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

Do GMM Phoneme Classifiers Perceive Synthetic Sibilants as Humans Do?

Gábor Pintér, Hiroki Watanabe

Graduate School of Intercultural Studies, Kobe University, Japan
g-pinter@port.kobe-u.ac.jp, watanabe@stu.kobe-u.ac.jp

Abstract

This study presents a psycholinguistically motivated evaluation method for phoneme classifiers by using non-categorical perceptual data elicited in a Japanese sibilant matching 2AFC task. Probability values of a perceptual [s]-[ʃ] boundary, obtained from 42 speakers over a 7-step synthetic [s]-[ʃ] continuum, were compared to probability estimates of Gaussian mixture models (GMMs) of Japanese [s] and [ʃ]. The GMMs, trained on the Corpus of Spontaneous Japanese, differed in feature vectors (MFCC, PLP, acoustic features), covariance matrix types (full, tied, diagonal, spherical), and numbers of mixtures (1-20). Using ten-fold cross validation, it was found that GMMs trained on MFCC features had the best sibilant classification accuracies (87.4-90.4%), but their correlations with human perceptual data were non-conclusive (0.35-0.98). Acoustic feature-based GMMs with tied covariance matrices had near human-like synthetic stimuli perception (0.957-0.996), but their classification performance was poor (71.3-80.4%). Models trained on perceptual linear prediction (PLP) features were on par with the acoustic feature-based models in terms correlation to the perceptual experiment (0.884-0.995), while losing slightly on classification performance (86.1-88.9%) compared to MFCC models. Across the board correlation tests and mixture-effect models confirmed that GMMs with better sibilant classifying performance produced more human-like probability estimations on the synthetic sibilant continuum.

Index Terms: perception, Japanese sibilants, synthetic continuum, ASR-HSR comparison

1. Introduction

Studies in both human and automatic speech recognition (HSR, ASR) are concerned with the problem of how linguistic information—in most of the cases phonemes and words—can be extracted from the acoustic speech signal. It is a well-known fact that despite the trivial similarity of the abstract goals, the two fields of HSR and ASR greatly diverge in terms of focus and methodology [1], [2], [3]. Traditional ASR research represents a holistic, performance-oriented approach to speech recognition. Higher accuracy and lower resource footprints are the ends that justify any means. Other than human-like recognition accuracy, psycholinguistic or perceptual aspects of speech recognition are usually not of direct concern in ASR systems. HSR studies, on the other hand, have a rather narrow focus, typically addressing some specific aspects of speech recognition, such as perceptual similarity of phonemes [4], temporal characteristics of recognition [5], or phonotactic effects [6]. Unlike ASR approaches, HSR studies do not aim to present an exhaustive explanation of the whole speech recognition process.

Despite all the palpable differences between ASR and HSR approaches, a convergence would be beneficial for both fields. For HSR studies, a comprehensive working model could be a proof of concept for psycholinguistic models. As for ASR, greater psychological plausibility, that is proximity to human-like recognition concerning various aspects of the process, has the promise of eventually leading to better recognition accuracies. In order to bring the two fields closer together, however, the differences have to be assessed first. Comparisons have been carried out among others in terms of recognition accuracy [7], effects of training data [1], reaction time [8], and phoneme confusion [9]. The current study aims to provide an addition to the ASR-HSR literature by comparing human and machine performance in a sibilant discrimination task—using synthetic sibilant continuum. The sibilant discrimination task was chosen to keep both human experiments and statistical models simple. Since sibilants can be reliably identified by spectral cues, their perception can be modeled with context independent models, such as Gaussian mixture models (GMMs). The human data for this study was obtained in a perceptual experiment using a 7-step [s]-[ʃ] continuum as stimuli. The GMM classifiers were trained over the Corpus of Spontaneous Japanese.

The current research raised the following questions. First, do GMM phoneme classifiers trained on Japanese utterances 'perceive' the synthetic stimuli from the human experiment as humans do? Concretely, do probability scores estimated by statistical sibilant classifiers correlate with probability scores from the human experiment? Second, what settings and features of the statistical models produce closer proximity to human perception? Third, would closer resemblance to human perceptual characteristics entail better classification performance? In other words, can recognition accuracies be predicted based on how close the resemblance to human performance in the sibilant classification task is?

2. Sibilant perception by humans

Sibilants are a relatively deeply researched topic in Japanese phonetics [10], [11], [12], [13], [14]. Japanese has two voiceless sibilants: the alveolar [s] and the alveo-palatal [ɕ] (for brevity transcribed here as [ʃ]). While Japanese and English [s] sounds are very similar, the Japanese [ʃ], differently from its English counterpart, lacks lip rounding and pronounced further

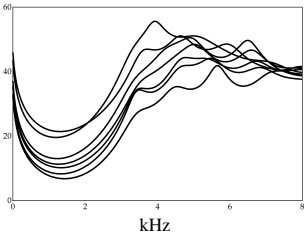


Figure 1: *Spectral envelopes for the [s]-[ʃ] continuum.*

http://dx.doi.org/10.21437/Interspeech.2016-325

back in the mouth [15]. As for perception, English listeners tend to identify sibilants by their spectral characteristics [16], and rely to a significantly lesser extent on formant transitional cues [17], [18]. In Japanese, formant cues weigh in more heavily than spectral ones [10]. Since sibilants can only appear in the onset constituent of the Japanese syllable, transitional cues are readily available. Vowel devoicing, however, can create forms in which sibilants end up in non-vocalic context, for example, at the end of the word (e.g., /desu/→[desu̥]→[des] COPULA). Despite the strong reliance on transitional cues, Japanese listeners have no problem identifying sibilants in these non-vocalic environments.

2.1. Stimuli

A 7-step [s]-[ʃ] continuum (S1-S7) was synthesized and appended as the final consonant to the carrier phrase /kono ka-/. The carrier phrase was obtained by truncating a natural utterance of [kono kasu tte nani]. The synthetic stimuli ranged between [kono kas] 'this residue' and [kono kaʃ] 'these lyrics'.

2.2. Experimental design

In order to avoid orthographic influence an audio-only XAB experiment was designed in which the participants were asked to listen to a sentence and two samples $\langle AB \rangle$, and select the sample that corresponded to the last word $\langle X \rangle$ of the sentence. The samples were presented in both $\langle AB \rangle$ and $\langle BA \rangle$ order. In order to avoid direct acoustic comparisons, and facilitate processing at phonological level the target and candidate words were separated by a relatively long (900ms) silence and a beep. A longer beep (500ms) was used to mark the beginning of the trial, a shorter one (200ms) to mark the samples to choose from.

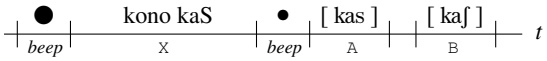
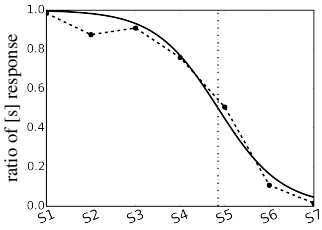


Figure 2: *Experiment trial with XAB design.*

The stimuli were presented through a custom desktop application with a graphical interface. Buttons with label A and B were prepared to collect the responses. The 7-step continuum was presented in two sample orders, repeated 6 times each, yielding in $7 \times 2 \times 6 = 84$ trials per session. The experiment, including a brief training session took around 20 minutes.

2.3. Results

The experiment was carried out in two PC rooms with 63 Japanese undergraduate students. After a brief orientation, the participants familiarized themselves with the experiment program using only the least ambiguous S1 : [s] and S7 : [ʃ] stimuli. After the practice, they performed the experiment individually. The participants were allowed to repeat the audio stimuli and to revise their responses. Some participants considered the experiment too repetitive, and reported difficulties in concentration. Thus, responses for those participants who did not reach 85% accuracy for the least ambiguous S1 and S7 stimuli were removed from the results. This clean-up left 42 participants in the dataset. Figure 3 summarizes the responses in the form of [s] response ratios averaged over 42 participants. The fitted sigmoid has the slope of -1.41 and a center between S4 and S5 (4.85).



stimuli	S %
S1	98.413
S2	87.500
S3	90.873
S4	75.794
S5	50.596
S6	10.714
S7	1.587

Figure 3: *Ratio of [s] responses with fitted sigmoid.*

3. Sibilant recognition with ASR

3.1. Sibilant recognition with GMMs

Representing the machine aspect of phoneme recognition, a range of GMMs were trained using the Corpus of Spontaneous Japanese. From the core part of the database 52,776 [s] and 33,748 [ʃ] labeled segments were extracted. The middle 30ms portion of these sibilants formed the basis of feature extraction. OpenSMILE [19] was used to calculate 13 dimensional MFCC and 6 dimensional PLP features over 20ms windows with 10ms shift. With the delta features added, 26 dimensional MFCC and 12 dimensional PLP feature vectors were created for each segment. In addition to these typical ASR features, six widely used phonetic features (center of gravity, skewness, spectral variance, root mean square and zero crossing rate) were also extracted over the same 30ms frame—using a custom Python script. These 6 values formed the third type of feature vector, labeled as ACU below.

A range of Gaussian mixture models for [s] and [ʃ] were trained with the three feature vectors (MFCC, PLP, ACU), combined with four different covariance types (full, tied, diagonal, spherical) and mixture numbers ranging from 1 to 20. In a ten-fold cross validation setup, the GMMs were trained on 9 folds, and evaluated against the unseen fold. The training/testing folds were rotated in 10 steps, resulting in 10 accuracy scores for each settings. The same randomly defined partitions were used across all feature vectors, covariance types and mixture number conditions.

Without aiming to draw general conclusions it can be claimed that for this particular sibilant recognition task full co-

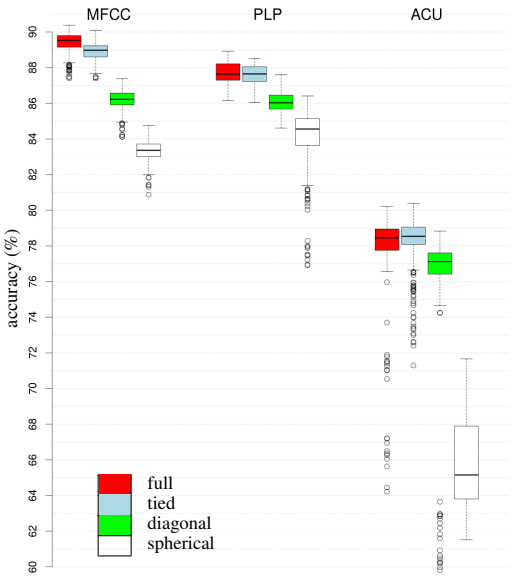


Figure 4: *Sibilant classification accuracies over 10 folds and 1-20 mixtures.*

variance GMMs trained on MFCC features provided the best recognition accuracy, topping at 90.372%. PLP-based models had just slightly lower performance than MFCC-based models, acoustic features produced the lowest scores. Spherical covariance models were the least accurate in all training feature conditions.

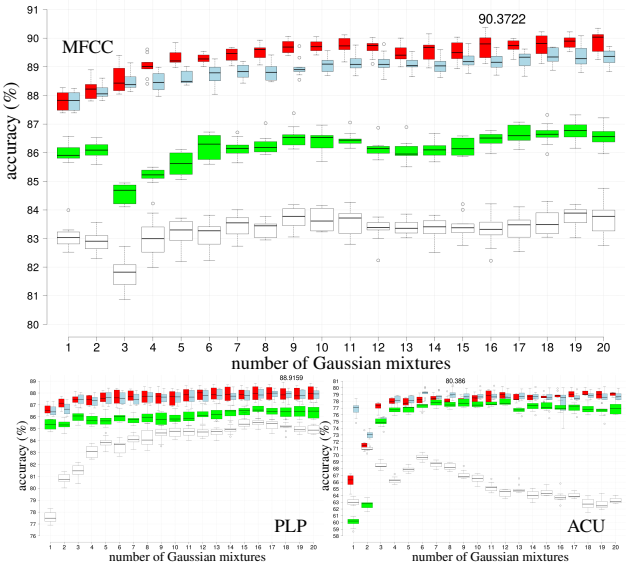


Figure 5: *Sibilant recognition accuracies. Each box summarizes results from 10 test folds.*

Figure 5 shows further details of the recognition performance with MFCC-based GMMs. While greater mixture numbers contributed to better recognition accuracies in general, the tendency saturates at around 10 mixtures. GMMs with spherical covariance matrices did not seem to benefit from higher mixture numbers at all.

3.2. Correlation with human data

Using the same feature extraction methods as explained in the previous subsection, the trained GMMs were tested against the synthetic S1-S7 stimuli used in the human experiment. By subtracting the log-probability scores estimated by Gaussian mixture models of [s] and [ʃ] the logarithm of probability ratios were calculated.

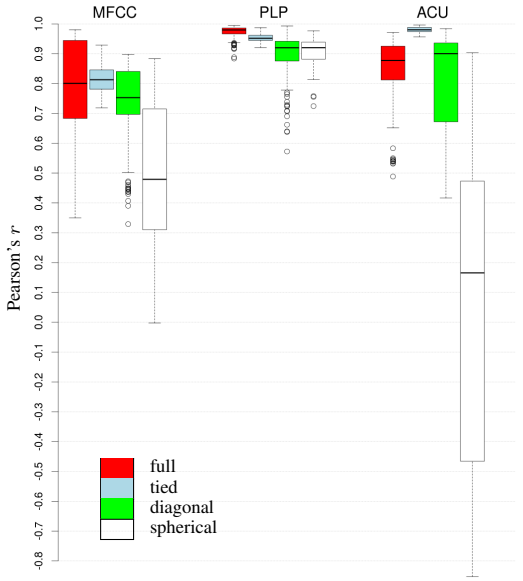


Figure 6: *Pearson correlation estimates over 10 folds and 1-20 mixtures.*

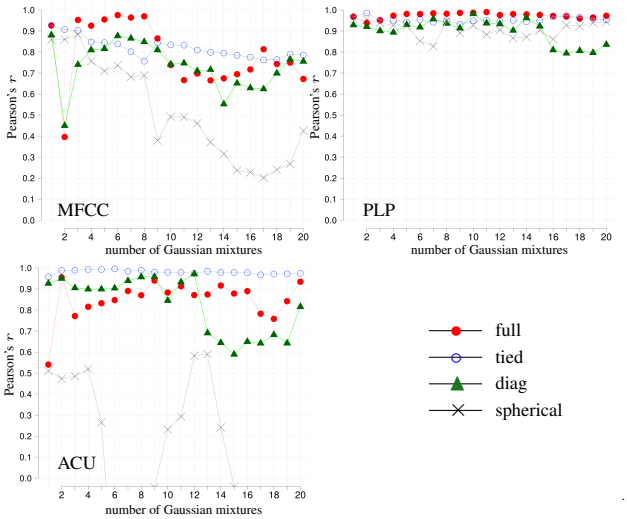


Figure 7: *Correlation with human sibilant classification. Each data point is the mean of Pearson's* rs *over 10 test folds.*

$$log(p([s])) - log(p([\int])) = log\frac{p([s])}{p([\int])} \quad (1)$$

The raw probability values from the human experiment were also converted into log probability ratios. Since there were only two alternatives in the forced choice task, probability values for [ʃ] could be calculated as $1 - p([s])$.

$$log\frac{p([s])}{1-p([s])} = log\frac{p([s])}{p([\int])} \quad (2)$$

Proximity to human perception was quantified as the correlation between (1) and (2). Pearson product moment correlation coefficients were chosen over rank correlations because it can test perceptual *distance* relations.

For example, as the data from the human experiment testify (cf. Figure 3) moving from stimuli S5 to S6 represents a relatively great increase in the [ʃ]-likeness of the percepts, while a move from stimuli S6 to S7 is relatively small. Rank correlation alone cannot explain this type of relations.

The strongest correlation with human data was achieved by acoustic features with tied covariance matrix models r={0.957, 0.996}. This high correlation, however, did not correspond to high sibilant recognition accuracies (cf. Figure 4: 71.3-80.4%). As the spherical covariance models demonstrate, GMMs trained on acoustic features were responsible not only for the highest, but also for the lowest levels of correlations with the sibilant experiment. PLP features, in contrast, showed a consistently good performance—with the means of correlation coefficients staying above 0.9 in all covariance conditions. Human correlation for MFCC-based GMMs were inconclusive. Full covariance GMMs with less than 9 mixtures had a steady and strong correlation with human responses, but in other cases Pearson's r showed relatively wide variation. As it is demonstrated in Figure 7, an increase in the number of mixtures did not necessarily improved correlation with the human data.

3.3. Score correlation

Figure 8 shows the relation between classification accuracy scores (vertical axes), and correlation coefficients calculated

in comparison with the perceptual [s]-[ʃ] boundary (horizontal axes).

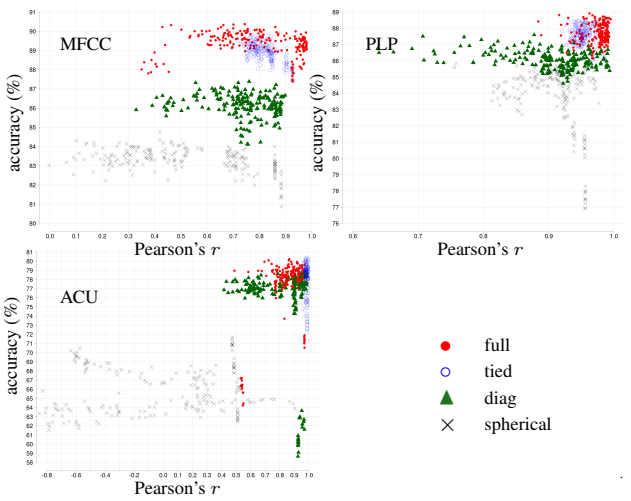


Figure 8: *Accuracy versus correlation to human perception of synthetic sibilants.*

Although the data points seem to lean towards the top-right corner of the plots, correlation tests outline a somewhat contradictory picture. As shown in Table 1, while correlation tests within covariance conditions tend to show negative or no significant correlations, sibilant classification accuracy and human-like perception display statistically significant positive correlation if conflating data within MFCC, PLP and ACU conditions. Regarding the whole data set, the correlation is statistically significant (Kendall's $\tau = 0.112$, $p < 0.001$).

Table 1: Kendall's τs for testing correlation between sibilant classification accuracy and human-like perception.

	MFCC		PLP		ACU	
	τ	p	τ	p	τ	p
full	−0.197	***	0.057	ns	0.124	**
tied	−0.411	***	−0.0003	ns	−0.034	ns
diag	−0.111	*	−0.194	***	0.025	ns
spher	−0.271	***	−0.192	***	0.076	ns
(all)	0.352	***	0.528	***	0.612	***

In order to explore the interaction between sibilant classification accuracy and perceptual correlation, a mixed-effect model was used on the data. The model used accuracy (`acc`) as the dependent variable; feature type (`feat`), covariance type (`cov`), number of mixtures (`mix`), and correlation to human responses (`corr`) as fixed effects; test fold (`fold`) as random effect. Since, as discussed above, types and mixture numbers showed variations as a function of feature type, their interactions `feat:cov` and `feat:mix` were inserted into the model.

```
acc~feat+feat:cov+feat:mix+corr+(1|fold)        (3)
```

The mixed effect model also found that correlation with human perception was statistically significant ($\chi^2(1)$=28.784, p<0.0001), though only to a small extent: 1.281 ± 0.238 (Table 2).

Table 2: Extract from the mixed-effect model summary.

```
                Est.  Err.   df      t Pr(>|t|)
(Intercept)   74.685 0.309  819 242.05  < 2e-16 ***
featmfc       12.829 0.254 2390  50.57  < 2e-16 ***
featplp       10.641 0.253 2390  41.99  < 2e-16 ***
corr           1.281 0.238 2391   5.38 8.11e-08 ***
...              ...   ...  ...    ...      ...
featacu:mix    0.168 0.012 2390  14.09  < 2e-16 ***
featmfc:mix    0.081 0.012 2390   6.73 2.11e-11 ***
featplp:mix    0.106 0.012 2390   9.20  < 2e-16 ***
...              ...   ...  ...    ...      ...
```

4. Summary

Did GMM phoneme classifiers perceive synthetic sibilants as humans do? The overall answer was yes, but closer inspection revealed a relatively complex picture.

First, most GMM sibilant classifiers—trained on the Spontaneous Corpus of Japanese—showed relatively strong correlation with human perceptual boundaries. Since the relatively strict Pearson correlation was used to calculate human-machine correlation, this result presents a strong argument for a positive answer to the question in the title.

Second, training parameters were found to have strong influence on the human-machine correlation. Acoustic features with tied covariance matrices had the best correlation with human perception (r={0.957, 0996}), although they performed badly in the classification task. The strong correlation with human-like perception was not unexpected as the synthetic stimuli for the human experiment had been created with specific acoustic features in mind (e.g., center of gravity). However, the poor accuracy with natural data implies that the chosen acoustic features describe only limited—albeit well-researched—perceptual aspects of the sibilants. It was also discovered that unlike in the case of classification accuracy, more mixtures did not necessarily provide better correlation with human perception of the synthetic stimuli. As for training feature vector types, PLP-based models provided consistently high correlation scores, outdoing MFCC models both in consistency and in correlation coefficient values.

Third, the relation between classification accuracy and correlation with human-like perception was investigated. The initial hypothesis was that better classifiers correlate more strongly to human-like perception of the synthetic stimuli. While statistical tests verified this hypothesis within larger partitions of the data, smaller partitions actually showed an inverse correlation. Within the four covariance conditions, classification accuracy was most cases found to be negatively correlated with proximity to human-like perception. This trade-off relation could be an indicator of a mismatch between statistical models and human perception. The acoustic details that statistical models are successfully relying on in classification tasks are most probably different from the ones human use in perception.

5. Acknowledgements

This work was supported by The Ministry of Education, Culture, Sports, Science and Technology KAKENHI (26770141) awarded to the first author. Special thanks to professor Michinao Matsui for his help during the preparation of this paper and to Ellen Rettig-Miki for her suggestions on the manuscript.

6. References

[1] R. K. Moore and A. Cutler, "Constraints on theories of human vs. machine recognition of speech," in *Proceedings of the Workshop*

on Speech Recognition as Pattern Classification, Nijmegen, The Netherlands, 2001, pp. 145–150.

[2] O. Scharenborg, J. M. McQueen, L. ten Bosch, and D. Norris, "Modelling human speech recognition using automatic speech recognition paradigms in SpeM," in *Proceedings of Eurospeech*, Geneva, Switzerland, 2003, pp. 2097–2100.

[3] O. Scharenborg, D. Norris, L. ten Bosch, and J. M. McQueen, "How should a speech recognizer work?" *Congnitive Science*, vol. 29, pp. 867–918, 2005.

[4] G. A. Miller and P. Nicely, "An analysis of the confusion among English consonants heard in the presence of random noise," *The Journal of the Acoustical Society of America*, vol. 26, no. 5, pp. 953–953, 1954.

[5] W. D. Marslen-Wilson, "Functional parallelism in spoken word-recognition," *Cognition*, vol. 25, pp. 71–102, 1987.

[6] J. Coleman and J. Pierrehumbert, "Stochastic phonological grammars and acceptability," in *Proceedings of the Third Meeting of the ACL Special Interest Group in Computational Phonology*. Somerset, New Jersey: Association for Computational Linguistics, 1997, pp. 49–56.

[7] R. P. Lippmann, "Speech recognition by machines and humans," *Speech Communication*, vol. 22, no. 1, pp. 1–15, 1997.

[8] L. ten Bosch, M. Ernestus, and L. Boves, "Comparing reaction time sequences from human participants and computational models," in *Proceedings of Interspeech*, Singapore, Singapore, 2014, pp. 462–466.

[9] B. T. Meyer, M. Wächter, T. Brand, and B. Kollmeier, "Phoneme confusions in human and automatic speech recognition," in *Proceedings of Interspeech*, Antwerp, Belgium, 2007, pp. 1485–1488.

[10] S. Funatsu and S. Kiritani, "Cross language study of perception of dental fricatives in Japanese and Russian," *Annual Bulletin of Research Institute of Logopedics and Phoniatrics*, no. 28, pp. 69–72, 1994.

[11] ——, "Effect of following vowel on perception of second language fricatives -native language interference in Russian learners of Japanese-," *Journal of the Phonetic Society of Japan*, vol. 4, no. 2, pp. 72–80, 2000.

[12] S. Hirai, K. Yasu, T. Arai, and K. Iitaka, "Acoustic cues in fricative perception for Japanese native speakers," *Technical Report of the Institute of Electronics, Information and Communication Engineers*, vol. 104, no. 696, pp. 25–30, 2005.

[13] H. Takeyasu, "Effects of the spectral properties of frication on perception of singleton/geminate fricatives," *Phonological Studies*, vol. 12, pp. 43–50, 2009.

[14] M. Matsui, "Shisatsu-on no yuuseisei to museikaboin no chikakuteki tegakari ni tsuite," in *The bulletin of the Phonetic society of Japan*, 2013, pp. 35–40.

[15] H. Y. Pan, A. Utsugi, and S. Yamazaki, "An acoustic phonetic study of voiceless alveolo-palatal fricatives in Japanese, Korean and Chinese," *Journal of General Linguistics*, vol. 7, pp. 1–27, 2004.

[16] K. S. Harris, "Cues for the discrimination of American English fricatives in spoken syllables," *Language and Speech*, vol. 1, pp. 1–7, 1958.

[17] W. Jassem, "The formants of fricative consonants," *Language and Speech*, vol. 8, pp. 1–16, 1965.

[18] V. A. Mann and B. H. Repp, "Influence of vocalic context on perception of the [ʃ]-[s] distinction," *Perception & Psychophysics*, vol. 28, pp. 213–228, 1980.

[19] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*, Barcelona, Spain, 2013, pp. 835–838.