A VOICE QUALITY ANALYSIS OF JAPANESE ANIME

Akira Utsugi[1], Han Wang[1], Ichiro Ota[2]

[1] Nagoya University, [2] Kagoshima University
utsugi@nagoya-u.jp, wangh199110@gmail.com, iota@leh.kagoshima-u.ac.jp

ABSTRACT

Japanese anime cartoons are characterized by their unique voices, which poses an interesting research opportunity in the study of phonetics. To identify the acoustic characteristics of Japanese anime voices, we analyzed voice quality in Japanese anime. Six *seiyu* (vocal actor/actress) voices from two recent anime cartoons were included in our analysis. We also analyzed normal Japanese speech from the Corpus of Spontaneous Japanese (CSJ) as control data. Using openSMILE, software for voice quality analysis, we measured several acoustic parameters related to voice quality. The results demonstrate that anime has significantly different values than normal speech does in spectral slope-related parameters such as H1-A3, F1~F3 relative energy, alpha ratio, Hammarberg index, and spectral slope (0.5–1.5 kHz). The results suggest that an anime voice has stronger harmonics than normal speech does.

Keywords: anime, *seiyu* (voice actor/actress), voice quality, phonation, spectral slope

1. INTRODUCTION

One of the most familiar genres of contemporary Japanese popular culture worldwide is anime cartoons. Approximately 60% of the world's anime cartoon series are produced in Japan [3]. One major characteristic of Japanese anime cartoons is the voices produced by *seiyu*s, that is, professional voice actors/actresses. These unique voices have gained attention from phonetic and sociophonetic studies [6][13][15]. In particular, female characters' voices are regarded as part of a distinctive female vocal style in Japan that Starr called the "sweet voice" [13]. This vocal style is widely observed in Japan and has been a subject of study for sociophonetics [13] as well as social and cultural studies about Japan [12].

Interestingly, circumstances related to anime are changing. As the Japanese anime industry grows, the number of *seiyus* also increases. Being a *seiyu* is a popular occupation for high school students in Japan, and the number of would-be *seiyus* is estimated to be around 300,000 [5]. We speculate that this increase is related to changes in culture, society, and media in Japan, and such essential changes may have affected the traditional anime voice. Our long-term goal is to understand how Japanese anime voices have changed over time, how they are different from cartoon voices in other countries, and which factors in culture, society, and media have influenced the formation of Japanese anime voice.

As the first step of this project, we identify the acoustic parameters that characterize anime voices. The literature has analyzed prosody [6] and voice quality [13][15], and voice quality is the focus of this study. In this study, we incorporate recent voice quality studies and include several acoustic parameters not measured in the literature on Japanese anime. Another novelty of our study is that we deal with relatively recent anime cartoons and compare them with normal speech.

One notable voice quality study proposed a standard parameter set called the "Geneva Minimalistic Acoustic Parameter Set" (GeMAPS) [2]. This parameter set comprises 18 low-level descriptors related closely to the production and perception of voice. GeMAPS was used, for example, in emotional speech studies of French opera [11] and of Japanese [16]. We use these parameters in our measurements.

2. METHODS

2.1. Data

We investigated two popular Japanese anime cartoons: *Kono Subarashii Sekai-ni Shukufuku-o!* (hereafter, "*Konosuba*," English title: "God's Blessing on this Wonderful World!" released in 2016) and *Gochuumon-wa Usagidesuka?* (hereafter, "*Gochiusa*," English title: "Is the Order a Rabbit?" released in 2014–15). Six *seiyu* voices (Sora Amamiya and Aki Toyosaki from *Konosuba* and Ayane Sakura, Inori Minase, Risa Taneda, and Satomi Sato from *Gochiusa*), for which permission for our academic use was granted by the parties concerned, were the targets of the analysis.

To compare anime speech with normal speech, we used the Corpus of Spontaneous Japanese (CSJ) developed by the National Institute for Japanese Language and Linguistics [8][9] to obtain the control data. Among several instances of data in the CSJ, tokens were extracted from eight female speakers' speech in the Simulated Public Speaking set. Speakers in their 20s and early 30s were selected to

make the control speakers' ages closer to the *seiyus*' ages.

2.2. Acoustic analysis

We used Praat (P. Boersma and D. Weenink) to label and segment and openSMILE (audEERING) to measure the acoustic parameters related to voice quality.

Since speech is often overlapped with background music in anime, the parts that were not overlapped with music were manually detected. For those parts, utterances were segmented. Then, the character name was labeled, and vowels were segmented for each utterance. Only vowels longer than 60 ms were extracted for the analysis since openSMILE fails to measure acoustic parameters when a vowel is shorter. As this paper is an interim report for the data for which segmentation has not yet been completed, 455 tokens (vowel parts) were subjected to the present analysis.

As for normal speech (CSJ), tokens were extracted based on distributed segmentation data. To make the number of tokens close to that of the anime group, 448 tokens were extracted for the present analysis.

Since the sampling rate of the anime sound files (48 kHz) was different from that of the CSJ (16 kHz), the anime files were downsampled to 16 kHz.

We then measured the 18 acoustic parameters composing the GeMAPS using openSMILE: frequency-related parameters such as pitch, jitter, formants 1, 2, and 3 frequency, formant 1 bandwidth, energy/amplitude-related parameters such as shimmer, loudness, harmonics-to-noise ratio (HNR), spectral (balance) parameters such as alpha ratio, Hammarberg index, spectral slope 0–0.5 kHz and 0.5–1.5 kHz, formants 1, 2, and 3 relative energy, harmonics difference H1-H2, and harmonics difference H1-A3.

Following Scherer et al. [11], we subtracted the alpha ratio and spectral slope (500–1500 Hz) values from 1 to make them comparable to other spectral slope-related parameters.

3. RESULTS

We calculated mean values for each acoustic parameter per speaker and vowel and then employed a two-way mixed ANOVA with two independent variables, including a between-subject variable TYPE (anime, normal) and a within-subject variable VOWEL (/a, i, u, e, o/) for each acoustic parameter. Since one speaker's data in anime lacked /e/ and /o/ tokens, this speaker's data were not included in the statistical analysis. The analysis revealed that the main effect of TYPE was significant for 12 parameters but insignificant for 6 parameters: loudness, jitter, shimmer, H1-H2, F2 frequency, and F3 frequency (Table 1).

Table 1: Main effect of TYPE on 18 acoustic parameters

Parameters	$F(1,11)$	p
Pitch	25.326	0.000
Loudness	0.010	0.922
Jitter	4.503	0.057
Shimmer	0.007	0.934
HNR	49.487	0.000
H1−H2	0.200	0.663
H1−A3	17.067	0.002
F1 frequency	32.465	0.000
F1 bandwidth	7.657	0.018
F1 relative energy	12.858	0.004
F2 frequency	1.988	0.186
F2 relative energy	19.001	0.001
F3 frequency	0.070	0.796
F3 relative energy	16.944	0.002
Alpha ratio	41.337	0.000
Hammarberg index	17.724	0.001
Slope (0–0.5k)	35.354	0.000
Slope (0.5–1.5k)	27.542	0.000

In the frequency-related domain, anime speech demonstrated a higher pitch, narrower F1 bandwidth, and higher F1 frequency than normal speech.

In the energy/amplitude-related domain, anime speech involved a lower HNR than normal speech. However, the effect of loudness was not significant.

In the spectral slope-related domain, anime speech had a lower spectral slope for all frequency bands (i.e., lower H1–A3, higher F1~F3 relative energy, lower alpha ratio, lower Hammarberg index, lower slope 0.5–1.5 kHz) than normal speech does, except for the spectral slope (0–0.5 kHz), in which amine demonstrated a higher slope than normal speech did.

Boxbar plots for spectral slope (0.5–1.5 kHz), alpha ratio, Hammarberg index, and HNR are shown in Figures 1–4.

4. DISCUSSION

We measured acoustic parameters that had not been treated in previous comparisons between anime and normal speech. The most remarkable results were found in the spectral slope-related domain: most of the parameters demonstrated that anime speech had a lower spectral slope than normal speech did. This result suggests that anime speech is characterized by higher energy in upper harmonics than normal

Figure 1: Spectral slope (0.5–1.5 kHz) results. Error bars exhibit ± 1 standard error.

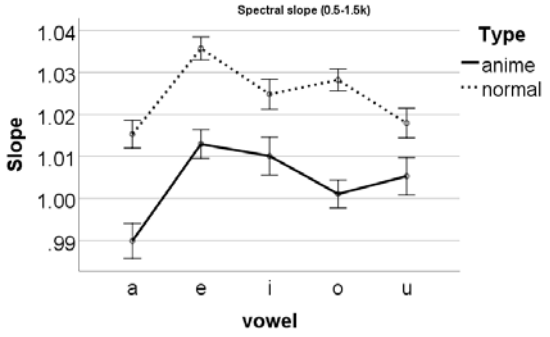


Figure 2: Alpha ratio results. Error bars exhibit ± 1 standard error.

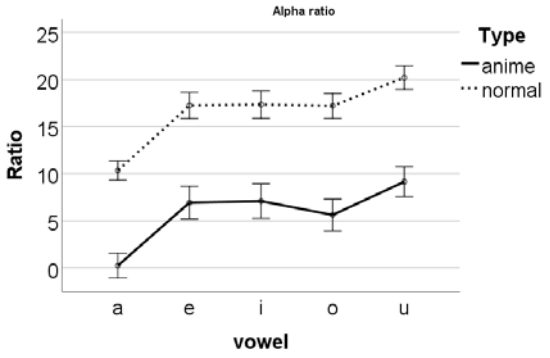


Figure 3: Hammarberg index results. Error bars exhibit ± 1 standard error.

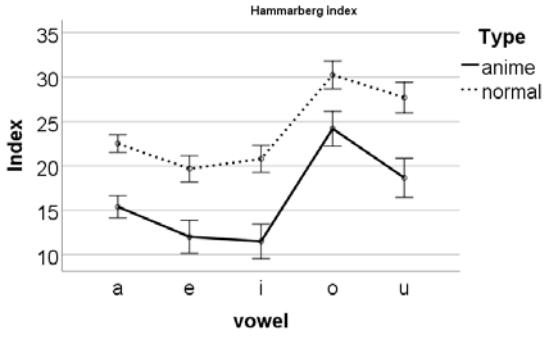
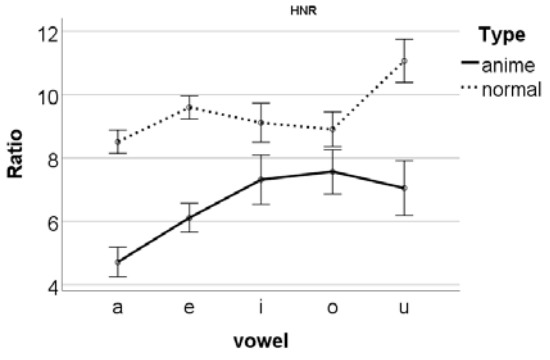


Figure 4: HNR results. Error bars exhibit ± 1 standard error.



speech has. An exception was spectral slope (0–500Hz), in which anime demonstrated a higher spectral slope than normal speech has. We are cautious in interpreting this parameter because F0 tends to affect this parameter. A token with F0 less than 250 Hz contains two harmonics in the 0–500 Hz range, whereas that with F0 between 250 Hz and 500 Hz contains one harmonic, and that with F0 greater than 500 Hz contains no harmonic. In anime speech, F0 values are greater than 250 Hz in many cases and are sometimes greater than 500 Hz. For this F0 characteristic, the parameter of the spectral slope (0–500 Hz) does not tend to reflect harmonic energy in anime; thus, we consider that this parameter is not appropriate for analyzing anime voices.

In the other domains, anime speech was characterized by a lower HNR. These results might be contradictory to the results of the spectral slope. A lower HNR has been considered as the reflection of a rougher and breathier voice (e.g., [1]), and a lower spectral slope has often been considered as an indicator of creaky phonation (e.g., [4]). However, more recent studies on voice quality interpret these parameters in a different way. According to Scherer and his colleagues, HNR is related to "phonation perturbation (which can be produced by both hyper- and hypotension of vocal fold adduction)" [10] or "vocal fold length and tension" [14], whereas some spectral slope-related parameters such as alpha ratio are related to subglottal pressure [14].

It is interesting to compare our results with Starr's study of Japanese anime voices [13]. She compared a "sweet voice" with the same *seiyu*'s "non-sweet voice" in anime. One may consider that Starr's comparison between sweet and non-sweet voices is equivalent to our comparison between anime and normal voices. Here, we would like to focus on four parameters in Starr's results: H1–H2, H1–A3, HNR, and 2k–4k. Although we did not measure 2k–4k, we consider the alpha ratio and Hammarberg index as equivalent parameters in that these reflect the spectral slope in the frequency region close to 2k–4k. Alpha ratio is defined as "the ratio between the summed energy from 50–1000 Hz and 1–5 kHz," and the Hammarberg index is defined as "the ratio of the strongest peak in the 0–2 kHz region to the strongest peak in the 2–5 kHz region" [2]. When we focus on these four parameters, several differences are found between the two studies. In Starr's study, the sweet voice showed higher H1–H2, H1–A3, and HNR values and lower 2k–4k values than the non-sweet voice. In our study, the anime voice showed lower H1–A3, HNR, alpha ratio, and Hammarberg index values than a normal voice, and H1–H2 was not significant. In short, the

spectral slope in the higher region showed a similar tendency, whereas H1-A3 and HNR showed opposite tendencies between the two studies. However, this is not surprising because we cannot identify our anime voice with Starr's sweet voice for the following two reasons. First, Starr's sweet voice characters and the characters we analyzed are different. In Starr's study, the sweet voice characters included a mother, an older girl, and a student council president while the non-sweet voice characters included a lieutenant, a marine officer, a princess, and a flight instructor. On the other hand, in our study, more than half of the characters were girls. Second, the two studies dealt with anime from different generations. Starr focused on anime produced in 1984 and 2005, whereas ours were produced in 2014 and 2016. In our impression, recent Japanese anime such as our targets are very different from older anime in terms of voice quality. For these two reasons, we consider that our targets do not fit Starr's sweet/non-sweet classification.

Interestingly, an acoustic analysis of a falsetto voice showed that it had lower HNR than a modal voice [7]. Although anime voices do not sound like falsetto, it is possible that anime voices have some similarities to falsetto voice.

Kawahara [6] analyzed the parameters pitch and loudness. His study compared two types of anime voices, "moe" and "tsun," with normal speech. He observed that the "moe" voice was higher (in terms of pitch) and louder than a normal voice, whereas the "tsun" voice was lower and quieter than a normal voice. In our study, on average, the anime voice had a higher pitch; however, we did not categorize voice type such as "moe" and "tsun." We speculate that high pitch was extremely high and low pitch was moderately low in anime compared with normal speech; thus, average pitch appeared higher in anime than in normal voice. Additionally, a loud voice and a quiet voice may simply cancel each other out. In any case, pitch and loudness were not observed to be consistent parameters in characterizing anime voices.

It should be noted that we have limitations in understanding the acoustic features of anime voices from this preliminary study. Voice quality parameters are affected by individual voice differences and recording settings. Since the two types of voice in this study came from different groups of speakers in a relatively small sample size, we cannot rule out individual bias. In addition, we cannot rule out the influence of recording settings in the two recording sources (anime and corpus). Our next step needs a different approach to minimize these biases.

5. CONCLUDING REMARKS

Our voice quality analysis revealed that contemporary Japanese anime voices were characterized by a lower spectral slope, narrower F1 bandwidth, and lower HNR compared with normal voices.

There are several possibilities regarding where these characteristics originate. For example, they might exist because of a *seiyu*'s voice control skills, a *seiyu*'s original physical voice traits (i.e., people with certain voice characteristics tend to become *seiyus*), or both (i.e., anime voice is achieved through physical voice talent and trained voice control skills). Further research to examine these possibilities could be realized through a within-speaker study to compare normal and anime voices by *seiyus*.

In addition, it is unclear whether the voice characteristics identified in this study are also observed in other countries' cartoons and whether they date back to past Japanese anime. If these characteristics are observed in contemporary Japanese anime only, another question arises of whether Japanese anime developed in that manner, which should be answered from interdisciplinary perspectives.

6. ACKNOWLEDGMENTS

We are grateful to KADOKAWA, NBC Universal Entertainment Japan, Konosuba Seisaku Iinkai (Konosuba Production Committee), and the voice actor agencies (I'm Enterprise, Music Ray'n, Office Osawa, Aoni Production, Accel One, and 81 Produce) for allowing us to use the voices in the anime DVDs for this research. This work is supported by JSPS KAKENHI Grant Number 17K18485. We also thank Yoshitaka Ota for his invaluable comments on the voice as a cultural product in contemporary Japanese society.

7. REFERENCES

[1] de Krom, G. 1995. Some spectral correlates of pathological breathy and rough voice quality for different types of vowel fragments. *J. Speech. Lang. Hear. Res.* 38, 794–811.

[2] Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L.Y., Epps, J., Laukka, P., Narayanan, S. S., Truong, K. P. 2015. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Trans. Affect. Comput.* 7, 190–202.

[3] Frederic, Jim. 2003. What's right with Japan. Time Asia. http://www.time.com/time/asia/2003/cool_japan/story.html. Retrieved December 9, 2018.

[4] Gordon, F., Ladefoged. P. 2001. Phonation types: A cross-linguistic overview. *J. Phon.* 29, 383–406.
[5] Iwata, M. 2017. *Seiyu-do*. Tokyo: Chuokoronshinsha.
[6] Kawahara, S. 2016. The prosodic features of the "moe" and "tsun" voices. *Journal of the Phonetic Society of Japan* 20 (2), 102–110.
[7] Keating, P.A. 2014. Acoustic measures of falsetto voice. Paper presented at the annual meeting of the Acoustical Society of America. Providence, RI.
[8] Maekawa, K., Kikuchi, H., Tsukahara, W. 2004. Corpus of Spontaneous Japanese: Design, annotation and XML representation. *Proc. International Symposium on Large–scale Knowledge Resources*, Tokyo, 19–24.
[9] Maekawa, K., Koiso, H., Furui, S., Isahara, H. 2000. Spontaneous speech corpus of Japanese. *Proc. 2nd Int. Conf. Lang. Res. Eval*, Athens, 947–952.
[10] Patel, S., Scherer, K. R., Björkner, E., Sundberg, J. 2011. Mapping emotions into acoustic space: The role of voice production. *Biol. Psychol.* 87, 93–98.
[11] Scherer, K. R., Sundberg, J., Fantini, B., Trznadel, S., Eyben, F. 2017. The expression of emotion in the singing voice: Acoustic patterns in vocal performance. *J. Acoust. Soc. Am*. 142, 1805–1815.
[12] Skov, L., Moeran, B. 1995. *Women, media, and consumption in Japan*. Honolulu: University of Hawai'i Press.
[13] Starr, R. L. 2015. Sweet voice: The role of voice quality in a Japanese female style. *Language in Society* 44, 1–34.
[14] Sundberg, J., Patel, S., Bjorkner, E., Scherer, K. R. 2011. Interdependencies among voice source parameters in emotional speech. *IEEE Trans. Affect. Comput.* 2, 162–174.
[15] Teshigawara, M. 2003. Voices in Japanese animation. Doctoral dissertation, University of Victoria.
[16] Wang, H., Utsugi, A. 2018. Nihongo kanjo-onsei no sanshutsu ni okeru koeshitsu no tokucho: GeMAPS o mochiita bunseki. *Proc. 2018 Spring Meeting of the Acoustical Society of Japan*, Miyashiro, 1261–1264.