VOICE TYPES AND VOICE QUALITY IN JAPANESE ANIME

Carlos T. Ishi[1,2], Akira Utsugi[3], Ichiro Ota[4]

[1] Guardian Robot Project, RIKEN, [2] ATR, [3] Nagoya University, [4] Kagoshima University
carlos.ishi@riken.jp, utsugi@nagoya-u.jp, iota@leh.kagoshima-u.ac.jp

ABSTRACT

Japanese anime is noteworthy for being dubbed by professional voice actors and is unique in its vocal characteristics. In this study, recordings were taken of students of voice actor course at a music college reading short texts and anime lines in various voice types. These recordings were analyzed by focusing on the acoustic parameters related to voice quality. The results showed that for a repeating task, when students were asked to repeat an actual animated voice, acoustic parameters such as F0 and sub-band harmonics-to-noise ratio (HNR) clearly captured the differences between several anime characters. When analyzing recordings of short texts in various voice types, a similar trend was found, although not as distinct as the repeating task. These results suggest that multiple anime voice types have been established in Japanese anime and voice quality helps to distinguish them.

Keywords: anime, voice quality, fundamental frequency, sub-band HNR, register

1. INTRODUCTION

Japanese anime voices are unique from the perspective of the voice culture. Unlike Western animation, which is often performed by actors, a majority of Japanese anime is dubbed by professional voice actors called “seiyu.” Further, anime voices of children are performed by adult female seiyus [1]; this indicates that Japanese anime voices have been unique from its early days. In contemporary Japan, seiyu is a popular profession among young people [2]. This background and contemporary trend are speculated to have influenced the characteristics of the anime voices. It has been suggested that multiple anime voice types are becoming established in contemporary anime, and can be viewed sociolinguistically as registers [3]. One of the aims of this study is to determine whether voices performed by different voice actors have similar vocal characteristics if the types of characters they perform are similar.

The study of anime voices has several approaches, that can be categorized by their descriptive and data-collection methods.

Impressionistic and quantitative methods form the descriptive categorization. Anime voices are often described impressionistically when discussed from the perspective of media studies. In contrast, sociophonetic studies have attempted to acoustically analyze voices via characteristics such as F0 [4-5] and voice quality [3,6-8].

Further, the data collection has two approaches: analyzing actual animated speech [5-8], and obtaining multiple types of speech from the same speaker through researcher intervention [3-4].

One of the goals of this study is to examine how the voices of different types of anime characters exhibit different characteristics in terms of acoustic features related to voice quality. Although previous studies have analyzed the acoustics of anime voices, few studies have been conducted on character-specific analysis of contemporary Japanese anime. Another goal of this study is to determine whether character voices have similar characteristics, regardless of the individual seiyus.

This study compared multiple types of speech recorded from the same speaker under the same conditions. As previous studies using a similar approach were ambiguous in terms of whether the listener was receiving the speech as intended by the speaker, this study conducted acoustic analysis only on recorded speech that met certain impression ratings.

2. METHOD

2.1. Data collection

Voice performances from 11 young females in their early 20s, who were students of a seiyu course at a music college in Japan, was collected. Their speech utterances were digitally recorded using an audio interface (PerSonus AudioBox iTwo) and a condenser microphone (PreSonus M7). The sampling and quantization bit rates were 44100 Hz and 24 bits, respectively.

Data collection comprised four different tasks structured in the following order: sentence reading (Part I), passage reading, dubbing, repeating, and sentence reading (Part II). Next, 15-minute informal interviews were conducted to collect tokens in a natural setting. This study only reports the results for the passage reading and repeating tasks.

For passage reading, the speakers read "The North Wind and the Sun" by performing four different characters (plus with the speaker's own voices). The characters were of the same types as the anime characters they were to perform in the dubbing and repeat tasks. However, at this stage, the characters were simply designated as a gentle adult woman or older sister ("ordinary"), a strong, bold type like a woman warrior ("brave"), a sexually neutral or immature girl ("boyish"), and a kawaii girl ("cute"), without mentioning the title of the anime and the names of the characters.

In the repeating task, the speakers imitated the performance of professional seiyus in several scenes from an anime immediately after hearing them. The anime for this task was *Kono Subarashii Sekai-ni Shukufuku-o!* ("God's Blessing on this Wonderful World!"), released in 2016. Six scenes from this anime were selected, and the statements spoken by the four characters in these scenes were analyzed. The targeted characters corresponded to the four character types defined in the passage reading task: Aqua (hereafter, "AQ," ordinary), Darkness ("DN," brave), Megumin ("MG," boyish), and Yunyun ("YY," cute).

2.2. Impression ratings

Three listeners (research assistants who did not know the anime characters) annotated the perceived character impressions, from the voices of all 11 speakers. The character impressions were graded over three levels (1 = slightly, 2 = clearly, 3 = strongly) for each of the four character types. Data where two or more annotators assigned level 2 or above to the target character, were used for acoustic analysis.

In the passage task, the number of speakers analyzed based on impression ratings was as follows: ordinary: 9 speakers, brave: 3 speakers, boyish: 8 speakers, cute: 9 speakers.

In the repeating task, the data of the utterances with very strong emotional expressions, when the impressions of the characters changed, were omitted from acoustic analysis.

2.3. Acoustic parameters

F0 and voice-quality-related parameters were extracted. Whereas a previous study analyzed acoustic parameters related to formants and spectral slopes [3], this study focused on harmonicity analysis.

F0 was extracted by filtering the maximum peak of the autocorrelation function (ACF) of the speech signal filtered through a low frequency band of 100–1200Hz. This band guaranteed the presence of one or more harmonics in this range, even for very high pitch voices. The ACF of the speech signal without traditional LPC inverse filtering is more robust, as anime voices include very-high-pitch voices. F0 was extracted at 32-ms frame lengths, with a 10-ms frame step. Subsequently, it was converted to the log scale with semitone intervals, according to:

$$F0[semitone] = 12 * log_2(F0[Hz]) + c \qquad (1)$$

where c = -36.376. Hence, $F0[semitone]$ = 57 corresponds to 220Hz (A3 on the musical scale) [9].

For voice quality, we tested several parameters related to harmonicity. By considering that spectral harmonicity vary at different frequency bands depending on vocal tension (stiffness), we analyzed sub-band harmonic-to-noise ratio (HNR).

The sub-bands were divided as follows: sub-band 1 (100–1200 Hz), 2 (1000–2200 Hz), 3 (1800–4000 Hz), 4 (4000–6000 Hz), 5 (5500–7500 Hz).

Sub-band HNR was computed in dB using an autocorrelation-based method [10], with adaptation to sub-band computation, according to:

$$HNR[k] = 10 * log_{10}\left(\frac{r[k][\tau_{max1}]}{1-r[k][\tau_{max1}]}\right) \qquad (2)$$

where k is the sub-band number (1 to 5), τ_{max1} is the lag of the maximum peak in the ACF of sub-band 1, and $r[k][\tau_{max1}]$ is the autocorrelation value of sub-band k at lag τ_{max1}, normalized by $(T - \tau_{max1})/T$, where T is the window length.

Note that τ_{max1} of sub-band 1 is used to compute the HNR of all the sub-bands assuming that harmonic components in all frequency bands are compatible with the harmonicity of the fundamental frequency obtained from the sub-band 1 peak. However, a small range around τ_{max1} was searched to account for slight deviations between the ACF peaks in different bands. Accordingly, HNR[k] will have positive values if harmonic components are predominant, and negative values if noise components are predominant in each sub-band k. Thus, high HNR values in high-frequency bands are expected for voices with high vocal tension/stiffness.

The mean values of each parameter were extracted as representative values for each utterance, and the mean values of the utterances for each character type performed by each speaker were used as representative values for each speaker.

3. RESULTS

Fig. 1 shows the scatter plots of the average F0 and HNR values of each speaker for each target character in the repeating task. The upper panel shows the distributions for HNR[3] (1800–4000Hz), whereas the lower panel shows the distributions for HNR[5]

(5500–7500Hz). The small red circles indicate the data of the professional seiyus (extracted from the original DVDs). The results for HNR[2] and HNR[4] were omitted because of redundancies.

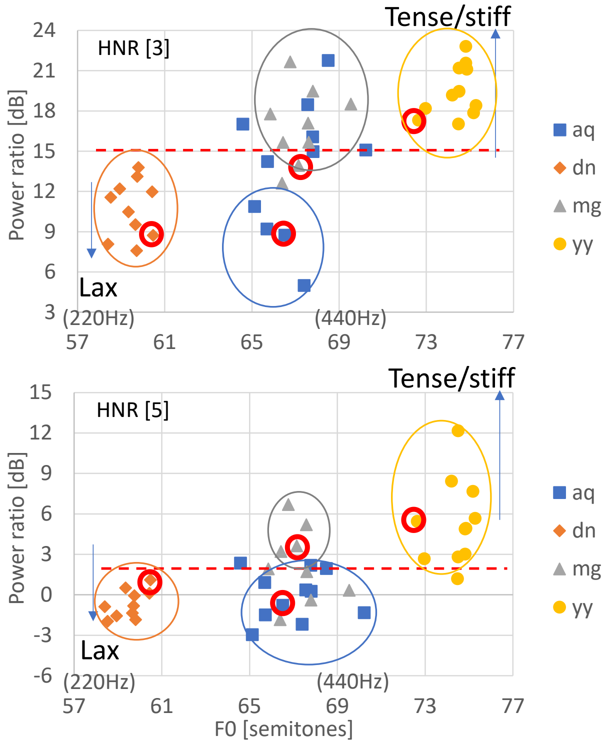

Figure 1: Distributions of the average F0 and HNR values (HNR[3] and HNR[5]) for each target character in the repeating task.

For each acoustic parameter, between-subject ANOVA was conducted with the speakers as random variables and character types as independent variables. Although all speakers performed all character types, a between-subject design was adopted as some samples were omitted through the impression ratings in Section 2.2. Multiple comparisons were performed using the Bonferroni method. Significant main effects were found for all parameters: F(3,33) = 273.75, $p < 0.01$, for F0; F(3,33) = 13.42, $p < 0.01$, for HNR[3]; and F(3,33) = 12.6, $p < 0.01$, for HNR[5].

From Fig. 1, DN and YY can be observed to show distinct distributions in F0, HNR[3] and HNR[5] (DN < YY; $p < 0.05$). The distributions of AQ and MG have some overlap, but MG tends to have a tenser voice than AQ, as observed in HNR[3] (AQ < MG; $p < 0.05$). However, the distributions in HNR[5] show that YY has a tenser voice quality than the other characters (YY > MG,DN,AQ; $p < 0.01$).

The results for the professional seiyus (small red circles in Fig. 1) indicate that the voices imitated by the students were close to the mean values of the seiyus.

Fig. 2 shows the distributions of passage reading task. The target characters were “brave”, “boyish”, “ordinary”, and “cute”. The distributions for “self” indicate the reading style in the speaker’s natural voice.

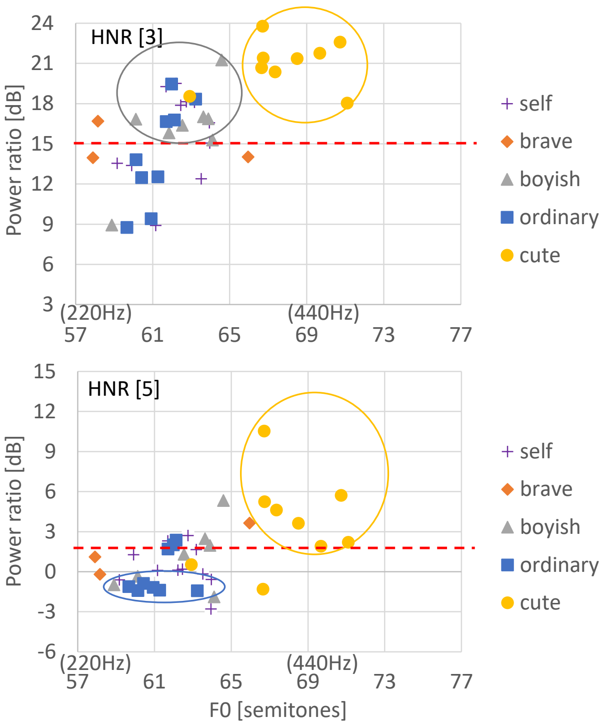

Figure 2: Distributions of the average F0 and HNR values (HNR[3] and HNR[5]) for each target character in the passage reading task.

Between-subject ANOVA and multiple comparisons were conducted similar to that done for the repeating task. Significant main effects were found for all parameters: F(4,35) = 14.5, $p < 0.01$, for F0; F(4,35) = 6.18, $p < 0.01$, for HNR[3]; and F(4,35) = 3.75, $p < 0.05$, for HNR[5].

The distributions in Fig. 2 differ from the results of the repeating task. Overall, the F0s in the passage reading task (Fig. 2) are lower than those in the repeating task (Fig. 1). This is because utterances in the repeating task were mostly accompanied by emotional expressions (such as surprise and excitement), whereas sentences in the passage reading task were mostly emotionless.

However, some similarities can be observed in the HNR[3] and HNR[5] distributions between some voice types. For example, the distributions of “cute” shows similar trends with YY, with a high-pitched tense voice (YY > MG,AQ,DN; $p < 0.05$). The distributions of “boyish” are also similar with MG character, i.e., mostly higher HNR[3] values and mostly lower HNR[5] values. However, the distributions between “ordinary” and AQ are less similar. A possible reason is that “ordinary” may not be an appropriate attribution for AQ. In fact, in the preliminary analysis, utterances in the repeating task were mostly perceived as “cute girl” rather than “adult female.” Regarding “brave”, data from only three of the speakers could be used for the acoustic analysis, of which two had very low F0s as in DN, while one of the speakers was perceived as “brave+boyish,” with higher F0 and tenser voice quality. Finally, regarding the distributions of “self”, most of the speakers have voices closer to “ordinary” or “boyish”. This is also in agreement with subjective impressions.

4. DISCUSSION AND CONCLUSION

4.1. Character types and acoustic parameters

A previous study on anime character types identified significant differences for F0 but not for HNR [3]. In contrast, the present analysis successfully detected between-character differences for HNR as well. This is because we conducted sub-band HNRs instead of full-band HNR in this study. Another reason would be that we conducted the acoustic analysis after impression ratings.

Character classification in this study was comparable to that in Kawahara's study [4]. He compared the F0s of two characters, "moe" and "tsun," and their normal voices. According to him, "moe" can be characterized by a set of adjectives like "pretty, cute, … and cheerful", whereas "tsun" was characterized as "beautiful, tall, … and inaccessible". Given that "moe" coincides with "cute" and "tsun" overlaps with "brave" in this study, the high and low F0 values for "cute" and "brave", respectively, are consistent with the results of Kawahara's study.

A variety of acoustic parameters related to voice quality were analyzed by Starr [7]. However, directly comparing Starr's character classification with that used in this study is difficult. She compares “sweet” and “nonsweet” voices in Japanese anime, where sweet voice characters include mothers, older women, etc. Although her classification is different from ours, it has a similar implication to this study in that it shows that pitch and voice quality are related to distinction in characters.

Regarding the interpretation of the results for HNR[3] and HNR[5], which reflect the levels of harmonicity components in the mid (1800–4000Hz) and high (5500–7500Hz) frequency bands, it is important to clarify that they can be affected by different factors. First, tense voices tend to have stronger harmonicity components in higher frequency bands than lax voices. Second, vocal tract length may also affect harmonicity in different frequency bands. Longer vocal tracts (male speakers) tend to have lower formant frequencies, and consequently weaker harmonicity components in higher frequency bands compared with shorter vocal tracts (females and children). The HNR[5] differences between the distributions of “cute” and “boyish” could be because both are tense qualities but “cute” is generated by raising the larynx (i.e. shorter vocal tract length), in comparison to “boyish”. Finally, HNR is also strongly affected by breathy or whispery voices, which contain higher levels of turbulent noise in mid- and high-frequency components. In the analyzed dataset, breathy and whispery voices were locally observed primarily at the end of phrases. However, they were not strongly reflected in the analysis in this present study, since mean values at utterance level were used. Such local breathiness is an important factor in emotional expression; therefore, a more detailed analysis is required in future investigations.

4.2. Consistency of distribution across speakers

Notably, voice characteristics of the participants were distributed close to each character. Such a distribution in the repeating task may not be surprising as the participants were asked to listen to actual anime voices and replicate them. However, this trend was also observed to some extent in the passage reading task, wherein participants were not asked to imitate a specific character. This suggests that anime character types are socially shared.

However, the results for the passage reading task were not as clear as those of the repeating task. In addition, many impressions were not judged as intended during impression ratings and were excluded from analysis. This may be either because the passage reading task was difficult for the participants, who were still in the process of training to become seiyus, or because some of the characters are difficult to be expressed through a reading task. In the passage reading task, the participants had to pronounce the passage based only on the instruction without professional model voices.

Finally, the balance of the emotions included in the utterances differed depending on the anime character type. Therefore, distinguishing character differences from differences in emotion in this study is difficult, and separating these two factors will be the subject of a future study.

5. ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Number JP21K18116 and JP22H04875. We would like to thank KADOKAWA and Konosuba Partners for their cooperation.

6. REFERENCES

[1] Ishida, M. 2020. *Anime to Seiyu no Mediashi: Naze Josei ga Shonen o Enjirunoka*. Tokyo: Seikyusha.

[2] Iwata, M. 2017. Seiyu-do. Tokyo: Chuokoron-shinsha.

[3] Ota, I., Ota, Y., Utsugi, A. 2021. The voice quality of anime as a sociocultural register. A paper presented at the 16th International Conference of the European Association of Japanese Studies.

[4] Kawahara, S. 2016. The prosodic features of the “moe” and “tsun” voices. *Journal of the Phonetic Society of Japan* 20 (2), 102–110.

[5] Marushima, A. 2020. Josei seiyu ni yoru yakugara no seibetsu no kotonaru onsei no onkyoteki tokucho: Kihonshuhasu ni chakumoku shite. *Osaka-keizaihokadaigaku Ronshu* 115, 23-33.

[6] Teshigawara, M. 2003. Voices in Japanese animation. Doctoral dissertation, University of Victoria.

[7] Starr, R. L. 2015. Sweet voice: The role of voice quality in a Japanese female style. *Language in Society* 44, 1–34.

[8] Utsugi, A., Wang, H., Ota, I. 2019. A voice quality analysis of Japanese anime. *Proc. 19th ICPhS* Melbourne, 1853–1857.

[9] Ishi, C.T., Ishiguro, H., Hagita, N. 2008. Automatic extraction of paralinguistic information using prosodic features related to F0, duration and voice quality. *Speech Communication* 50(6), 531-543.

[10] Boersma, P. 1993. Accurate short-term analysis of the fundamental frequency and the harmonic-to-noise ratio of a sample sound. *IFA Proceeding* 17, 97-110.