Phonetic fluency of Japanese learners of English: automatic vs native and non-native assessment

Mariko Kondo[1,2], *Lionel Fontan*[4], *Maxime Le Coz*[4], *Takayuki Konishi*[3], *Sylvain Detey*[1,2]

[1]SILS, [2]GSICCS, [3]GEC, Waseda University, Japan
[4]Archean LABS, Montauban, France
mkondo@waseda.jp, lfontan@archean.tech, mlecoz@archean.tech,
tkonishi@aoni.waseda.jp, detey@waseda.jp

Abstract

This study compared automatic assessments of L2 phonetic fluency of Japanese learners of English in read speech, with ratings by native and non-native English assessors, and considers whether the first language of the assessors affects the results.

Speech data of 183 Japanese and 25 native English speakers' readings of "the North Wind and the Sun" were assessed for phonetic fluency by 16 trained assessors with different first languages (four American English, four Japanese and eight other languages). They rated segmental accuracy, prosody, fluency and nativelikeness. A subset of 97 of the speakers' data (the 25 native English speakers and 72 randomly selected Japanese speakers) was also used to develop an automatic fluency assessment system. The 97 speakers' data were re-assessed by four different trained American raters. The correlation between the automatic evaluation and the four raters was 0.83. When the automatic system was then tested on the remaining original 111 speakers' data and the original 16 assessors' scores, it showed correlations of 0.62-0.67 for the American, Japanese and other language raters.

The results suggest that the automatic assessment system can assess phonetic fluency of Japanese-accented English quite reliably, and that native and non-native evaluators used different phonetic cues to evaluate fluency.

Index Terms: second language fluency, Japanese accented English, native assessment of speech, non-native assessment of speech, automatic assessment of speech, English learner corpus

1. Introduction

During the process of second language (L2) phonological acquisition, the first language (L1) influences the development of the interlanguage in various ways, and L1 transfer occurs at both the segmental and prosodic levels. Among the different dimensions that L2 learners must master (e.g. segmental accuracy, speech rhythm and intonation) in recent years the concept of nativelikeness has been considered to be less important, with more emphasis on the notion of intelligibility. In that respect, L2 prosody acquisition has also been attracting attention, since prosodic features play an essential role in speech intelligibility, whereas segmental accuracy is more important for correct word recognition. Also, suprasegmental accuracy, such as speech timing, seems to have greater influence than segmental accuracy on native speakers' judgement of L2 speaking proficiency [1], [2]. However, prosodic characteristics in L2 speech have not been studied as much as segmental features [3].

Within the realm of prosodic features, speech fluency is often mentioned as an essential and ultimate achievement for L2 learners, even though its definition may encompass different dimensions depending on the research area. In the field of L2 pronunciation, it may be defined as "the degree to which speech flows easily without pauses and other disfluency markers" [4, p.5]. Therefore, acquiring a native-like (or rather highly intelligible) speech fluency in an L2 is extremely important to ensure proper spoken interaction between L2 users, whether they are native or non-native speakers. However, the evaluation of L2 fluency is not straightforward and it is necessary to take into account issues which differ from those found in L1 fluency evaluation (e.g. potential gaps between segmental and suprasegmental accuracy).

Since human evaluation of phonetic fluency is time consuming, and inter-rater disagreements may hinder its reliability, building a fast and robust automatic L2 phonetic fluency evaluation system may greatly help researchers and practitioners in areas such as L2 education and speech pathology. In the field of natural speech processing, several researchers have tried to develop such a system, usually relying on automatic speech recognition systems [5], [6]. Others have explored methods based on low-level signal segmentation algorithms [7], [8], [9], comparing native human scores and automatic scores. However, very few studies have investigated the differential perception of L2 phonetic fluency by native and non-native assessors, and how this could impact the development of unbiased intelligibility-oriented automatic assessment systems, especially considering the increasing use of some languages as lingua franca, especially English. Therefore, in this paper we describe a four-step study which compares native, non-native and automatic assessment scores of phonetic fluency of read-out speech by Japanese learners of English: (1) description of the original development of the automatic assessment system with Japanese learners of French [7]; (2) development of an ad hoc automatic assessment system for Japanese learners of English [8]; (3) testing the robustness of the system with different sets of data and different raters; and (4) examining the relationship between phonetic features that human assessors use to evaluate Japanese accented English and the acoustic cues that the automatic assessment system is trained to detect.

The automatic L2 phonetic fluency measurement system was originally designed for Japanese learners of French [7]. It

10.21437/SpeechProsody.2020-160

was trained with the data of 8 Japanese students of different proficiency levels, elicited in the reading task of the *InterPhonologie du Français Contemporain* (IPFC) recording protocol [9]. It was subsequently tested with the longitudinal data of the *Corpus Longitudinal Interphonologique de Japonais Apprenants de Français* (CLIJAF) [10] corpus [11]. Each time, four different dimensions were examined, and scored by human assessors on a 5-point scale: global fluency (perceived ease of speech), speech rate (perceived speed of speech), regularity of speech rate (perceived changes in the tempo, i.e. accelerations, decelerations and breaks), and speech fluidity (perceived fluidity of coarticulation, i.e. smoothness of transitions between phones). During the training phase, strong correlations were observed between the human and automatic ratings, with the automatic estimators of speech rate and regularity of speech rate standing out as the two most important predictors of fluency ratings. During the testing phase with the longitudinal data (12 Japanese university students recorded in the same task four times over two years after 4, 7, 12 and 19 months of study) as well as native material (recorded in the same task), speech rate emerged as the main predictor of overall fluency, in tune with other research in the field [5], [12]. The system's predictions reflected the progress made in overall fluency by the learners [13], which motivated us to extend its application to other groups of learners, i.e. Japanese learners of English, in collaboration with the Japanese-Asian English Speech cOrpus Project (J-AESOP) project [14], [15].

2. Assessing L2 phonetic fluency of Japanese learners of English

2.1. Using the J-AESOP corpus to train an ad hoc automatic assessment system

2.1.1. Method of human assessment

We used the speech data of the English version of "the North Wind and the Sun" from the J-AESOP corpus which contains English speech data of 183 native Japanese speakers of varied English levels and 25 native English speakers of varied accents [15]. To build the automatic fluency assessment system, a subset of the J-AESOP data was used, 72 randomly selected Japanese speakers and all 25 native English speakers (Fig 1 Left-side in solid line box) [8]. The text consists of five sentences, so we created a separate file for each recorded sentence uttered by each speaker, yielding 485 sentence files (97 speakers x 5 files). The files were randomized and presented to four native speakers of American English who were trained to assess L2 English speech. They evaluated the fluency level of all the files on a 5-point scale between 0 (not fluent) and 4 (native-level fluency), so that the rating scale matched the Japanese-accented French fluency study described in the introduction (Fig 1a) [7]. Fluency was evaluated separately for each sentence in order to avoid any carryover impression of one sentence to the next sentences uttered by the same speaker, e.g. stuttering, repetition, or overlong pauses.

An average score for the five sentences was obtained for each speaker. The inter-rater correlations of the four raters' fluency scores for the Japanese speakers' data (360 sentences) were between 0.62 and 0.7, but the correlation rose to 0.77-0.83 for the 485 sentences of the combined Japanese and English speakers because almost all English speakers' utterances were rated as "highly fluent".

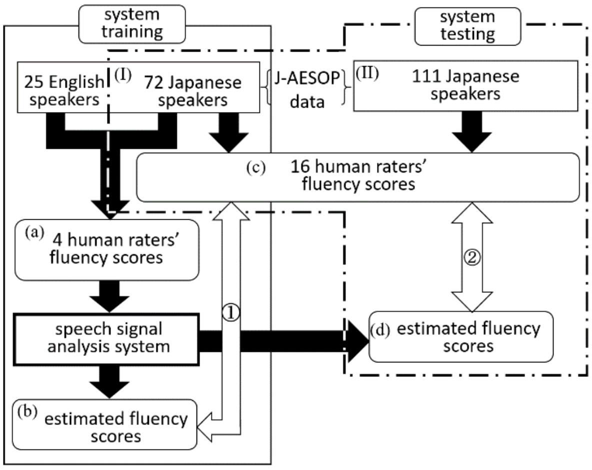
system training
system testing
25 English speakers
(I)
72 Japanese speakers
J-AESOP data
(II)
111 Japanese speakers
(c) 16 human raters' fluency scores
(a) 4 human raters' fluency scores
②
①
speech signal analysis system
(d) estimated fluency scores
(b) estimated fluency scores

Figure 1: *Experimental design*: *the solid line box (left) shows the system training experiments and the dotted line box (right) shows the system testing experiments.*

2.1.2. The speech signal analysis system

This section describes the analysis process leading to the automatic estimation of a fluency score from a speech signal (Fig 1b). Figure 2 shows a flow diagram of the whole process.

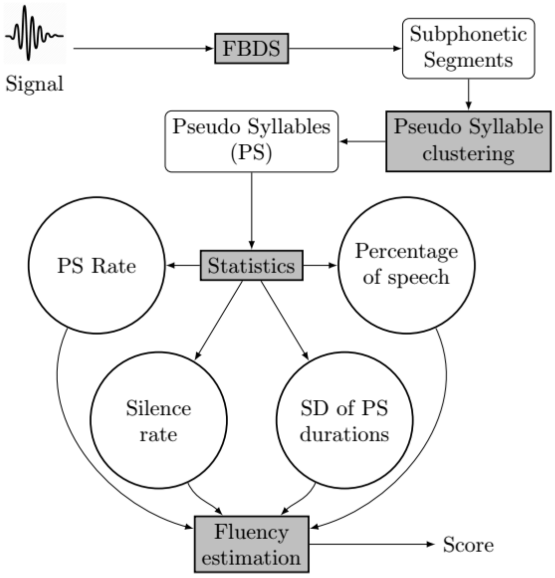

Figure 2: *Flow of data (white boxes and circles) through different processes (gray boxes) used to compute a speech fluency score from a speech signal*

2.1.2.1 Signal segmentation using the Forward-Backward Divergence Segmentation (FBDS).

The analysis starts with the temporal segmentation of the signal using a forward-backward divergence segmentation (FBDS) algorithm [16]. This algorithm has been used in recent studies to automatically extract temporal predictors of speech fluency, both for L2 speech [7] and pathological speech [17]. When applied to speech signals, the FBDS algorithm results in a subphonemic segmentation: phones are generally split into different FBDS-derived segments corresponding to different articulatory phases (e.g. attack, sustain and release of a vowel).

2.1.2.2 Identification of pseudo syllables and silent breaks

The subphonemic FBDS-derived segments are then clustered to form pseudo syllables (PS). This is done by aggregating the FBDS-derived segments according to the evolution of the energy along the speech signal: adjacent FBDS-derived segments are considered to be part of the same PS if, when passing from one to another, the energy does not decrease under a given threshold.

Also, when the energy of an FDBS-derived segment is found to be very low compared with the whole signal energy, the FDBS-derived segment is considered as silent. Adjacent silent FBDS-segments are considered as parts of the same break.

This automatic process leads to a higher-level segmentation of the speech signal into PS and silent breaks.

2.1.2.3 Computation of statistics

Four outcome measures are computed on the basis of the PS and silent breaks segmentation:

- PS rate, as the number of PS divided by the phonation duration (in seconds). This measure is assumed to be highly correlated with the perceived rate of speech;
- Standard deviation of PS duration. This measure is expected to be influenced by the presence of filled pauses and hesitations in speech: the more variable the PS duration, the more likely the presence of such disfluencies;
- Rate of silent breaks, as the number of silent breaks exceeding 200ms divided by the total number of words in the recording. This measure is assumed to be negatively correlated with perceived speech fluency;
- The percentage of speech, as the duration of all PS, divided by the total duration of the speech signal, in seconds. This measure is expected to be positively correlated with perceived speech fluency.

2.1.2.4 Estimation of perceived speech fluency

Finally, the model combines the four predictors of speech fluency into a single score. This model was created by computing a step-by-step linear regression over the average ratings of speech fluency for the 485 sentences pronounced by the 97 speakers of the training corpus (72 Japanese speakers and 25 native English speakers) (Fig 1(I)). The average human ratings were used as the dependent variable and the four automatic predictors as the independent variable.

The best regression (R = 0.73) used the four automatic predictors. An analysis of the normalized coefficients showed that the PS rate explained 42.3% of the variance, compared with 28.1% for the rate of silent breaks, 15.0% for the percentage of speech, and 14.6% for the standard deviation of PS duration.

When averaging the automatic scores and the human ratings as a function of the speaker (N = 97), the model achieved a correlation of 0.83 between predicted and automatic ratings of speech fluency (Fig 1①) [8].

2.2. Using the J-AESOP corpus to test the system: comparing automatic evaluations with native and non-native human evaluations

2.2.1. Method

We tested the reliability of our automatic fluency evaluation system by comparing it with human evaluation scores of the remaining data of J-AESOP corpus (Fig 1 Right-side in dotted box). The J-AESOP corpus has human evaluation scores on fluency, segmental accuracy, prosody, and nativelikeness based on the speakers' performance of "the North Wind and the Sun". The read speech of the entire prose (five sentences) was divided into three files of approximately equal durations (two sentences, one sentence and two sentences, respectively) so that evaluations could be conducted without any carryover impression from one speech file to the next file by the same speaker. Sixteen trained raters with ten different first languages assessed the speech fluency of each speech file (208 speakers x 3 files) on a scale from 1 (not fluent) to 10 (native-level fluency), considering aspects such as intonation, speech rate, pausing and stumbling. The first languages of the 16 raters were American English (4 raters, different from those in section 2.1.1), Japanese (4 raters) and 8 other languages (OT) (Cantonese, French, German, Korean, Mandarin Chinese, Polish, Punjabi, and Spanish) (Fig 1c).

First, the automatically calculated predicted fluency scores of the 72 Japanese speakers' data were compared with the 16 human raters' scores. Then, the predicted fluency scores of the speech data of the other 111 Japanese speakers in the J-AESOP corpus (Fig 1(II)) were calculated using the method explained in section 2.1.2 (Fig 1d). The predicted scores of these 111 speakers were also compared with the human evaluation scores (Fig 1②).

2.2.2. Results

The correlation coefficients of the automatically predicted and human evaluated speech fluency scores of the original 72 Japanese speakers are shown in Table 1. The correlation coefficients were fairly high: ranging from 0.82 for the American English raters to 0.72 for the Japanese speaking raters. The overall correlation for all the raters was 0.79. In addition, Pearson's product-moment correlation showed all four correlations were significant (*p<.001*).

Table 1: *Pearson's product-moment correlation coefficients (r) of predicted perceived fluency rating cores and human rating scores of 72 Japanese speakers' English by each rater's first language (*** p<.001)*

Rater groups	*r*
English (American)	0.82***
Japanese	0.72***
Other languages (OT)	0.77***
All raters	**0.79*** **

Next, the correlation coefficients of the automatically predicted and human evaluated speech fluency scores of the other 111 J-AESOP Japanese speakers are shown in Table 2. These correlations ranged from 0.67 for the American English raters to 0.62 for the Japanese raters. The overall correlation for all the raters was 0.65. Pearson's product-moment correlation showed all four correlations were significant (*p<.001*).

Table 2: *Pearson's product-moment correlation coefficients (r) of predicted perceived fluency rating scores and human rating scores of 111 Japanese speakers' English by rater first language (*** p<.001)*

Raters groups	***r***
English (American)	0.67***
Japanese	0.62***
Other languages (OT)	0.63***
All raters	**0.65*****

The results suggest that the automatic evaluation system predicted the evaluation scores of the model subset (N = 72) of the corpus fairly successfully, but the prediction of the scores of the other set (N = 111) of the corpus was a little less accurate. However, Pearson's product-moment correlation showed that these four correlations were still also significant (*p<.001*).

The equality of the system training (Table 1) and testing (Table 2) correlation coefficients were tested by rater L1groups. A Fisher *r*-to-*z* transform [18] was used to check for significant differences in the correlations obtained for evaluations by English native speakers, Japanese speakers and other language speakers. There were significant differences between the three groups for the system training (Table 3a), but for system testing the only significant difference was between the English and Other language groups (Table 3b).

Table 3: *Equality test of correlation coefficients of predicted perceived fluency rating scores in system training and testing by raters' L1 groups (1-tailed)*

Rater groups	***(a) system training***		***(b) system testing***	
	z-score	***p-value***	***z-score***	***p-value***
Eng vs. Jp	3.642	0.0001	1.56	0.059
Eng vs. OT	2.659	0.0039	1.953	0.025
Jp vs. OT	-2.418	0.0078	-0.168	0.433

3. Discussion

Both experiments showed that the automatic evaluation system could assess phonetic fluency of Japanese-accented English quite reliably. In the first experiment, after the automatic assessment system was trained with the 72 Japanese and 25 English speakers' data, it was tested with the same speech data scored by 16 different human raters (Figure 1c). The automatically predicted fluency scores of the 72 Japanese speakers were reliable with very high correlation scores. Then, in the second experiment the automatic system was tested on 111 new speakers' data and the predicted fluency scores were compared against the evaluation scores by the same 16 raters. The correlation coefficients between the predicted scores and human assessment scores were not as high as the results for the original 72 speakers, but the automated and human rater scores were significantly correlated (*p<.001*). These results show the potential for the automatic fluency evaluation system to be used to assess fluency of non-native speakers' English speech.

There was some effect of the raters' L1 on fluency score. There were significant differences between L1 groups of the 16 raters in the correlation coefficients of the data of the 72 speakers used to originally train the system (Table 3a). This result is probably to be expected because the system was trained using only American raters. On the other hand, when the system was tested with the other 111 speakers' data, there was a large significant difference in the correlation coefficients between the English and OT rater groups, but not between the JP and OT rater groups. The difference between the correlation coefficients for the English and Japanese raters was not significant but the p-value was quite close to 0.05. Therefore, these results suggest that English raters tend to assess differently from non-native English raters. However, when the fluency scores of the same American and Japanese raters had been compared previously [19], looking at the fluency ratings of "the North Wind and the Sun" speech data of all 183 Japanese and 25 native English speakers, the native and non-native raters gave similar ratings for fluency.

Therefore, given that there were differences in the current study between the native and non-native rater correlations with the automatic evaluation system ratings, it suggests that the native and non-native raters used different cues to evaluate fluency. It also suggests that the fluency cues used in the automatic evaluation system, such as silent breaks, speech rate and pseudo syllables may not be important fluency features for non-native listeners.

It must be noted, however, that in the previous study [19] one pronunciation error which both native and non-native raters marked down was vowel epenthesis. Epenthetic vowels change syllable structure and so affect speech rhythm. The automatic speech signal analysis system assessed in the current study uses pseudo syllables and silent breaks as part of the fluency rating system. Vowel epenthesis is likely to alter pseudo syllables and silent breaks. Therefore, the good correlations between the human and automatic ratings suggest that the automatic system was able to match well the human raters' sensitivity to vowel epenthesis pronunciation errors and how they affect fluency.

These results suggest that fluency is a feature which can be automatically evaluated fairly accurately. So, next we need to validate the results with different speech data, and also assess further the phonetic features that affect fluency evaluation in relation to rater L1, and how phonetic training affects evaluation.

4. Conclusions

This study tested the automatic assessments of Japanese learners' L2 English phonetic fluency in read speech by comparing ratings by native and non-native English raters.

The results showed that (1) there was a correlation between the automatic and human fluency evaluation scores in both training and testing L2 corpora; (2) the automatic system's phonetic features like pseudo syllable rate, standard deviation of pseudo syllable duration, and silent breaks could reflect the human raters' sensitivity to fluency; and (3) native and non-native evaluators used different phonetic cues to evaluate fluency.

5. Acknowledgements

This research was supported by the Japanese Society for the Promotion of Science, Grant-in-Aid (Challenging Exploratory Research) No. 19K21638 to the first author and (B) No. 23320121 and No.15H03227 to the last author. We would like to thank Xavier Aumont of Archean Technologies for his support for this work, and the audience at New Sounds 2019 for their useful comments about Fontan et al. (2019).

6. References

[1] J. Anderson-Hsieh, R. Johnson, R., and K. Koehler, “The relationship between native speaker judgments of non-native pronunciation and deviance in segmentals, prosody, and syllable structure,” *Language Learning*, vol. 42, no. 4, pp. 529-555, 1992.

[2] U. Gut, “Non-native speech rhythm in German”. *Proceedings of 15th International Congress of Phonetic Sciences*, Barcelona, Spain, 3-9 August 2003, pp. 2437-2440.

[3] M. Jilka, “Different Manifestations and Perceptions of Foreign Accent in Intonation”. In J. Trouvain, and U. Gut (Eds.), *Non-native Prosody: Phonetic Description and Teaching Practice*, pp. 77-96. Berlin: Mouton de Gruyter, 2007.

[4] M. Derwing and M. J. Munro, *Pronunciation Fundamentals. Evidence-based Perspective for L2 Teaching and Research.* Amsterdam. Netherlands: John Benjamins, 2015.

[5] C. Cucchiarini, H. Strik, and L. Boves, “Quantitative assessment of second language learners’ fluency by means of automatic speech recognition technology,” *The Journal of the Acoustical Society of America*, vol. 107, no. 2, pp. 989–999, 2000.

[6] ——, “Quantitative assessment of second language learners’ fluency: Comparisons between read and spontaneous speech,” *The Journal of the Acoustical Society of America*, vol. 111, no. 6, pp. 2862–2873, 2002.

[7] L. Fontan, M. Le Coz, and S. Detey, “Automatically measuring L2 speech fluency without the need of ASR: A proof-of-concept study with Japanese learners of French,” in *INTERSPEECH 2018 — 19th Annual Conference of the International Speech Communication Association, September 2-6, Hyderabad, India, Proceedings*, 2018, pp. 2544-2548.

[8] L. Fontan, M. Le Coz, and M. Kondo. “Building an ASR-free automatic tool for measuring the speech fluency of Japanese learners of English,” presented at NEW SOUNDS 2019 - 9th International Symposium on the Acquisition of Second Language Speech, August 30th-September 1st, Tokyo, Japan, 2019.

[9] S. Detey, I. Racine, Y. Kawaguchi, and F. Zay, “Variation among non-native speakers: the InterPhonology of Contemporary French,” in *Varieties of Spoken French*, S. Detey, J. Durand, B. Laks, and C. Lyche, (Eds.) Oxford, U.K.: Oxford University Press, 2016, pp. 491–502.

[10] S. Detey and I. Racine, “Towards a perceptually-assessed corpus of non-native French: the InterPhonology of Contemporary French (IPFC) project illustrated with a longitudinal study of Japanese learners’ /b-v/ production,” *International Journal of Learner Corpus Research*, vol. 3, no. 2, pp. 223–249, 2017.

[11] S. Detey, “CLIJAF: corpus longitudinal interphonologique de Japonais apprenants de français”. Projects Kakenhi (B) no. 23320121 & no. 15H03227, Japanese Society for the Promotion of Science, Tech. Rep., 2011-2019.

[12] A. Ginther, S. Dimova, and R. Yang, “Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring,” *Language Testing*, vol. 27, no. 3, pp. 379–399, 2010.

[13] S. Detey, L. Fontan, M. Le Coz and S. Jmel, “Computer-assisted assessment of phonetic fluency in a second language: a longitudinal study of Japanese learners of French,” (under revision).

[14] H. Meng, C. Tseng, M. Kondo, A. Harrison, and T. Visceglia, “Studying L2 Suprasegmental Features in Asian Englishes: A Position Paper,” in *INTERSPEECH 2009 — 10th Annual Conference of the International Speech Communication Association, September 6-10, Brighton, U.K., Proceedings*, 2009, pp. 1715-1718.

[15] M. Kondo, H. Tsubaki, and Y. Sagisaka, “Segmental Variation of Japanese Speakers’ English: Analysis of “the North Wind and the Sun,” in AESOP Corpus”, *Journal of the Phonetic Society of Japan*, vol. 19, no.1, pp. 3-17, 2015.

[16] R. André-Obrecht, “A new statistical approach for the automatic segmentation of continuous speech signals,” *IEEE Transactions on Audio, Speech, and Signal Processing*, vol. 36, no. 1, pp. 29–40, 1988.

[17] D. Sztahó and I. Valálik, “Speech Fluency Measurement of Patients with Parkinson's Disease by Forward-Backward Divergence Segmentation,” in *10th IEEE International Conference on Cognitive Infocommunications. Naples, Italy*, 2019.

[18] Lee, I. A. and K. J. Preacher, Calculation for the test of the difference between two dependent correlations with one variable in common [Computer software]. Available from http://quantpsy.org., 2013.

[19] M. Kondo and T. Konishi, “Tsuujiru Eigo no tameno hatsuon kyouiku (What we need to teach for better communication in English),” (in Japanese), *Conference Handbook of the 35th Conference of the English Linguistic Society of Japan*, pp. 232-237, Tohoku University, Sendai, Japan, 18-19 November, 2017.