24th INTERNATIONAL CONGRESS ON SOUND AND VIBRATION
23–27 July 2017, London

ICSV24
23-27JULY 2017
LONDON CALLING

FUNDAMENTAL RESEARCH OF SPEECH DISCRIMINATION FOR TONGUE THRUST USING ZERO-CROSSING AND MEL-FREQUENCY CEPSTRUM COEFFICIENTS

Masashi Nakayama, Shunsuke Ishimitsu

Graduate School of Information and Sciences, Hiroshima City University, Hiroshima, Japan
email: masashi@hiroshima-cu.ac.jp

Kimiko Yamashita, Kaori Ishii, Kazutaka Kasai, Satoshi Horihata

School of Dentistry at Matsudo, Nihon University, Chiba, Japan

This study investigates the differences in speech between the acoustic features of healthy people and patients of tongue thrust. Compared with healthy participants, patients of tongue thrust exhibit broadband spectral frequencies in speech accompanied by the non-stationary sound of a specific consonant. This paper proposes and tests two well-established acoustic features, zero-crossing and the mel-frequency cepstrum coefficients (MFCCs), to distinguish healthy speakers from those with tongue thrust. Zero-crossing here is the point where the frequency parameter changes sign, and is used to discriminate between voiced and unvoiced sections. The MFCCs represent cepstrum coefficients calculated on the mel scale of frequency followed by tone pitch. According to the results of tests of the proposed method on clinical datasets, significant differences between healthy participants and tongue thrust patients were observed in terms of both acoustic features, with $p < 0.2$ and $p < 0.01$, when conditioned on oral habits, gender, and age.
Keywords: Oral habit, tongue thrust, discrimination, zero crossing, MFCC

1. Introduction

Certain oral habits, such as tongue protrusion, adversely affect speech as well as the performance of some functions, such as chewing and the enunciation of words. In general, the acquisition of proper oral habits early on is important in a child's development. These habits include the protrusion of the tongue and the manner of breathing. In particular, tongue protrusion affects all teeth and can hinder their arrangement. Dentists have sought to improve people's oral habits over the years using oral myofunctional therapy (MFT) [1]. However, a number of people are negligent of their oral health in general, and hence do not often resort to dental consultation. Against this backdrop, the authors of this paper propose a method to distinguish people tongue habits or oral habits using differences in the acoustic features of their speech. A user introduces the system as application software, and then articulates a word, the acoustic characteristics of which are easy to capture. It is expected that the implementation of such a method can help improve people's oral habits at early stages of development.

A number of studies have been conducted on distinguishing tongue habits and positions, estimating tongue position using formant analysis [2], and measuring the duration of the articulation of consonants and sound pressure in patients with mandible deformities [3]. Of methods that address the evaluation of enunciation and the estimation of tongue position, the speech pronunciation evaluation system [4], Ami Voice [5] and the ATR CALL pronunciation challenge are representatives [6]. Other methods have investigated the identification of improper tongue protrusion, and the extraction and evaluation of similarities in pronunciation between trainees and native speakers using

acoustic features, such as the mel-frequency cepstrum coefficients (MFCC), formants, fundamental frequencies and utterance speed. A representative method in this vein involves pronunciation training using acoustic features in an articulator [7]. However tongue position could be estimated using a pronunciation map, where the system did not focus on clinical participants. Articulatory feature analysis has also been proposed for tongue position estimation [8, 9]. This paper proposes a novel approach to tongue habit discrimination using acoustic features from speech recognition [10]. To discriminate tongue habits, the algorithm requires the significance of and/or differences in the acoustic parameters between normal speakers and those suffering from tongue thrust. The authors also tested the proposed system through experiments using the acoustic features of zero-crossing and the MFCC for discrimination.

The remainder of this paper is organized as follows: Section 2 provides an overview of the proposed system and the algorithm for discrimination, and Section 3 gives details of two datasets, a position-dependent dataset and a clinical dataset. Section 4 discusses the acoustic features used and evaluates significance of and space separation between healthy speakers and those with tongue thrust. Finally, Section 5 concludes this paper and offers directions for future work.

2. Tongue habit discrimination

This section provides an overview of tongue habit discrimination based on a distance minimization criterion [10]. The system can be implemented by extending conventional speech recognition, which aims to estimate a word or a word sequence in speech using feature parameters and a decoding algorithm. However, it focuses on distinguishing tongue protrusion. The system and its algorithm are hence extended on an acoustic model conditioned on tongue position, as shown in Figure 1. With the extension, the system can discriminate a tongue habit and the position of the tongue during pronunciation. Figure 2 shows an overview of the system for tongue habit discrimination. The proposed system not only implements statistical modeling through a probability model, but can also introduce comparisons based on distance criteria if the parameters and their distributions slightly vary. Equation (1) expresses tongue habit discrimination based on the minimized distance criterion:

$$\hat{v} = \arg\min_{v \in V} d\left(\left| y_v - y \right|\right) \tag{1}$$

where y_v is the feature of the acoustic model, and v represents a protruding tongue in tongue space V. By Equation (1), the estimated candidate for protruding speech is minimized based on the distance between the acoustic features of speech y and model parameter y_v. The scalar value of the acoustic feature in Equation (1) also extends into vectors if the user requires robust discrimination. Considering this idea, Equation (2) defines the vectors and weights:

$$\hat{v} = \arg\min_{v \in V} \left[\lambda_1 d\left(\left| y_{v,1} - y_1 \right|\right) + \cdots + \lambda_x d\left(\left| y_{v,x} - y_x \right|\right)\right] \tag{2}$$

where λ_x represents a weight at the x-th dimension of the acoustic feature; the weight can be adjusted according to effectiveness. A method based on Equation (2) is shown as an example for introducing a vector parameter. One approach has represented acoustic features thus [11].

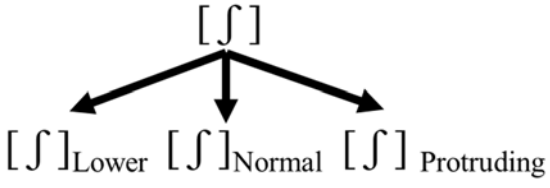


Figure 1: Acoustic modeling for tongue habits.

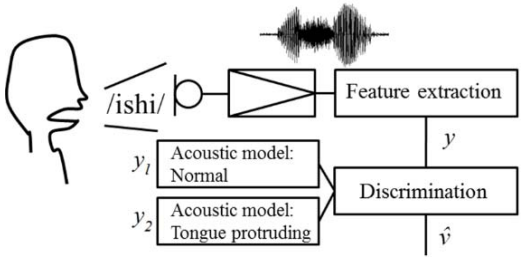


Figure 2: Tongue habit discrimination system.

3. Speech dataset of tongue thrust patients

Speech datasets were collected for this research. There were two datasets, a tongue position-dependent dataset and a clinical dataset. Table 1 shows the recording conditions for them. Speech was recorded in a soundproof consultation room for a hearing experiment. The recoding conditions for speech were prepared with a condenser microphone, Audio Technica ATM31a, located between the participant and the microphone. The sampling frequency with resolution was 22.05 kHz in 16 bits, and the down-sampled speech was analyzed at 16 kHz. Each participant pronounced a word /ishi/ chosen as suitable for discrimination the acoustic characteristics of protruding speech. The word were composed several smaller words featuring the characteristic consonant /s/, the IPA representation of which is [ʃ]. Because of the use of the word, acoustic characteristics could be measured as a sound as a speaker sandwiched the tip of the tongue between the upper and lower frontal teeth. Table 2 shows the conditions for Datasets 1 and 2. Dataset 1 consisted of speech conditioned on protruding speech pronounced by dentists. Dataset 1 was composed from nine trials times five conditions: normal, protruding, mandibular, lower and mandibular with lower. Dataset 2 consisted of speech recorded by children and adults, some of whom were healthy and others were patients of tongue thrust. Dataset 2 was composed of three trials times two conditions: normal and protruding. The healthy adult speakers had had narration training for two years or more, and those with protruding speech were those who tongue pronouncing utterances have over bite within 0 mm or less, and children with pronouncing with tongue protruding and swallowing who are discriminated by dentists.

Table 1: Recording conditions for Datasets 1 and 2

Recording position	Consultation Room, School of Dentistry at Matsudo, Nihon University
Environment	Calm room
Vocabulary	/ishi/ for pronunciation by tongue thrust patients
Microphone	Autio Technica ATM31a
Recording equipment	Roland EDIROL UA-25EX
Condition	Recorded at 22.05kHz, 16 bits, and analyzed at 16kHz, 16 bits.

Table 2: Participants' conditions for Datasets 1 and 2

Dataset	Subject	Condition
Dataset 1	Dentist, Adult	No protruding tongue. Normal, Protruding, Mandibular, Lower and Mandibular with Lower. 1 Male and 1 Female (Average: 27.0 years old) × 5 manner × 9 trials = 90 sample
Dataset 2	Normal, Adult	No protruding tongue with narrator training for over 2 years. 10 Males and 10 Females (Average: 26.3 years old) × 1 manner × 3 trials = 60 samples
	Protruding, Adult	Protrude tongue with over bite less than 0 mm. 5 Males and 10 Females (Average: 29.4 years old) × 1 manner × 3 trials = 45 samples
	Normal, Child	No protrude tongue. 10 Males and 10 Females (Average: 9.7 years old) × 1 manner × 3 trials = 60 samples
	Protruding, Child	Protrude tongue. 9 Males and 11 Females (Average 9.6 years old) × 1 manner × 3 trials = 60 samples

4. Acoustic feature analysis

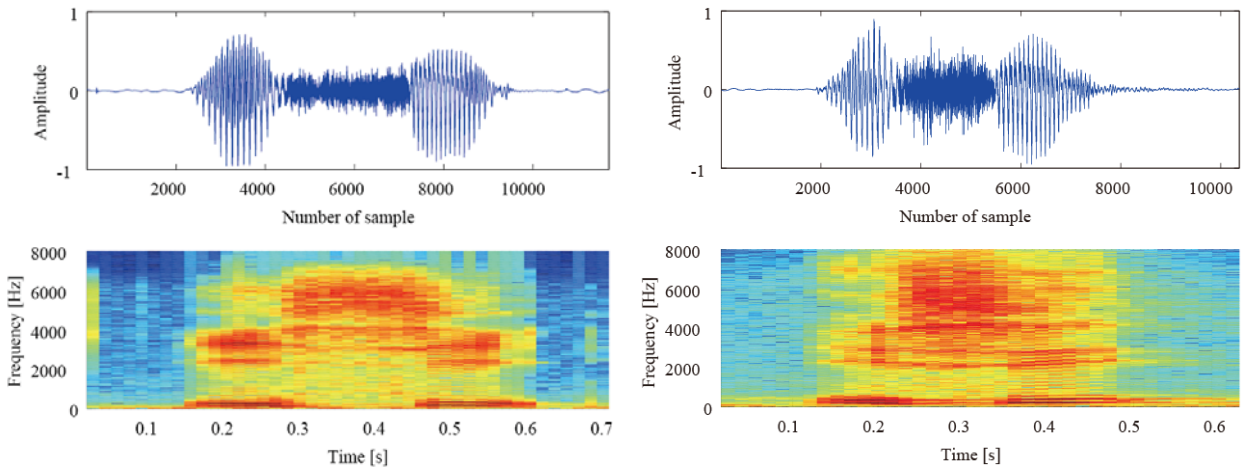
Measurement experiments were performed for time frequency analysis and the extraction of acoustic features with normal and protruding speech. The time frequency analysis confirmed the characteristics of frequency in protruding and normal speech as well as the procedures to extract signals and features.

4.1 Time frequency analysis

Time frequency analysis locates differences in sounds using a spectrogram. Speech features continuously changing characteristics because it is composed of several words. Hence, a short time frequency analysis is commonly used for it because the sound of speech in short intervals can be considered stationary by using the window functions of Hamming or Hanning. The method in Equation (3), Spectrogram $Y(t,f)$, was employed in the experiment to find the differences:

$$Y(t,f) = \int_{-\infty}^{+\infty} y(\tau) w^*(\tau - t) e^{-j2\pi f\tau} d\tau \quad (3)$$

where $y(\tau)$ is the analysis target, t is the time sequence, f is frequency and $w(t)$ is the window function. An STFT (short-time Fourier transform) was used because it ranges between –Inf. and +Inf. at sequence signal $y(t)$. The STFT calculates, with a Fourier series expansion, speech $y(t)$ in the range of -Inf. to +Inf., and assumes that it is framed with window function $w(t)$ of finite length T_0 and the section periodically continues. Figure 3 shows an example /ishi/ of time frequency analysis using STFT. The protruding speech in Figure 3 confirmed a broadband frequency with the consonant duration, since the consonant was intonated at the frontal teeth. The characteristics were quite different from those evident in the enunciation of vowel sounds, which had a resonance frequency called the formant frequency.

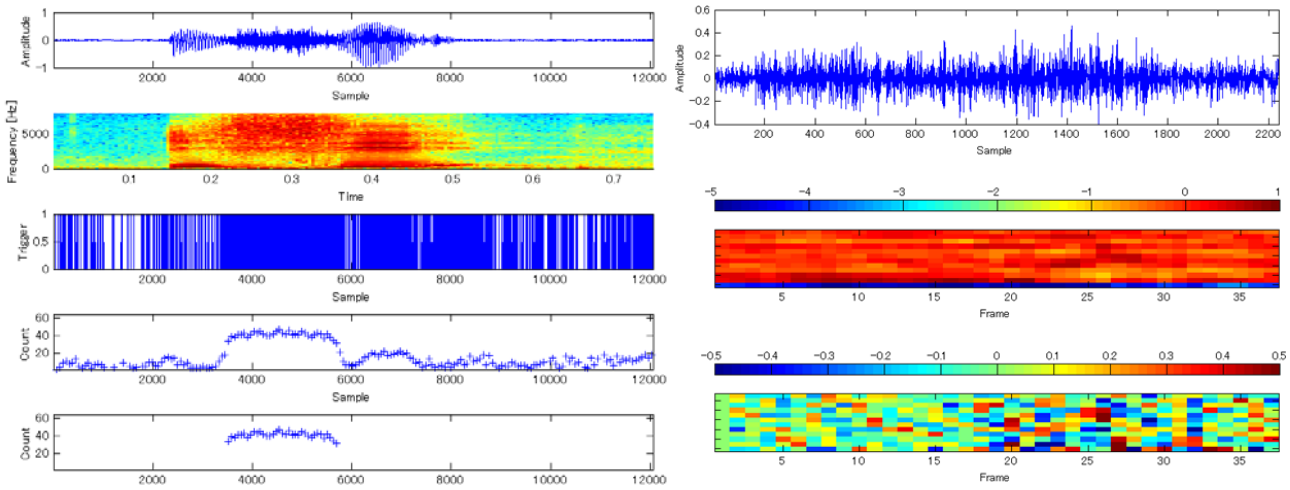


(a) Normal speech. (b) Protruding speech.

Figure 3: Time frequency analysis using spectrogram.

4.2 Acoustic feature analysis using zero-crossing and MFCC

Figure 4 shows the feature extraction for zero-crossing, MFCC and ΔMFCC [12, 13]. The waveform of /ishi/ and its time frequency analysis are shown in the first and the second parts of Figure 4(a). As the basis of the waveforms, the measured zero-crossing is shown in the middle row of Figure 4(a), is confirmed detection of consonant duration referred counting number of zero-crossing where the consonant section increased. Followed the result, consonant duration extracts and decides are shown in lowest row at Figure 4(a). By this procedure, the consonant section of the sound can be extracted as shown in the sub-figure of Figure 4(b). Finally, the extracted sound calculated MFCC and ΔMFCC are shown in middle and lowest parts of Figure 4(b).



(a) Zero-crossing. (b) MFCC and ΔMFCC.

Figure 4: Feature extraction using zero-crossing, MFCC and ΔMFCC.

5. Significance and space separation of acoustic features

To capture the sound characteristics of tongue protruding with unvoiced sound that is a fricative, zero-crossing and MFCC were used.

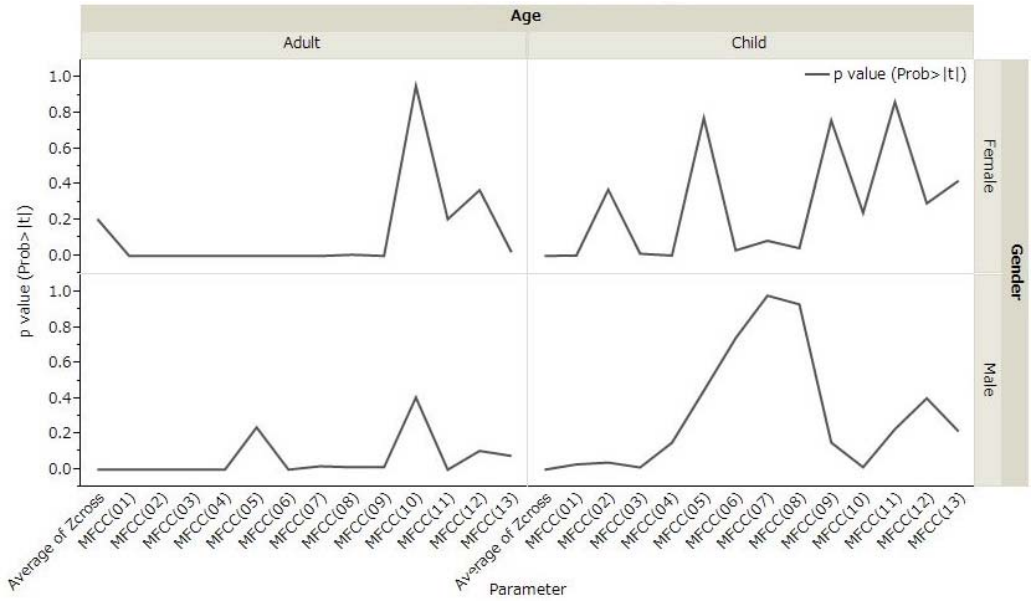
5.1 Experimental conditions

Zero-crossing, MFCC and ΔMFCC were extracted using each data item from Datasets 1 and 2. Table 3 shows the experimental conditions. Each parameter obtained from the vectors of the acoustic features at the time sequences was analyzed at each frame. Therefore, the representative feature vector had 27 dimensions (zero-crossing (1) + MFCC (13) + ΔMFCC (13) = 27). The statistical analysis them confirmed the significance and the space separation of each acoustic feature.

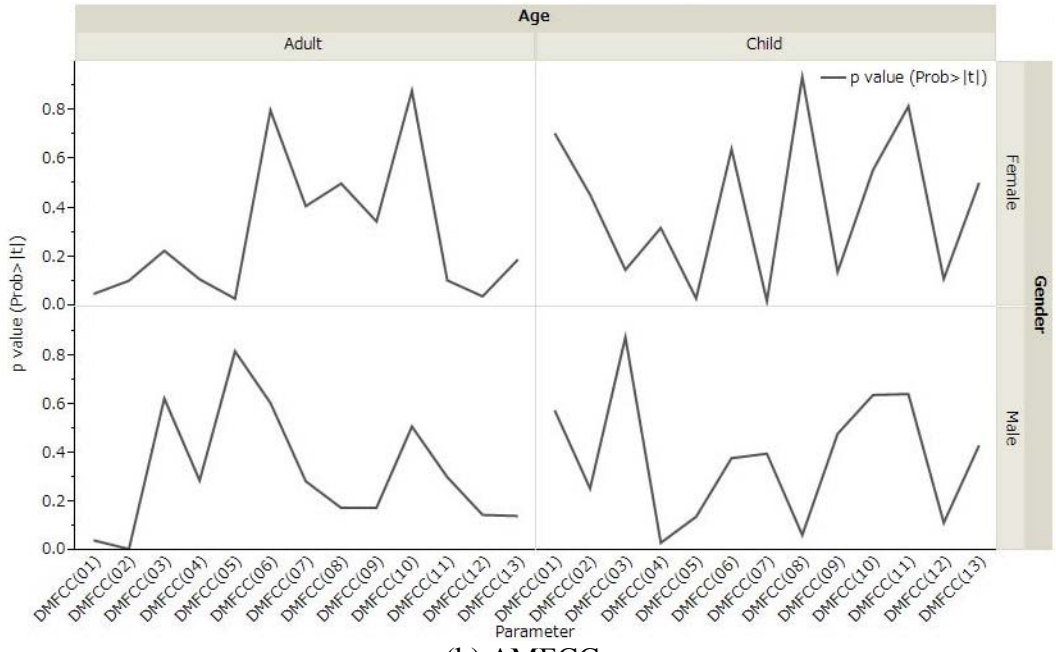
5.2 Experimental results

5.2.1 Significance obtained by one-way ANOVA

Experiments using one-way ANOVA were performed on Datasets 1 and 2. This had little value for acoustic features, but significance and difference are important for speech discrimination. In other words, speech discrimination requires a significant difference between the speeches being examined. Figure 5 shows the distribution of p-values of Datasets 1 and 2. It confirmed that the significance of zero-crossing was $p < 0.2$ and that of MFCC was $p < 0.01$ in the first to the fourth, and sixth to the ninth dimension. On the other hand, significance was not always obtained with ΔMFCC. There was little value to ΔMFCC, since the parameter represents fluctuation and difference in the time dimension, which are canceled out by the calculation of the mean. To obtain the effectiveness of ΔMFCC, the parameter needs to be introduced and used directly, without mean calculation.



(a) Zero-crossing and MFCC.



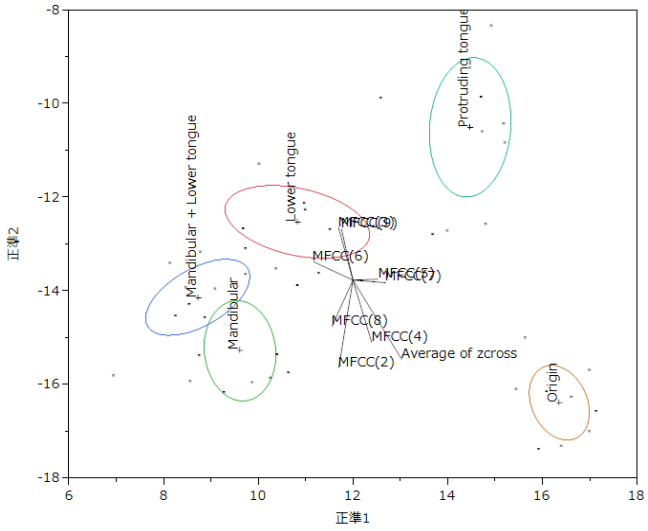
(b) ΔMFCC.

Figure 5: Distribution of p-values across instances of speech.

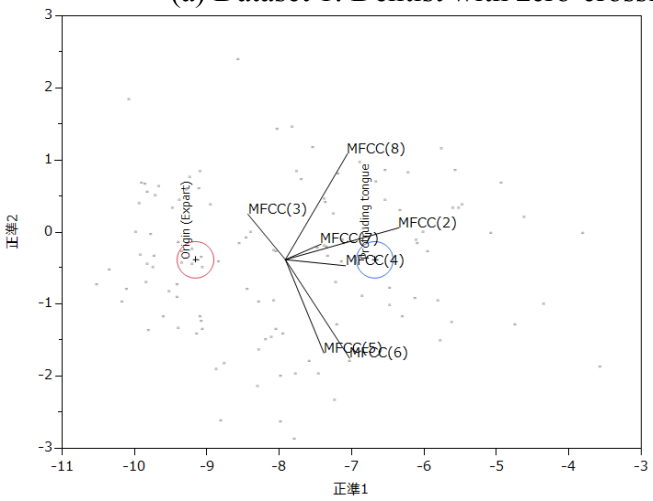
5.2.2 Space separation based on acoustic features for discriminant analysis

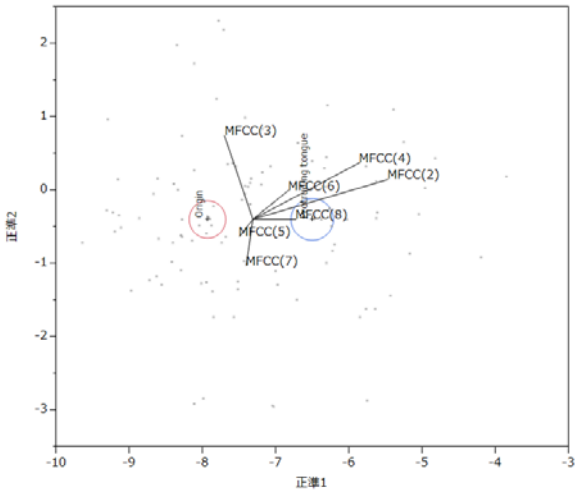
According to the results of Figure 5, the discrimination analysis as space separation focused on MFCC. The experiments employed 2nd- to 8th-order MFCC, since the first dimension contained power information for the sound data. Figure 6 shows the discriminant analysis at the canonical space of the 2nd- to 8th-order MFCC. Figure 6(a) shows the distribution of Dataset 1, and Figures 6(b) and 6(c) show the distributions of Dataset 2. The points in each figure were placed in canonical space, where x-y represents the 1st and 2nd orders of canonical space, the circles the multivariate least-squares mean, and the biplots represent the directions of the covariates in canonical space for each acoustic feature. In both results, it was possible to separate spatial separation in multivariate least-squares means each tongue habit discrimination in Datasets 1 and 2 with MFCC. It was thus confirmed that sufficient discrimination is possible by using seven dimensions of the MFCC, from the second to the eighth dimension, even if conditioned on tongue habit, gender and age.

The experiments confirmed that we can expect five kinds of tongue habits given the gender and age conditions using zero-crossing and MFCC. Although it is possible to mix zero-crossing and MFCC, it is necessary to consider the setting of the scaling factor, etc., since the parameter spaces are different. Therefore, tongue habit discrimination can be realized with conditions where consonants [ʃ] capture tongue habit characteristics using zero-crossing and seven dimensions of the MFCC.



(a) Dataset 1: Dentist with zero-crossing and 2nd- to 9th-order MFCC.





(b) Dataset 2: Adult with 2nd- to 8th-order MFCC. (c) Dataset 2: Child with 2nd- to 8th-order MFCC.

Figure 6: Distribution of discriminant analysis.

6. Conclusion and future work

This research proposed a system for tongue habit discrimination using acoustic feature analysis. To compare the characteristics of protruding and normal speech, Databases 1 and 2 were constructed and compared in terms of acoustic features using zero-crossing and ΔMFCC. The following are the contributions of this study:

- A method for tongue habit discrimination was proposed based on a conventional speech recognition system.
- Datasets 1 and 2 were collected as samples conditioned on tongue habit, gender and age.
- The effectiveness of discrimination of the proposed method was confirmed as significant and

with space separation when acoustic features, zero-crossing and MFCC, were employed excluding ΔMFCC with mean calculation.

- The average value of ΔMFCC failed to obtain a useful level for tongue habit discrimination. However, it was effective for each frame.
- It was confirmed that spatial separation by acoustic features based on each tongue condition is possible by using a seven-dimensional MFCC between the 2nd and 8th orders.
- It was shown that the discrimination can be expected the 7 dimensional MFCC with five kinds of tongue habit conditioned on genders or ages, and then zero-crossing can be performing the consonant detection and its duration period.

In the future work, we will investigate models needed for tongue habit recognition and the identification of acoustic feature for robust discrimination, and implement a system for tongue habit discrimination. We are plans to make the discrimination system available online.

References

1 Hanson, M. L., Oral Myofunctional Therapy, *Am J Orthod*, **73** (1), pp. 59–67, (1978).

2 Ishii, K., Saito, K. and Kasai, K., Clinical Application of Acoustic Analysis in Evaluation of Tongue Function, *Orthod Waves-Jpn*, **71** (3), pp. 170–177, (2012) (in Japanese).

3 Shigemasa, R., Takano, N. and Nakano, Y., The Influence of Orthognathic Surgery on Japanese Consonant Sounds in Cases with Mandibular Retrognathia, *The Japanese Journal of Jaw Deformities*, **25** (4), pp. 241–248, (2015) (in Japanese).

4 Nakagawa, S. and Ohta, K., A Statistical Method of Evaluating Pronunciation Proficiency for Presentation in English, *Proceedings of Interspeech 2007*, pp. 2317–2320, (2007).

5 Advanced Media. AmiVoice SP2, [Online], (2011), Available: http://sp.advanced-media.co.jp

6 Media Navi. ATR CALL Pronunciation LABO, (2017) [Online] http://www.medianavi.co.jp/product/atr-call-challenge/atr-call-challenge.html

7 Nitta, T. and Iribe, Y., Applying Speech Recognition Technology to Pronunciation Training, *Journal of Multimedia Education Research*, **9** (1), pp. S19−S28, (2012) (in Japanese).

8 Saito, M., Ishimitsu, S., Kasai, K., Ishii, K., Nishio, I., Yamashita, K. and Horihata, S., Study for Tongue Position Evaluation of Tongue Habit Voice Using Articulatory Feature Analysis, *Proceedings of ASJ 2015,* Spring meeting, pp. 365–366, (2015) (in Japanese).

9 Saito, M., Ishimtisu, S., Yamanaka, T., Kasai, K., Ishii, K., Nishio, I. and Horihata, S., Study for the Evaluation of Clarity Tongue Protrusion Habit Voice considering Phonological Features, *Proceedings of ASJ 2014,* Spring meeting, pp. 923–924, (2014) (in Japanese).

10 Ishimitsu, S., Nakayama, M., Kasai, K., Horihata, S., Ishii, K. and Yamashita, K., Tongue Position and Tongue Manner Discrimination System, and these Discrimination Methods and Procedures, *Japanese Patent Application*, No. 2016-167180, (2016) (in Japanese).

11 Nakata, K., Recognition and Generation of Speech, *IPSJ Magazine*, **13** (4), pp. 247–255, (1972) (in Japanese).

12 Furui, S, Comparison of Speaker Recognition Methods Using Static Features and Dynamic Features, *IEEE Transactions on Acoustics, Speech and Signal Processing*, **29** (3), pp. 342–350, (1981).

13 Nakagawa, S., A Survey on Automatic Speech Recognition, *The IEICE Transactions on Information and Systems* (Japanese Edition), **J83-D-II** (2), pp. 433–457, (2002) (in Japanese).