# Phonetic characteristics of filled pauses: a preliminary comparison between Japanese and Chinese

*Kikuo Maekawa [1], Ken'ya Nishikawa [1] and Shu-Chuan Tseng [2]*
*[1]National Institute for Japanese Language and Linguistics, Tokyo, Japan*
*[2]Institute of Linguistics, Academia Sinica, Taipei, Taiwan*

## Abstract

Filled pauses in spontaneous Chinese and Japanese were analyzed to examine if there is systematic phonetic difference between the vowels of filled pauses and those occurred in ordinary lexical items. Also, the effect of the category of filled pauses (simple vocalic fillers versus fillers derived from demonstratives) was examined in both languages. Random forests analysis revealed that it was possible to construct automatic classifiers that achieved F-measure values of .7-.9. It turned out also that, in both languages, vowels in simple vocalic filled pauses showed higher F-values than the filled pauses derived from demonstratives. Lastly, it turned out that acoustic features distinguishing filled pauses from ordinary lexical items differ depending on both the category of filled pauses and languages.

## Aim of the study

Filled pauses (FP hereafter) transmit various pragmatic/para-linguistic information, but at the same time, there are strong phonological constraints in the form of FP. It is accordingly expected that sub-phonemic phonetic details and voice quality play important roles in the information transmission by FP. It was reported recently for Japanese that there is systematic difference of voice quality between the vowels in FP and those in ordinary lexical items like nouns and verbs (LX hereafter), and it was possible to construct automatic classifiers of significant performance for the task of FP-LX vowel classification (Maekawa & Mori, 2015, 2016). The aim of the current paper is two-folds. First, to examine if similar conclusion could be obtained in language other than the Japanese, and second, to examine if the conclusion differs depending on the category of filled pauses.

## Data

Two corpora of spontaneous speech were used for analysis. The core part of the Corpus of Spontaneous Japanese (CSJ-Core, 505455 words and 984092 morae spoken by 139 speakers; see Maekawa (2003) for details) and Mandarin Conversational Dialogue Corpus (MCDC8, 93533 words and 136229 syllables spoken by 16 speakers, see Tseng (2014) for details). An important difference between these corpora is that CSJ is a corpus of monologue, while MCDC8 is a corpus of dialogue. See conclusion on this issue.

In both Japanese and Chinese, the simplest and typical FP consists of a single vowel. In Japanese, all five vowels are used as FP but with considerably different frequencies. In Chinese, schwa vowel is used exclusively for simple fillers. Moreover, in both Chinese and Japanese, there are FPs that are morphologically derived from demonstratives. In Japanese, /ano/ ('that') and /sono/ ('this') are such FP. In Chinese /nà nà ge/ ('that', 'that+classifier') and /zhè zhège/ ('this', 'this+classifier') are used as FP (Zao & Jurafsky 2005). In the rest of this paper, these two FP categories will be referred to as simple-fillers (SF) and demo-fillers respectively.

In the Chinese corpus, there were 104 simple-fillers vowel (/ə/), 244 demo-filler vowels of /ə/, and 475 demo-filler vowels of /a/. Note only monophthong (excluding diphthongs and filler-like particles) were analysed in this study for the sake of comparison with the Japanese vowel.

Japanese corpus had 3450 simple-filler vowels of /eH/ (long /e/ vowel), 136 simple-filler vowels of /aH/, 1519 demo-filler vowels of /a/ (derived from /ano/) and 252 demo-filler vowels of /o/ (derived from the first syllable of /sono/).

## Acoustic analysis

Acoustic features listed in Table 1 were computed for each vowel using Praat (Boersma & Weenink, 2013). Some features were Z-transformed using the mean and SD of each speaker. H1, H2 and A3 values were corrected by the method of Hanson (1997). Note only those vowels having at least ten cycles were selected for analysis. Vowels having one or more missing values in acoustic features were also excluded.

Tables 2 and 3 summarizes the results of a series of t-tests applied to the acoustic features of Table 1, the null hypothesis being no difference of means between the LX and FP vowels.

Note the test was applied to a data set consisting of 100 samples of LX vowels and 100 samples of FP vowels chosen randomly from the corpus. These data sets were prepared for random forests analysis presented in the next section.

*Table 1: List of acoustic features.*

| FEATURE | GLOSS |
|---|---|
| TL | Spectral tile [dB] estimated from cepstrum |
| Jitter | Mean Jitter (PPQ5) [%] |
| Shimmer | Mean shimmer (APQ5) [%] |
| AutoCorr | Mean autocorrelation |
| Harm2noise | Mean harmonics to noise ratio |
| FOLZ | Mean Z-value of log10 F0 in Hz |
| F1LZ | Mean Z-value of log10 F1 in Hz |
| F2LZ | Mean Z-value of log10 F2 in Hz |
| F3LZ | Mean Z-value of log10 F3 in Hz |
| IntensityZ | Mean Z-value of intensity in dB |
| Duration | Duration [sec] |
| H1*-H2* | Difference of the first and second harmonics [dB] |
| H1*-A1 | Difference of the fisrt harmonics and level of F1 |
| H1*-A2 | Difference of the fisrt harmonics and level of F2 |
| H1*-A3* | Difference of the fisrt harmonics and level of F3 |

*Table 2:* T*-test of acoustic features, Japanese vowels.*

| FEATURE | SF/aH/ | SF/eH/ | Demo/a/ | Demo/o/ |
|---|---|---|---|---|
| TL | n.s. | - LX<FP | * LX<FP | ** LX<FP |
| Jitter | * LX<FP | n.s. | n.s. | n.s. |
| Shimmer | n.s. | n.s. | n.s. | n.s. |
| AutoCorr | - LX>FP | n.s. | *** LX<FP | * LX>FP |
| Harm2noise | n.s. | n.s. | *** LX<FP | * LX>FP |
| FOLZ | *** LX>FP | n.s. | n.s. | *** LX>FP |
| F1LZ | * LX<FP | ** LX<FP | n.s. | n.s. |
| F2LZ | ** LX<FP | n.s. | n.s. | n.s. |
| F3LZ | n.s. | n.s. | n.s. | n.s. |
| IntensityLZ | *** LX>FP | *** LX>FP | *** LX>FP | *** LX>FP |
| Duration | *** LX<FP | *** LX<FP | *** LX>FP | * LX>FP |
| H1*-H2* | n.s. | * LX<FP | - LX<FP | * LX<FP |
| H1*-A1 | n.s. | n.s. | *** LX<FP | n.s. |
| H1*-A2 | ** LX<FP | n.s. | ** LX<FP | n.s. |
| H1*-A3* | n.s. | - LX<FP | *** LX<FP | ** LX<FP |

*** p<.001; ** p<0.01; * p<0.05; - p<0.1

Tables 2 and 3 show that FP and LX vowels showed significant differences in many acoustic features, but the set of significant features differ depending on the category of FP. Moreover, it is interesting that the magnitude relationship of a given significant feature can be in opposite direction depending on the FP category. For example, in Table 2, FP vowels are significantly longer than LX vowels in simple fillers, while LX vowels are longer than FP vowels in demonstrative FPs. Inverted relationships can be found in each table and across two tables. We will return to this issue in the discussion section.

*Table 3:* T*-test of acoustic features, Chinese vowels.*

| FEATURE | SF/ə/ | Demo/ə/ | Demo/a/ |
|---|---|---|---|
| TL | *** LX<FP | n.s. | n.s. |
| Jitter | n.s. | ** LX>FP | n.s. |
| Shimmer | - LX>FP | *** LX>FP | * LX>FP |
| AutoCorr | n.s. | ** LX<FP | * LX<FP |
| Harm2noise | - LX<FP | *** LX<FP | * LX<FP |
| FOLZ | n.s. | *** LX<FP | *** LX<FP |
| F1LZ | *** LX<FP | ** LX>FP | ** LX>FP |
| F2LZ | *** LX>FP | ** LX<FP | n.s. |
| F3LZ | n.s. | * LX>FP | * LX<FP |
| IntensityLZ | * LX>FP | - LX<FP | ** LX<FP |
| Duration | *** LX<FP | n.s. | n.s. |
| H1*-H2* | n.s. | ** LX<FP | ** LX<FP |
| H1*-A1 | ** LX<FP | * LX<FP | n.s. |
| H1*-A2 | n.s. | n.s. | * LX<FP |
| H1*-A3* | ** LX<FP | n.s. | n.s. |

## Random forests analysis

The results of t-tests in Tables 2 and 3 do not provide direct evaluation on the effectiveness of the features for the classification of FP and LX vowels. Random forests analysis was used to examine this issue. Random forests is a machine learning technique to construct statistical classifier like the support vector machines. A crucial difference from the support vector machines is that random forests provides information on the contribution of features used for learning. The RandomForest package (Ver. 4.6-12) of the R language (Ver. 3.3.1) was used for computation. The data sets were the same as the ones used in the previous section.

Table 4 and 5 summarize the performances of random forests analyses. Note these are the performance of cross-validation. In cross validation, data set was randomly split into two subsets; one of them contained 90% of data (i.e., 180 samples) and was used for training of classifier, and the other set containing the resulting 10% of data (20 samples) was used as a test set. The performance of the classifier was evaluated by the success rate of the classification when it is applied for the test set. This process was repeated 10 times for each class of FP and LX vowels. Numbers in Table 4 and 5 are the means of such repeated cross-validations.

The top 15 rows of Table 4 and 5 stand for the relative contributions of acoustic features. The numbers shown in each cell of the table is

called MDG (Mean Decrease in Gini). MDG shows the decrease in the value of Gini index caused by the exclusion of the predictor variable in question. The greater the MDG, the greater the contribution of the variable. In each FP category, features of top three MDG are shown by shading. Note that these are the mean values of MDGs over 10 repetitions of cross-validation.

*Table 4: Results of random forests analysis. Cross-validation of the Japanese vowels. Mead MDG values.*

| FEATURE | SF/aH/ | SF/eH/ | Demo/ano/ | Demo/sono/ |
|---|---|---|---|---|
| TL | 2.25 | 2.62 | 3.68 | 5.60 |
| jitt_ppq5 | 3.51 | 4.26 | 4.35 | 3.77 |
| shim_apq5 | 2.37 | 2.75 | 3.98 | 4.40 |
| autoCorr | 2.15 | 3.82 | 5.44 | 4.67 |
| harm2noise | 1.97 | 3.63 | 9.08 | 4.10 |
| F0LZ | 6.28 | 4.78 | 7.50 | 8.25 |
| F1LZ | 3.44 | 5.51 | 3.94 | 4.53 |
| F2LZ | 3.28 | 4.06 | 6.52 | 9.83 |
| F3LZ | 2.70 | 2.93 | 5.30 | 4.08 |
| IntensityLZ | 14.14 | 19.21 | 8.83 | 7.05 |
| DurLZ | 37.25 | 24.81 | 8.37 | 9.36 |
| H1*-H2* | 2.49 | 3.17 | 5.48 | 6.99 |
| H1*-A1 | 2.07 | 2.58 | 6.50 | 7.08 |
| H1*-A2 | 3.21 | 2.98 | 3.84 | 4.30 |
| H1*-A3* | 2.35 | 2.31 | 6.62 | 5.45 |
| F-measure | 0.87 | 0.89 | 0.73 | 0.76 |

*Table 5: Results of random forests analysis. Cross-validation of the Chinese vowels. Mead MDG values.*

| FEATURE | SF/ə/ | Demo/ə/ | Demo/a/ |
|---|---|---|---|
| TL | 2.62 | 4.20 | 4.45 |
| jitt_ppq5 | 4.26 | 4.46 | 6.58 |
| shim_apq5 | 2.75 | 4.12 | 4.73 |
| autoCorr | 3.82 | 4.53 | 3.91 |
| harm2noise | 3.63 | 5.11 | 5.22 |
| F0LZ | 4.78 | 11.87 | 13.89 |
| F1LZ | 5.51 | 6.73 | 5.10 |
| F2LZ | 4.06 | 6.16 | 3.48 |
| F3LZ | 2.93 | 4.48 | 6.50 |
| IntensityLZ | 19.21 | 9.78 | 7.44 |
| DurLZ | 24.81 | 5.35 | 3.87 |
| H1*-H2* | 3.17 | 7.76 | 9.87 |
| H1*-A1 | 2.58 | 6.88 | 5.22 |
| H1*-A2 | 2.98 | 4.33 | 5.96 |
| H1*-A3* | 2.31 | 3.63 | 3.24 |
| F-measure | 0.89 | 0.69 | 0.72 |

The last rows of Tables 4 and 5 show F-measure, a commonly used measure of classification performance. This is also the mean over the ten repetitions.

## Discussion

In tables 4 and 5, F-measure distributes in the range .69-.89, suggesting the effectiveness of the 15 acoustic features as the predictor variables. There is, however, difference of F-measure between the simple-filler and demo-filler categories in both languages. Vowels belonging to the simple-filler category showed higher F-values (.87-.89) than those belonging to demo-fillers (.69-.76) in both Japanese and Chinese.

In Table 4, as far as the vowels of simple-filler category are concerned, intensity and duration are among the most important. The importance of these prosodic features coincides with the conclusion reported in Maekawa and Mori (2016). In the category of demo-fillers, on the other hand, spectral features like F2 and harmonics to noise ratio make certain contribution to the classification; as the result, the contributions of prosodic features are smaller compared to simple-fillers.

In Chinese (Table 5), the situation is different. Prosodic features of duration is not as important as in Japanese in demo-fillers. F0 made large contribution in demo-fillers as in Japanese, but the influence of the variable is not the same as in Japanese. Table 2 shows that, in Japanese and where F0 showed significant difference, mean F0 is always lower in FP than in LX vowels, while in Chinese F0 is always higher in FP than in LX vowels (see Table 3).

Inter-language dissimilarities like this can be found in other acoustic features as well. When we compare the contribution of intensity in two languages, we found that in Japanese (See Table 2), mean intensity is always lower in FP than in LX vowels, while in the demo-fillers of Chinese, FP showed higher intensity than in LX vowels (see Table 3). In the same vein, autocorrelation and harmonics to noise ratio are lower in FP than in LX in Japanese, but in Chinese, they are higher in FP than in LX. Moreover, features concerning the spectral tilt of voice source (H1*-H2*) made certain contribution in Chinese, while their role in Japanese is limited.

Lastly, changes in formant frequencies between the FP and LX vowels is of some interest. Figure 1 compares the mean first (F1LZ) and second (F2LZ) formant frequencies of seven classes of vowels analysed so far. Rectangles and circles stand for the Japanese and Chinese vowels respectively. It can be seen in Figure 1 that while Japanese vowels do not show large displacement between the LX and corresponding FP vowels, Chinese vowels show larger displacements, especially in the /ə/ vowel.

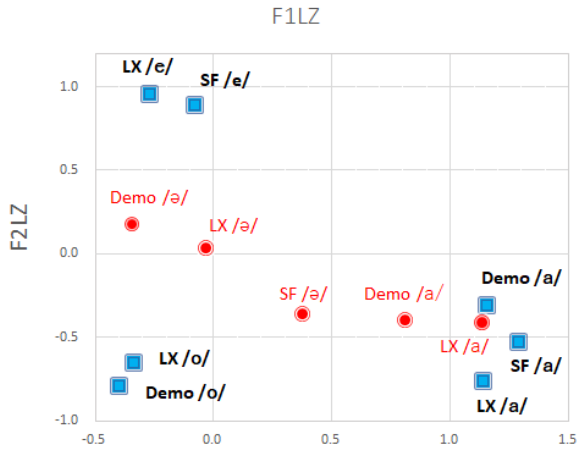



*Figure 1: Mean formant frequency of seven vowel classes. Unit is standard deviation.*

This difference is due probably to the typological difference of prosody between the two languages.

As some researchers believe, Chinese polysyllabic words have the specification of stress in addition to that of lexical tones (Lee, Tseng & Ouh-Young 1989), while Japanese is a pure pitch-accent language lacking any kind of stress.

To sum up, simple-fillers of Japanese and Chinese, as well as the demo-fillers of Japanese are similar in their behaviour. They are characterized by longer duration, lower intensity and lower F0. On the other hand, demo-fillers of Chinese are very different from all other fillers in that they are characterized by higher F0, higher intensity and they are not longer than LX vowels.

Also, they are characterized by lower shimmer, higher autocorrelation and higher harmonics to noise ratio. These features are characteristic of so-called clear and crisp speech. Our speculation is that this can be partly a result caused by the high-falling tone the lexical counterparts of the demo-fillers originally carry.

## Conclusion

The study reported here revealed three new findings.

First, as reported in previous studies dealing with Japanese, phonetic characteristics of the vowels in FP are systematically different from those in LX in Chinese as well. It is possible to automatically classify the vowels by means of random forests classifiers with the mean F-measure of about 0.8.

Second, vowels in simple-fillers are much easier to classify than the vowels in demo-fillers. This tendency is found in both languages.

Third, the substantial phonetic difference between the FP and LX items can be different dependent on both the category of FP and the languages. Especially, the Chinese demo-fillers make a phonetic class that is drastically different from all other filled pauses across two languages.

To conclude, we found both language-independent and language-dependent aspects of the phonetics of FP in the present study. More analysis is needed, however, for the fuller understanding of the issue. Especially, the analyses of Japanese dialogue data and Chinese monologue data are badly needed, although unfortunately, it is currently impossible due to the lack of suitable (i.e. phonetically annotated, large-scale, spontaneous speech) corpora other than the CSJ-Core and MCDC8.

## Acknowledgements

This work is supported by the JSPS Kakenhi grants to the first author (26284062). It was also supported by the MOST project grant to the third author (105-2410-H-001-084).

## References

Boersma, P. & D. Weenink. 2013. *Praat: doing phonetics by computer* [Computer program]. Version 6.0.21, retrieved from http://www.praat.org/

Hanson, H. 1997. Glottal characteristics of female speakers: Acoustic correlates". *Journal of the Acoustical Society of America*, 101 (1):466–481.

Lee. L.-S, C.-Y. Tseng & M. Ouh-Young. 1989. The synthesis rules in a Chinese text-to-speech system. *IEEE Transactions. Acoustics, Speech, and Signal Processing*, 37(9):1309–1320.

Maekawa, K. 2003. Corpus of Spontaneous Japanese: Its Design and Evaluation. *Proceedings of the ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR2003)*, 13–16 April 2003, Tokyo, Japan, 7–12.

Maekawa, K. & H. Mori. 2015. Voice-quality analysis of Japanese filled pauses: A preliminary report. Paper presented at DiSS 2015, Edinburgh, Scotland (no page numbers).

Maekawa, K. & H. Mori. 2016. Voice-quality difference between the vowels of filled pauses and ordinary lexical items. *Proceedings of Interspeech 2016,* 8–12 September 2016, San Francisco, USA, 3171–3175.

Tseng, S-Ch. 2014. Chinese disyllabic words in conversation. *Chinese Language and Discourse*, 5(2):23–51.

Zao, Y. & D. Jurafsky. 2005. A preliminary study of Mandarin filled pauses. *Proceedings of DiSS 2005*, 10–12 September 2005, Aix-en-Provence, France, 179–182.

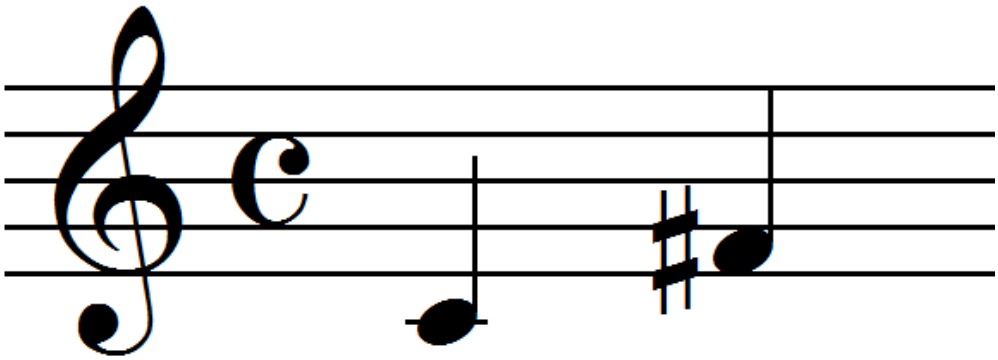
Proceedings of

# DiSS 2017

## The 8th Workshop on Disfluency in Spontaneous Speech

KTH Royal Institute of Technology
Stockholm, Sweden
18–19 August 2017

**TMH-QPSR**
Volume 58(1)



Edited by
Robert Eklund & Ralph Rose

Conference website: http://www.diss2017.org
Proceedings also available at: http://roberteklund.info/conferences/diss2017

Cover design by Robert Eklund
Graphics and photographs by Robert Eklund (except ISCA and KTH logotypes)

Proceedings of DiSS 2017, Disfluency in Spontaneous Speech
Workshop held at the Royal Institute of Technology (KTH), Stockholm, Sweden, 18–19 August 2017
TMH-QPSR volume 58(1)
Editors: Robert Eklund & Ralph Rose
Department of Speech, Music and Hearing
Royal Institute of Technology (KTH)
Lindstedtsvägen 24
SE-100 44 Stockholm, Sweden

ISSN 1104-5787
ISRN KTH/CSC/TMH–17/01-SE