VOICE-QUALITY ANALYSIS OF JAPANESE FILLED PAUSES: A PRELIMINARY REPORT

Kikuo Maekawa[†] and Hiroki Mori [††]

[†] Dept. Corpus Studies, National Institute for Japanese Language and Linguistics
[††]Graduate School of Engineering, Utsunomiya University

ABSTRACT

Using the Core of the *Corpus of Spontaneous Japanese*, acoustic analysis of F1, spectral tilt (TL), H1-H2, jitter and F0 was conducted to examine the voice-quality difference between the vowels in filled pauses and those in ordinary lexical items. It turned out by simple SVM analysis that the two classes of vowels could be discriminated with the mean accuracy of higher than 70%.

1. INTRODUCTION

Recently, analysis on the cognitive functions of filled pauses (FP hereafter) has made remarkable progress [1-3]. Speech production mechanisms of FP, on the other hand, remained largely unclarified even now. In this paper, results of preliminary voice-quality analyses of Japanese FP will be presented.

In Japanese, voice-quality of FP seems to be controlled by the speakers to transmit various “paralinguistic” information [4,5]. For example, FP /eH/ (prolonged /e/, the most frequent FP in Japanese) is often produced with noticeable creaky phonation, and the listeners tend to perceive the FP as transmitting strong ‘hesitation’ of the speaker. Similarly, /eH/ with breathy phonation tends to be perceived as implying the speaker’s attitudes like ‘politeness’, ‘weariness’, or ‘lack of self-confidence’.

Needless to say, similar control of voice-quality can be observed in ordinary lexical items (LX hereafter) like nouns and verbs [6]. But it seems that the control of voice-quality is more salient (and frequent) in FP than in LX. It is presumably because the number of possible word-form is considerably limited in FP than in LX.

The aim of the present study is to show that there is systematic acoustic difference related to voice-quality between the vowels in FP and LX in Standard (Tokyo) Japanese. In this study, the word “voice-quality” is used to refer to both laryngeal and supra-laryngeal features of speech. The phonation characteristics mentioned above belong to laryngeal voice-quality, while the variation in vowel articulation (measured in terms of vowel formant frequencies) belongs to supra-laryngeal, or segmental, voice-quality. Both of them are analysed in this paper.

2. DATA

Monologue talks in the CSJ-Core, i.e., the X-JToBI annotated part of the *Corpus of Spontaneous Japanese* was analysed [7]. 79 male and 58 female speakers were involved in this data. Among more than 30,000 FP recorded in the CSJ-Core, vocalic FP (namely, those FP consisting exclusively of monophthong vowels, /iH/, /eH/, /aH/, /oH/, and /uH/.) were analysed and compared to the corresponding LX long vowels. Only the LX vowels located in the word-initial positions like /eHgo/ “English”, and /koHgi/ “lecture”) were analyzed. Vowels that were estimated to have less than eight pitch cycles were omitted from the analysis, because jitter analysis (see below) required the duration of at least 5 pitch cycles.

Table 1 shows the number of samples analysed in the current study. Omitting the cases where numbers of samples were too small, male’s /aH/, /eH/, and /oH/, and female’s /aH/ and /eH/ were analysed.

Table 1: Numbers of analysed samples *

Speaker sex	Vowel	Filled Pause (FP)	Lexical Item (LX)
Male	/aH/	108	113
	/eH/	2411	764
	/iH/	16	281
	/oH/	66	2177
	/uH/	16	561
Female	/aH/	40	61
	/eH/	1049	529
	/iH/	2	248
	/oH/	10	1910
	/uH/	12	500

*Samples in the shaded cells are omitted.

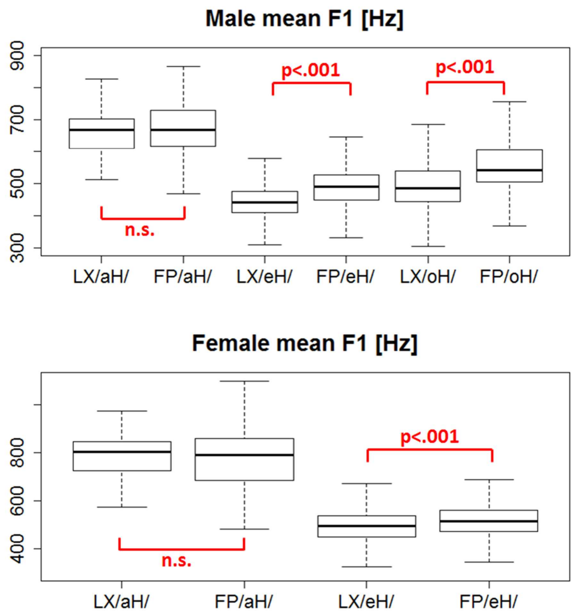
3. ANALYSIS

3.1. Formant frequency

Formant analysis was conducted using Praat [8]. Mean F1, F2, and F3 of vowels were computed by LPC method (number of poles was set to 12). An interesting result was obtained with respect to the first formant frequency (F1). F1 was significantly higher in FP than in LX in the male samples of /eH/

and /oH/ (See Fig.1). Female vowels showed significant difference in /eH/, but not in /aH/.

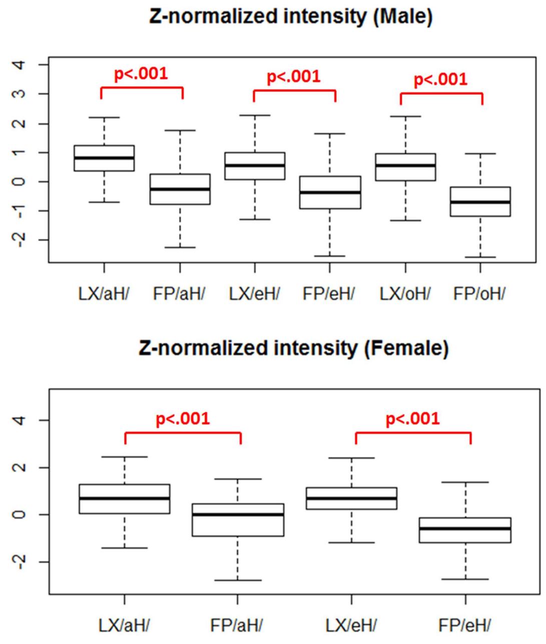
Figure 1: Mean F1[Hz]] in LX and FP.



3.2. Intensity

Intensity information of vowels was also analysed by Praat. Mean vowel intensities were z-normalized based upon the distribution of all vowel samples encompassing both FP and LX for each subject. Mean intensity was significantly smaller in FP than in LX in all vowels across male and female samples (Fig.2).

Figure 2: Mean z-transformed intensity in LX and FP.



3.3. Spectral tilt (TL)

Spectral tilt (TL) is a measure of phonation types. In order to estimate the TL of vowels in an efficient way, cepstrum-based method was used for analysis. Overall trend of a spectrum was approximated by the first cepstrum component, and the difference between the estimated amplitudes at 0 and 3000 Hz was used as the estimated TL [9]. Fig. 3 shows an example of the analysis where blue curve is the FFT spectrum, green curve represents the first cepstrum coefficient, and the red line stands for the TL. Note the TL estimated by this method is not the tilt of laryngeal source. Rather, it is the tilt of speech sound radiated from the mouth of speaker.

In male samples, TL was significantly larger in FP than in LX in all vowels (Fig. 4), while in female samples significant difference was not observed in any vowel.

Figure 3: Ceptstrum-based estimation of TL.

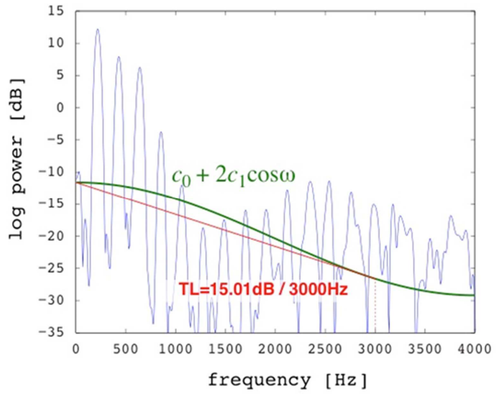
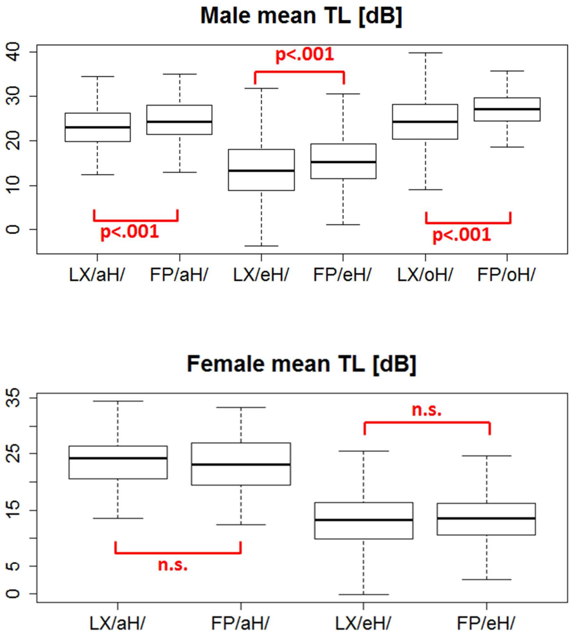


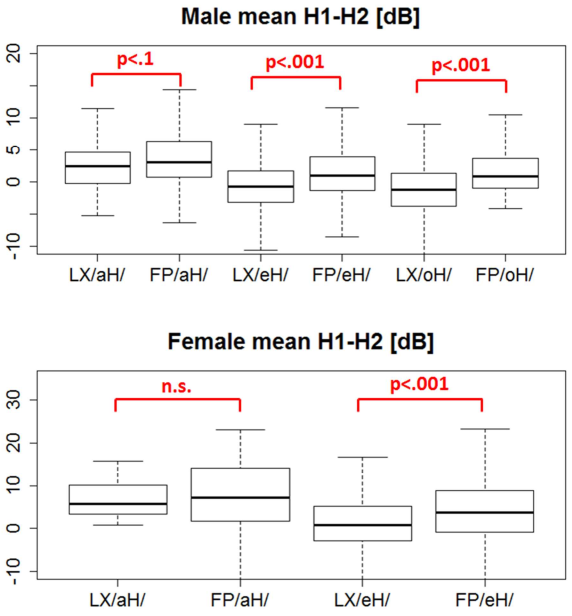
Figure 4: Mean TL [dB] in LX and FP



3.4. H1-H2

H1-H2 (the difference between the levels of the first and second harmonics) is another measure of phonation types, which is popular among phoneticians [10]. An open source script of Praat developed by Chad Vicenik was used for the computation [11]. Fig. 5 shows the mean H1-H2 values for male and female speech. In male speech, mean H1-H2 was significantly higher in FP than in LX (significance is marginal in the case of /aH/). In female speech, mean H1-H2 was significantly higher in /eH/, but not in /aH/.

Figure 5: Mean H1-H2 [dB] in LX and FP.



3.5. Jitter

Jitter is an index of the fluctuation of speech fundamental frequency (F0). In speech pathology, jitter is used frequently for the evaluation of pathological voices, and measured on the basis of the analysis of long sustained vowels. In this study, however, jitter analysis was applied for vowels in running spontaneous speech. Among various definitions of jitter, PPQ5 was computed using the voice report function of Praat. Mean jitter in logarithmic scale was significantly higher in FP than in LX in all vowels of male and female samples (Fig. 6). See the discussion below for the problem of this analysis.

3.6. F0

F0 contours of FPs is believed to be simpler compared to that of LX [12]. Moreover, the F0 values of FPs are supposed to be predictable, to some extent, from the surrounding phonological (tonal) context [13,14]. Here, mean F0 of FP vowels is compared to that of LX. As shown in Fig. 7, mean F0 was significantly lower in FP than in LX in all vowels across male and female samples.

Figure 6: Mean jitter in log scale in LX and FP.

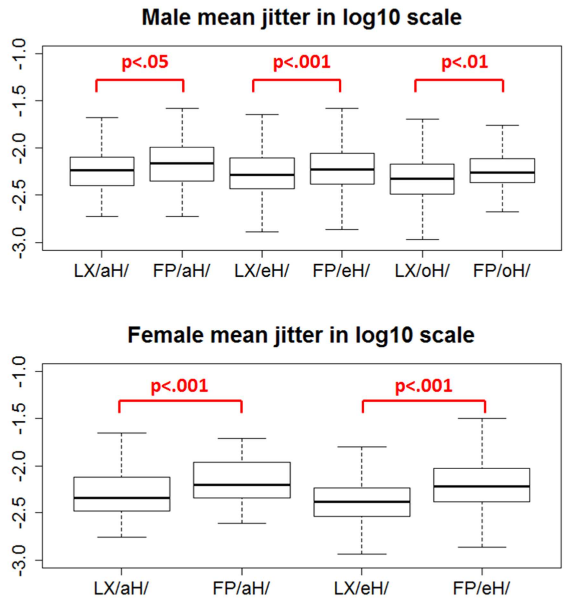
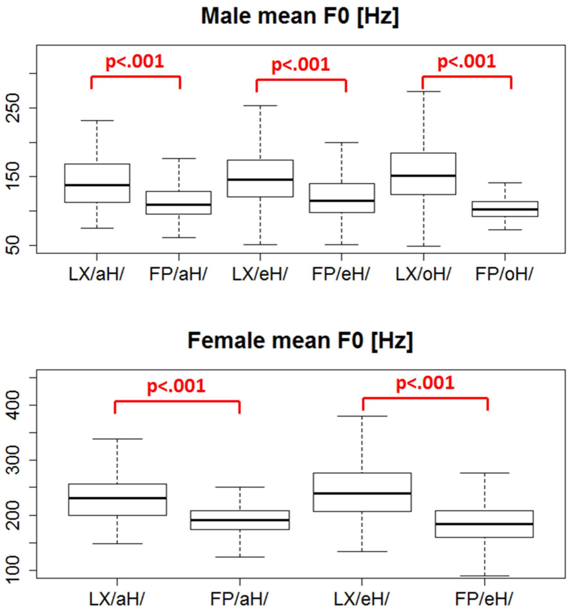


Figure 7: Mean F0 [Hz] in LX and FP.



3.7. Automatic classification

Acoustic analyses presented above suggested strongly the presence of voice-quality difference between the FP and LX. Automatic classification by means of SVM was carried out using the e1071 package of R language (version 3.1.3) in order to know how systematic the differences were.

Table 2 shows the results of ten-fold cross-validation. Note that SVM parameters were fixed to the same values (cost=3.0, gamma=0.5, and kernel is radial) for all vowels. Note also that the samples of FP and LX vowels were resampled from the data set of Table 1 so that each class has the same number of samples. Therefore, the baseline accuracy was 50% in all cases. The number of samples thus resampled is shown in the third column ('N') of the table.

Mean accuracy of classification (%) was computed under four combinations of explanatory variables: 'All' stands for the case where all six acoustic features were used as explanatory variables. '-P' and '-J' stand respectively for the cases where pitch and jitter were removed, and '-PJ' mean the removal of both of them. The success rates are considerably higher than the baseline in all cases.

Table 2: Accuracy (%) of classification by SVM.

Sex	Vowel	N	All	-P	-J	-PJ
M	/aH/	200	72.0	65.0	70.5	60.5
	/eH/	400	76.0	70.3	76.3	69.5
	/oH/	100	71.0	61.0	74.0	70.0
F	/aH/	60	73.3	63.3	78.3	73.3
	/eH/	400	80.0	64.3	78.5	64.5

4. DISCUSSIONS

Systematic difference was observed between the voice-quality of vowels in FP and LX in spontaneous speech. As a general tendency, vowels in FP had relatively "softer" (as indicated by larger TL and H1-H2 values) and unstable (as indicated by larger jitter values) phonation. They were also marked by lower intensity, lower F0, and higher F1. Moreover, SVM analysis using these features as explanatory variable revealed that it was possible to make distinction between the FP and LX vowels with the mean accuracy of higher than 70%.

There are, however, two important cautions about the current analysis. First, computation of jitter values may be influenced seriously by F0 tracking error, which is inevitable in the current data that contains many samples with heavy creaky / breathy phonation. In this respect, it is noteworthy that in Table 2 the highest mean accuracy was often achieved when the jitter data was removed from the explanatory variables (i.e., the '-J' condition).

Second, the difference of F1 is not easy to interpret. There are at least two interpretations that seem to be equally plausible. For one, it can be the consequence of incomplete glottal closure of 'soft' phonation that makes vocal tract open at both ends [15]. For another, it can be the consequence of speakers' deliberate control over vowel articulation. Prior studies reported the case where speakers control their segmental articulation (in addition to the control of prosody) to transmit certain paralinguistic information [5,16].

As suggested in the introduction, it is the present authors' belief that the observed acoustic difference is related, in some way, to the transmission of paralinguistic information. But this remains as a mere conjecture at the current stage of inquiry. In the current analysis, samples having different paralinguistic information were analysed altogether. Because of this, it is highly probable that the observed acoustic characteristics of FP were biased strongly by the properties of the samples whose paralinguistic information have high occurrence frequencies. For example, higher values of TL and H1-H2 observed in FP samples could be the result of frequent use of FP to transmit 'politeness' and/or 'lack of self-confidence'.

Manual annotation of perceived phonation types and intended paralinguistic message is currently underway in view of automatic annotation. The classification data will help us to understand the production mechanism of FP more precisely.

5. REFERENCES

[1] Sadanobu, T. & Y. Takubo. "Danwa ni okeru monitā kinō." *Gengo Kenkyū*, 108, 74-93, 1995.

[2] Clark, H. & J. Fox Tree. "Using *uh* and *um* in spontaneous speaking." *Cognition*, 84, 73-111, 2002.

[3] Watanabe, M. *Features and roles of filled pauses in speech communication*. Tokyo: Hituzi, 2009.

[4] Ladd, D. R. *Simultaneous structure in phonology*. Oxford: Oxford Univ. Press, 2014.

[5] Mori, H, K. Maekawa, & H. Kasuya "*Onsei wa nani o tsutaete iru ka*". Tokyo: Korona-sha, 2014.

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

[7] Maekawa, K. "Corpus of Spontaneous Japanese: Its Design and Evaluation". *Proc. SSPR2003*, 7-12, 2003.

[8] http://www.fon.hum.uva.nl/praat/

[9] Maekawa, K. & H. Mori. "Filā no sēshitsujō no tokuchō ni kansuru yobiteki bunseki." *Proc. Spring Meeting of Acoust. Soc. Japan*, 3-2-9, 2015.

[10] Gordon, M. & P. Ladefoged. "Phonation types: a cross-linguistic overview." *J. Phonetics*, 29, 383-406, 2001.

[11]http://www.linguistics.ucla.edu/faciliti/facilities/acoustic/PraatVoiceSauceImitator.txt

[12] Shriberg, E. *Preliminaries to a theory of disfluencies*. Ph.D. Diss. Univ. California at Barkeley, 1994.

[13] Shriberg, E. & R. Lickley. "Intonation of clause-internal filled pauses". *Phonetica*, 50, 172-179, 1993.

[14] Maekawa, K. "Prediction of F0 height of filled pauses in spontaneous Japanese". *Proceedings of DiSS 2013*, Stockholm, 41-44, 2013.

[15] Honda, K. et al. "Visualisation of hypopharyngeal cavities and vocal-tract acoustic modelling." *Computer Methods in Biomechanics and Biomedical Engineering*. 13(4), 443–453, 2010.

[16] Maekawa. K. "Production and perception of 'paralinguistic' information." *Proc. Speech Prosody 2004*, Nara, 367-374 , 2004.