INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Voice-Quality Difference between the Vowels in Filled Pauses and Ordinary Lexical Items

Kikuo Maekawa [1], *Hiroki Mori*[2]

[1] National Institute for Japanese Language
[2] Utsunomiya University

kikuo@ninjal.ac.jp, hiroki@speech-lab.org

Abstract

Acoustic differences between the vowels in filled pauses and ordinary lexical items such as nouns and verbs were examined to know if there was systematic difference of voice-quality. Statistical test of material taken from the Corpus of Spontaneous Japanese showed that, in most cases, there was significant difference of acoustic features like F0, F1, F2, intensity, jitter, shimmer, TL, H1-H2, H1-A2, duration, etc. between the two classes of vowels. Random forest classification of open data sets showed higher than 0.8 F-values on average. It turned out intensity, F0, F1, jitter, and H1-H2 were the most important acoustic features for the expected voice-quality difference.

Index Terms: filled pauses, voice-quality, random forest, Corpus of Spontaneous Japanese

1. Introduction

The aim of this paper is to show that there is systematic difference of voice quality between the vowels in filled pauses and ordinary lexical items in spontaneous Japanese.

Recently, there is growing consensus among researchers about the important cognitive roles played by filled pauses (FP hereafter) in speech communication. FPs such as English *uh* and *um*, and Japanese *eH* (prolonged /e/) and *anoH*, are used to transmit various pragmatic information like hesitation, floor holding, indication of ongoing message planning and lexical search, and so forth [1-4].

On the other hand, the production mechanism of FP is left largely unclarified except for a couple of preliminary studies on the F0 shape of FPs in English and Japanese [5, 6].

The absence of interest in the production aspect of FPs is not surprising in a sense, because it is often the case that the morphological / phonetic properties of FPs are quite restricted compared to ordinary lexical items like nouns and verbs. The inventory of FPs is quite small; less than ten in most European languages, and less than twenty in a language like Japanese whose FP inventory is known to be quite rich.

Moreover, there is often a restriction about the possible phonetic shapes of FPs. In Japanese, for example, FPs are either vowels (i.e., /iH/, /eH/, /aH/, /oH/, /uH/, and their short vowel counterparts), or cognates of demonstratives like /ano/ and /sono/ (both correspond to "that" in English).

As for prosody, preceding studies on English and Japanese showed that the F0 shape of FPs was largely underspecified; it can be predicted to a large extent from the intonational environment in which they occur [5, 6].

The presence of strict morphological restriction as well as prosodic underspecification give rise to one important question about the production of FPs: why can FPs transmit various pragmatic messages under such strong restrictions? A plausible answer to this question is the contribution of voice quality features like phonation types. Recently, the present authors tried to show by means of acoustic analysis that there was systematic difference between the voice quality of FPs and ordinary lexical items (LX hereafter) [7]. In the study, FPs and LXs were compared with respect to various acoustic features (see below). The results revealed significant difference between the FPs and LXs in most features.

In the present study, the same data as in the previous study is reanalyzed with respect to more acoustic features. Moreover, a machine learning technique is used to evaluate relative importance of acoustic features for the discrimination of FPs and LXs. Lastly, the performance of classifier is evaluated by means of cross-validation.

2. Data

Monologue talks in the X-JToBI [8] annotated part of the Corpus of Spontaneous Japanese [9], known as the CSJ-Core, was analyzed. The data was spoken by 79 male and 58 female speakers. Among the more than 30,000 FPs involved in the data, vocalic FP /eH/ and /aH/ (namely, those FPs consisting exclusively of monophthong vowels) were analyzed and compared to the corresponding LX long vowels.

As for the LX vowels, only the vowels located in the word-initial position like /eHga/ "movie" and /raHmeN/ "ramen noodle" were chosen for analysis. Vowels that were estimated to have less than ten pitch cycles were omitted from the analysis, because jitter and shimmer analyses required the duration longer than five pitch cycles. Table 1 shows the number of samples analyzed in the current study. Vowels whose numbers of samples were less than 50 were omitted from the analysis.

Table 1. Number of samples

Speaker	Vowel	FP	LX
Male	/aH/	108	113
	/eH/	2411	764
Female	/aH/	40	61
	/eH/	1049	529

3. Analysis

3.1. Acoustic analysis

Acoustic features analyzed in this study included duration, mean-intensity, mean F0, mean first formant frequency (F1), mean second formant frequency (F2), mean jitter, mean

http://dx.doi.org/10.21437/Interspeech.2016-1309

shimmer, mean harmonic to noise ratio (Harm2noise), mean spectral tilt (TL), mean difference of the first two harmonics H1-H2), and, mean difference between the first harmonic and the level of second formant (H1-A2).

In the following subsection, some selected results will be presented in detail. Due to the limitation of space, the whole results will be summarized later in Table 2.

3.1.1. Duration

As shown in Figure 1, FP has longer vowel duration than LX, both in male and female samples.

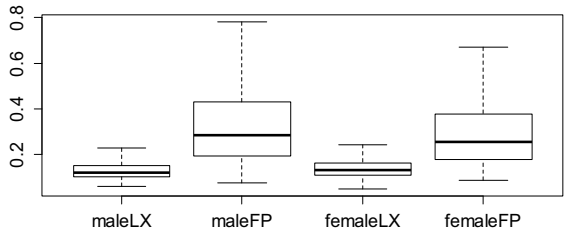

Figure 1. Mean duration [sec] of /eH/

3.1.2. Intensity

Intensity information was z-normalized using the mean and SD of each speaker. As shown in Figure 2, FP has weaker intensity values compared to LX.

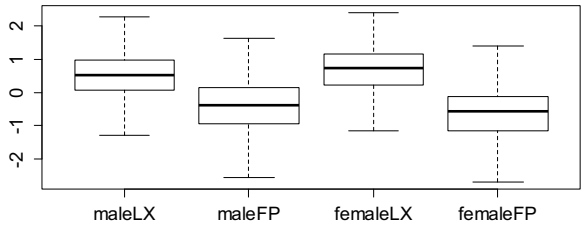

Figure 2. Mean intensity (z-transformed) of /eH/

3.1.3. F0

F0 was computed for every 10ms using autocorrelation method and averaged over the whole duration of a vowel using Praat (Ver. 5.3.76) [10]. Log mean F0 values were z-transformed for each speaker based upon the all F0 values obtained for the speaker. Logarithm (base 10) of F0 values is z-transformed for each speaker. As shown in Figure 3, FP has lower mean F0 than LX in male speech. On the other hand, female speech tends to have slightly higher mean F0 in FP than LX.

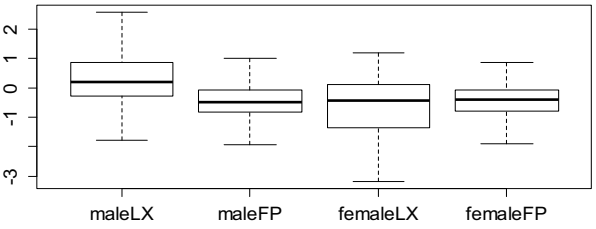

Figure 3. Mean F0 (z-transformed) of /eH/

3.1.4. F1 and F2

Formant frequencies were estimated by LPC method (number of poles was set to 12), and then z-transformed for each speaker. Like F0, logarithm of F1 value is z-transformed for each speaker.

As shown in Figure 4, FP has higher mean F1 value than LX in the case of /eH/ in both male and female speech.

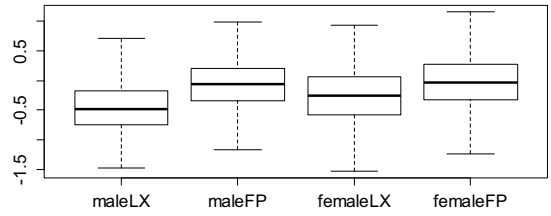

Figure 4. Mean F1 (z-transformed) of /eH/

3.1.5. Jitter and shimmer

Among various definitions of jitter, PPQ5 (five-point period perturbation quotient) [11] was computed using the voice report function of Praat. As shown in Figure 5, FP has higher mean jitter than in LX.

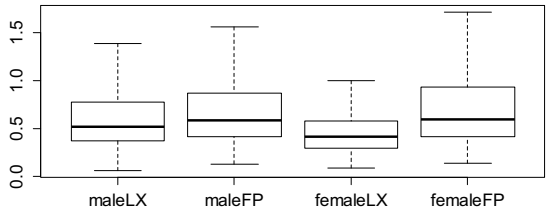

Figure 5. Mean jitter [%] of /eH/

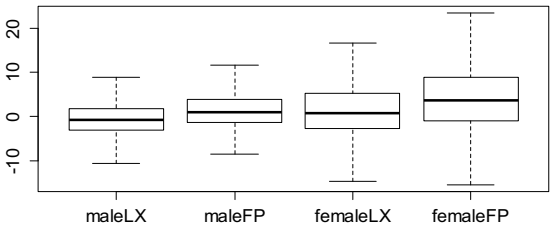

Figure 6. Mean H1-H2[dB] of /eH/

3.1.6. Spectral tilt, H1-H2, H1-A2 and harmonic to noise ratio

TL, or spectral tilt, was estimated by a cepstrum-based method described in [7, 12]. 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.

H1-H2 (the difference between the levels of the first and second harmonics) and H1-A2 (the difference between the levels of the first harmonic and second formant) are also the measures of spectral tilt. They are often utilized in the literature of phonetics [13]. An open source script of Praat developed by Chad Vicenik [14] was used for the computation with minimum necessary modifications. As shown in Figure 6,

FP has tendency to have higher spectral tilt (H1-H2, in this case) than LX.

Lastly, harmonics to noise ratio is an acoustic measure that can be related to voice hoarseness [15]. It was computed using the voice report function of Praat.

3.1.7. Summary of acoustic analyses

Table 2 summarizes the results of t-tests applied for all eleven acoustic features. Significance levels are shown by stars, and the observed magnitude relationships between the FP and LX samples are indicated by the '>' and '<' marks.

All features showed statistical significance of some sort in at least two columns. Observed magnitude relationships were stable in most cases, but there were cases where the relationship turned out to be unstable (F0, F2, and Harm2noise).

Table 2. Results of acoustic analysis

Variables	Male /eH/	Male /aH/	Female /eH/	Female /aH/
Duration	**** LX < FP	**** LX < FP	**** LX < FP	**** LX < FP
Intensity	**** LX > FP	**** LX > FP	**** LX > FP	**** LX > FP
F0	**** LX > FP	**** LX > FP	**** LX < FP	NS
F1	**** LX < FP	**** LX < FP	**** LX < FP	**** LX < FP
F2	**** LX > FP	**** LX < FP	**** LX < FP	NS
Jitter	**** LX < FP	* LX < FP	**** LX < FP	~ LX < FP
Shimmer	* LX < FP	NS	**** LX < FP	~ LX < FP
Harm2noise	**** LX < FP	NS	**** LX > FP	NS
TL	**** LX < FP	* LX < FP	~ LX < FP	NS
H1-H2	**** LX < FP	~ LX < FP	**** LX < FP	NS
H1-A2	** LX < FP	**** LX < FP	*** LX < FP	**** LX < FP

**** p<.0001, *** p<.001, ** p<.01, * p<.05, ~ p<.1, NS Not significant

3.2. Classification by random forest

Table 2 provides little information about the relative importance of acoustic features for the sake of the classification of FP and LX vowels. It is impossible to know which one of intensity and H1-A2 is more important for the classification, for example.

A machine-learning technique known as random forest was utilized to evaluate the relative importance of acoustic features [16]. Four vowel classes, i.e. male /eH/, male /aH/, female /eH/, and female /aH/, were analyzed separately in this section.

In each vowel class, a classification tree for the FP-LX distinction was constructed using the eleven acoustic features listed in Table 2 as the predictor variables.

In order to set the baseline of classification to 0.5 precisely, the same number of samples were randomly extracted from the data of Table 1. As for male /eH/, male /aH/, and female /eH/, 100 vowels were extracted for each of FP and LX categories. As for female /aH/, 30 vowels were extracted for FP and LX categories.

The randomForest package (4.6-12) of the R language (3.1.3) was used for the analysis. Parameter tuning was done by the tuneRF command of the package.

Table 3 summarizes the results. The first eleven rows of the table show the relative importance of acoustic features. The number in each column is called mean decrease in Gini index (MDG) and shows the relative importance of the feature for the classification of the FP and LX vowels. The larger the MDG is, the greater the contribution of the feature for the classification is. Shaded cells stand for the top-five acoustic features in each vowel class.

The last 4 rows of the table, on the other hand, show the accuracy, precision, recall, and F-measure (the harmonic mean of precision and recall) of the classification. Precision and recall are those of FP rather than LX, here, and in the rest of this paper.

It was the features of duration and intensity that contributed the most to the classification, and, the features like F0, F1, Jitter, and H1-A2 made secondary contributions depending on the vowel class. Judging from the F-values, the performance of the classification was fairly good.

There is, however, an important phonetic question: duration is not usually regarded to be a constituent of voice-quality features of vowels in traditional phonetic [17]. To answer this question, random forest analyses that excluded duration from the predictor variables were conducted.

As summarized in Table 4, the results were very similar to the ones reported in Table 3. The only notable exception was the considerably lowered classification performance in the case of female /aH/.

Table 3. Results of random forest (Including duration)

Feature	Male/eH/	Male/aH/	Female/eH/	Female/aH/
Duration	35.10	38.62	26.39	10.89
Intensity	24.04	16.51	23.61	3.12
F0	7.44	13.39	5.58	1.13
F1	9.49	5.97	7.40	1.69
F2	3.73	5.14	4.43	1.27
Jitter	3.66	2.67	7.00	2.15
Shimmer	3.28	3.31	4.30	1.68
Harm2noise	2.93	2.94	4.54	1.05
TL	2.78	2.91	4.23	0.91
H1-H2	2.95	3.10	5.66	2.55
H1-A2	4.14	4.96	6.35	3.08
Accuracy	0.84	0.90	0.88	0.80
Precision	0.84	0.88	0.86	0.77
Recall	0.84	0.92	0.90	0.82
F-measure	0.84	0.90	0.87	0.79

Table 4. Results of random forest (Excluding duration)

Feature	Male/eH/	Male/aH/	Female/eH/	Female/aH/
Intensity	26.95	24.83	36.87	4.16
F0	12.88	22.09	7.68	2.27
F1	15.03	10.27	10.94	3.25
F2	6.50	7.87	5.55	2.89
Jitter	6.23	4.23	8.62	2.33
Shimmer	5.89	5.93	5.22	2.48
Harm2noise	6.20	6.18	5.23	2.72
TL	6.03	5.36	5.71	1.84
H1-H2	6.59	4.76	6.54	3.81
H1-A2	7.20	8.00	7.14	3.72
Accuracy	0.79	0.81	0.85	0.60
Precision	0.84	0.87	0.85	0.53
Recall	0.76	0.78	0.84	0.62
F-measure	0.80	0.82	0.85	0.57

3.3. A unified model

So far, four vowel classes were analyzed separately using different classifier models. In this section, a unified model that covers both male and female speakers, and /aH/ and /eH/

altogether is proposed. For this purpose, data of all vowel classes were merged into one, and two new predictor variables, viz., speaker sex (male and female) and vowel category (/a/ and /e/), were added to the existing predictor variables of acoustic features. The procedure of random forest analysis was the same as in the previous analyses.

The unified model showed slightly better performance than in the separate analyses. Accuracy, precision, recall, and the F-value of the unified model were 0.89, 0.87, 0.90, and 0.88 respectively. The top-five features and the corresponding MDG values were duration (114.23), intensity (68.24), F0 (27.68), F1 (19.73), and, Jitter (16.52). The two new variables did not make any noticeable contributions, the MDG for sex and vowel category being 1.10 and 0.85 respectively.

Exclusion of duration feature from the predictor variables did not change the result considerably. Accuracy, precision, recall, and the F-value of the unified model without duration feature were 0.80, 0.82, 0.79, and 0.81 respectively; the top-five features included intensity (MDG 84.88), F0 (44.19), F1 (32.47), Jitter (24.46), and H1-H2 (24.40), whereas the contributions of sex (3.64) and vowel category (1.85) stayed nearly negligible.

3.4. Cross-validation

So far, the performance of random forest classifiers was evaluated by using so-called closed data, i.e. the case where the same data set was utilized both for training and evaluation of a classifier. As is well known, classifiers constructed in this way tend to run the risk of overlearning or overfitting to the training data.

To avoid the risk, ten-fold cross-validation was conducted using the data set that we used in 3.3 (i.e. the data for unified model). Cross-validation was conducted in the following manner. First, 90% of the new data set (i.e. 594 samples) was extracted randomly as a training data, and the remaining 10% (i.e. 66 samples) was used as a test data (or open data). This sampling process was repeated ten times so that there were ten independent pairs of training and test data sets. With respect to these test data sets, the chance level of FP classification differed from set to set, but on average the chance level is about 0.50.

Second, a random forest classifier, which was a unified model, covering both male and female samples and both /aH/ and /eH/ samples, was constructed based upon a training data. Then the classifier was applied for the task of FP/LX classification of corresponding test set. The performance of classifier was evaluated by means of accuracy, precision, recall, and the f-measure as in previous analyses. This process was repeated ten times for all pairs of training and test data sets.

Moreover, the whole process described above was repeated twice. In the first round, data including the feature of duration was utilized, and in the second turn, data excluding the duration feature was utilized.

Variation of accuracy, precision, recall, and f-measure are shown in Figures 7 (including the duration feature) and 8 (excluding the feature).

In Figure 7, performance of classifies was as high as 0.89 on average in terms of the f-measure. Exclusion of duration feature lowered the performance to some extent, but the performance shown in Figure 8 remained still fairly high; the f-measures were 0.81 on averages.

As for the relative importance of acoustic features, the top five features included duration (118.5), intensity (56.0), F0 (23.5), F1 (15.7), and jitter (15.1) in the case of data including duration (The parenthesized digits are mean MDG).

In the case of data excluding duration, the top five features included intensity (85.3), F0 (36.9), F1 (30.4), jitter (23.7), and H1-H2 (21.0).

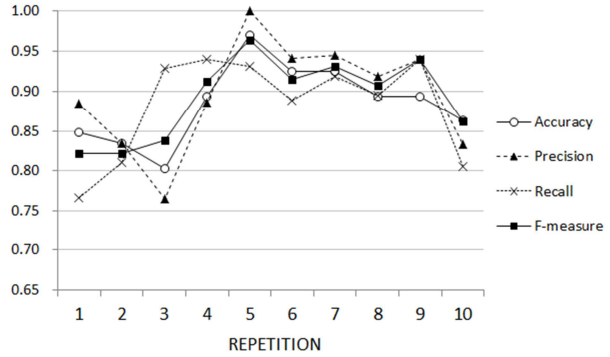

Figure 7. Result of cross-validation (Including duration)

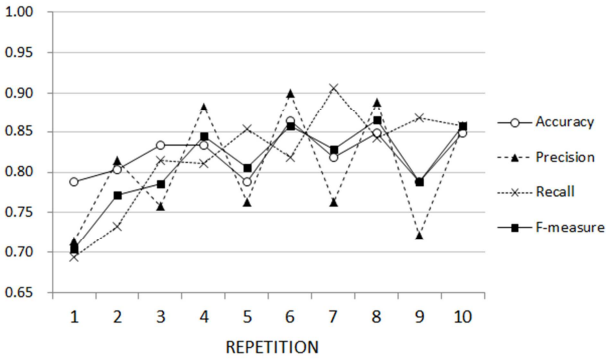

Figure 8. Result of cross-validation (Excluding duration)

4. Conclusion

Classification of FP and LX vowels by means of random forest technique unveiled relative contribution of various acoustic features to the classification task, which was left unrevealed in the previous studies [7, 12].

Cross-validation of random forest classifiers revealed high performance in the classification of open data sets. Although the feature of duration contributed greatly in all models, analysis of data excluding the duration feature also showed high classification performance.

This fact suggests strongly the conclusion that there is systematic difference in voice-quality between the vowels in FP and LX. Acoustic features like intensity, F0, F1, jitter, and H1-H2 seem to be the most relevant for the difference. Compared to LX vowels FP vowels seem to be characterized by lower intensity and F0, higher F1, larger jitter, and, larger H1-H2 as shown in Figures 2-6.

Further clarification of the phonetic details of voice-quality difference should be the theme of future investigation.

5. Acknowledgements

This work is supported by the JSPS *Kakenhi* grant to the first author (No.26284062).

6. References

[1] Swerts, M. "Filled pauses as markers of discourse structure". *J. Pragmatics*, 30, 485-496,1998.

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

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

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

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

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

[7] Maekawa, K. & Mori, H. "Voice-quality analysis of Japanese filled pauses: A preliminary report". *Proc. Diss 2015*, Edinburgh, 2015.

[8] Maekawa, K. Kikuchi, H., Igarashi, Y., & Venditti, J. "X-JToBI: An extended J_ToBI for spontaneous speech". *Proc. ICSLP2002*, Denver, 1545-1548, 2002.

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

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

[11] Gelzinis, A., Verikas, A. & Bacauskiene, M. "Automated speech analysis applied to laryngeal disease categoraization". *Computer Method and Programs in Biometicine*, 91, 36-47, 2008.

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

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

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

[15] Yumoto, E., Gould, W. & Baer, T. "Harmonics-to-noise ratio as an index of the degree of hoarseness." *JASA*, 71 (6), 1544-1550, 1982.

[16] http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#overview

[17] Laver, J. *Principles of Phonetics*. Cambridge Univ. Press, 1994.