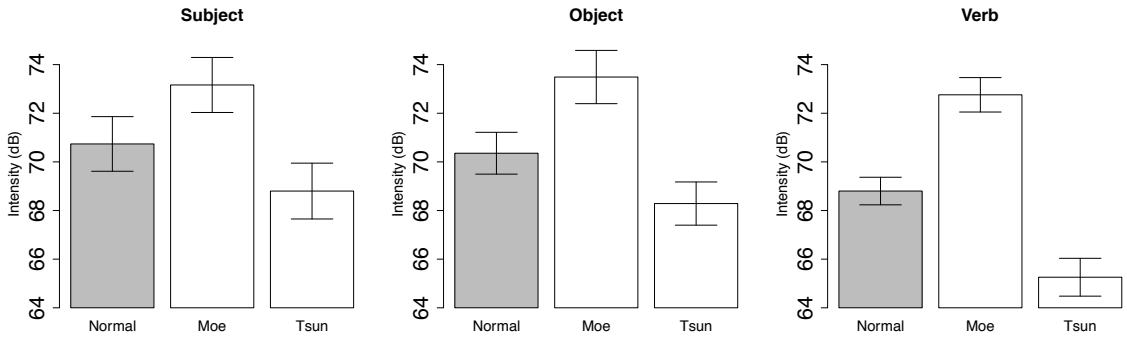

Figure 9: The average maximum intensity of three voice types for Speaker 1. The error bars are 95% confidence intervals.

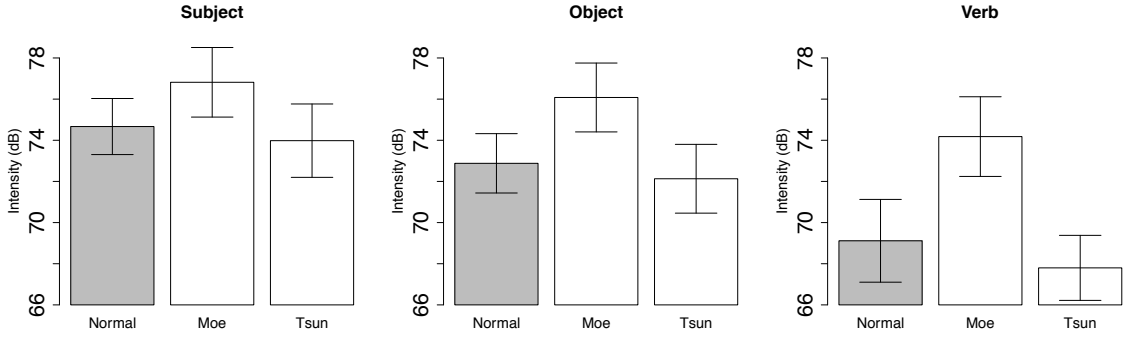

Figure 10: The average maximum intensity of three voice types for Speaker 2.

5 Discussion

For both speakers, the moe voice is generally characterized by higher voice with respect to their normal voice, whereas the tsun voice is characterized by lower voice. These patterns make intuitive sense: the moe voice, which capitalizes on "cuteness," involves higher voice. On the other hand, the tsun voice, which is associated with "inaccessibility", involves lower voice. We can perhaps understand these associations in terms of Ohala's (1984; 1994) frequency code hypothesis: the higher voice implies smallness and the lack of animosity, whereas the lower voice implies threat or inaccessibility. Also relevant is the finding that lower f0 is often associated with a "cold" angry voice (Chuenwattanapranithi et al. 2008, cf. Erickson 2005 for a "hot" angry voice), and that higher f0 is considered to be more attractive than lower voice for women's voice (Collins & Missing, 2003; Feinberg et al., 2008).

The two voice actresses recorded in this project, especially Speaker 1, change H-tone targets more than L-tone targets. This patterning is compatible with Fujisaki's model of intonation (Fujisaki, 1983; Fujisaki & Hirose, 1984), in which much of the intonational variation is made at high frequency levels rather than at the intonational baseline level (see also Pierrehumbert & Beckman 1988: Ch. 7). The moe voice of Speaker 2 may deviate from this model, but it is nevertheless true that this speaker changes H-tone targets more than L-tone targets. It could be the case that Speaker 2 is using falsetto—or what is known as "uragoe" in Japanese[5]—for the moe voice, which allows her to change her L-baseline more drastically than normal speech, but there is nevertheless a general pressure to keep L-tones constant across different voice styles, at least more so than H-tones. Alternatively, we can also understand the situation as follows: Speaker 2 wants to keep the L-tone baselines, but her H-tone targets for the moe voice are simply too high to start with her L-tone baseline. At any rate, the descriptive generalization is clear—the H-tone targets are raised more than the L-tone targets.

The intensity differences observed in this experiment can perhaps be interpreted in the following way. The moe voice represents a "lively" character who enunciates lucidly. The tsun voice on the other hand represents a "cool, calm and inaccessible" character, who would speak in a quieter voice. Admittedly, these are post-hoc explanations, and are not as explanatory as one wishes, because, as stated in the introduction, moe and tsun can be defined in terms of so many different parameters. At the very least, the intensity patterns seem to make intuitive sense.

Overall, this paper focused on the prosodic aspects of tsun and moe voice, which are relatively easy to study. However, it is certainly likely that there are other dimensions in which these voice types may differ; e.g., impressionistically speaking, the moe voice sounds more nasalized than the normal voice; the tsun voice sounds breathier than the normal voice. Studying these aspects is left for future research (see Sadanobu 2015, Starr 2015, Teshigawara 2003a, and Teshigawara 2003b in this regard). For example, Teshigawara (2003a,b) argues that villains' voices in anime are characterized by pharyngeal constriction—would we observe the same characteristics in tsun voice (although tsun girls are not exactly villains)? What about other acoustic dimensions that potentially affect "vocal attractiveness/asethetics"? (see Babel & McGuire 2015, Starr 2015 and references cited therein.)

One final remark: in terms of research value, the findings of this paper may not be groundbreaking. However, this sort of research has two additional benefits. One is that this kind of research is useful for phoneticians to reach out to the general public. Showing that phonetics

[5]An anonymous reviewer pointed out that the definition of *falsetto* "is not a term readily agreed upon". The precise vocal fold characteristics of this type of voice, however we call it, is an interesting topic for future study.

provides a way to scientifically analyze differences between tsun voice and moe voice can be interesting to those with no prior background in phonetics. Second, this sort of research is a useful teaching tool to attract students' attention. Phonetics is not always easy to teach, but having "interesting stuff" in class always helps to lower the resistance to learning something difficult. Indeed, two people have already asked me if they can use this material for their teaching. To that end, I am happy to share the real speech samples to those who are interested in using the voices of professional voice actresses for their teaching. For the time being, a sample of each type of sentence can be downloaded from `http://bit.ly/1WCu5DA`.

References

Babel, Molly & Grant L. McGuire (2015) Perceptual fluency and judgments of vocal aesthetics and stereotypicality. *Cognitive Science* **39**(4): 766–787.

Boersma, Paul (2001) Praat, a system for doing phonetics by computer. *Glot International* **5**(9/10): 341–345.

Chuenwattanapranithi, Sythathip, Yi Xu, Bundit Thipakorn, & Songrit Maneewongvatana (2008) Encoding emotions in speech with the size code. *Phonetica* **65**: 210–230.

Collins, Sarah A. & Caroline Missing (2003) Vocal and visual attractiveness are related in women. *Animal Behaviour* **65**: 997–1004.

Déprez, Viviane, Kristen Syrett, & Shigeto Kawahara (2013) The interaction of syntax, prosody, and discourse in licensing French *wh-in-situ* questions. *Lingua* **124**: 4–19.

Erickson, Donna (2005) Expressive speech: Production, perceptin and application to speech synthesis. *Ascouatical Science and Technology* **26**(4): 317–325.

Feinberg, David R., Lisa M. DeBruine, Benedict C. Jones, & David I. Perrett (2008) The role of femininity and averageness of voice pitch in asethetic judgments of women's voices. *Perception* **37**(4): 615–623.

Fry, D. B. (1955) Duration and intensity as physical correlates of linguistic stress. *Journal of the Acoustical Society of America* **27**: 1765–1768.

Fujisaki, Hiroya (1983) Dynamic characteristics of voice fundamental frequency in speech and singing. In *The Production of Speech*, Peter F. Macneilage, ed., New York: Springer, 39–55.

Fujisaki, Hiroya & Keikichi Hirose (1984) Analysis of voice fundamental frequency contours for declarative sentences of Japanese. *Journal of the Acoustical Society of Japan* **5**: 233–242.

Kawahara, Shigeto (2013) The phonetics of Japanese maid voice I: A preliminary study. *On-in Kenkyuu [Phonological Studies]* **16**: 19–28.

Kawahara, Shigeto, Kazuko Shinohara, & Joseph Grady (2015) Iconic inferences about personality: From sounds and shapes. In *Iconicity: East meets west*, Masako Hiraga, William Herlofsky, Kazuko Shinohara, & Kimi Akita, eds., Amsterdam: John Benjamins, 57–69.

Kawahara, Shigeto & Takahito Shinya (2008) The intonation of gapping and coordination in Japanese: Evidence for Intonational Phrase and Utterance. *Phonetica* **65**(1-2): 62–105.

Myers, Jerome & Arnold Well (2003) *Research Design and Statistical Analysis*. Mahwah: Lawrence Erlbaum Associates Publishers, mahwah.

Ohala, John J. (1984) An ethological perspective on common cross-language utilization of F0 of voice. *Phonetica* **41**: 1–16.

Ohala, John J. (1994) The frequency code underlies the sound symbolic use of voice pitch. In *Sound Symbolism*, Leane Hinton, Johanna Nichols, & John J. Ohala, eds., Cambridge: Cambrdige University Press, 325–347.

Pierrehumbert, Janet B. & Mary Beckman (1988) *Japanese Tone Structure*. Cambridge: MIT Press.

Plag, Ingo, Gero Kunter, & Mareile Schramm (2011) Acoustic correlates of primary and secondary stress in North American English. *Journal of Phonetics* **39**(3): 362–374.

R Development Core Team (1993–2016) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Sadanobu, Toshiyuki (2015) Four types of linguistic resources for variable speaking units in common Japanese. *Journal of the Phonetic Society of Japan* **19**(2): 109–114.

Shinohara, Kazuko & Shigeto Kawahara (2013) The sound symbolic nature of Japanese maid names. *Proceedings of Japan Cognitive Linguistics Association* **13**: 183–193.

Starr, Rebecca L (2015) Sweet voice: The role of voice quality in a Japanese feminie style. *Language in Society* **44**: 1–34.

Teshigawara, Mihoko (2003a) *Voices in Japanese animation*. Doctoral dissertation, University of Victoria.

Teshigawara, Mihoko (2003b) Voices in Japanese animation: How people percieve voices of good guys and bad guys. *Proceedings of ICPhS 15* : 2413–2416.

Togashi, Jun-ichi (2009) A case study of tsundere expression: Tsundere attributes and their linguistic expressions. Ms. Daito Bunka University.