2018 RISP International Workshop on Nonlinear Circuits,
Communications and Signal Processing (NCSP2018)
Honolulu, Hawaii, USA, March 4-7, 2018

Perceptual grouping with prosodic features in Japanese dialects

Ling Zhang and Masato Akagi

School of Advanced Science and Technology,
Japan Advanced Institute of Science and Technology
1–1 Asahidai, Nomi, Ishikawa, 923–1292, Japan
E-mail: {momo233, akagi}@jaist.ac.jp

Abstract

Previous studies found that prosodic perception in the dialects affects listeners' impression. However, relationships and differences among the dialects have not discussed yet using their impressions. In this paper, perceptual distances of stimuli preserving prosodic information of dialects were discussed in a 3-dimension space constructed by using Multi-Dimensional Scaling (MDS). 3 physical features were superimposed using Multiple Regression Analysis (MRA) into the space. As a result, it is indicated that the factors among different perceptual distances of prosodies were almost depended on the difference of F0. The dialects can be grouped in 2 parts with the prosodic features of them.

1. Introduction

In daily conversation, the same spoken content generally conveys different impression by incorporating different dialect prosodies. It has been found that the prosodic differences among some Japanese dialects affect their impressions on listeners even if they do not know the dialect [1, 2]. It means that the prosody is important in human communication. If reasons why prosodies affect human's perception of dialects can be found out, it is thought that this study will lead to solution of the perception mechanism about prosody.

Several researches focusing on speech production revealed that Japanese dialects have different prosodic features each other. Kubozono found that Tokyo dialect and Kagoshima dialect are different in their accent characteristics [3]. Moriyama also found that Tokyo dialect and Osaka dialect are different in their accent position and mora-characteristics of vowels [4].

Prosodic characteristics of each dialect have been investigated also by a perspective of speech perception. Nagase found that Japanese dialects are perceptually divided into 2 types; the dialects with distinct images (e.g. Osaka, Tokyo) and those with vague images (e.g. Kagoshima, Hakata), using subjective evaluation [2]. Another subjective investigation by Chan also found that even beginning Japanese learners have different impressions for each dialect [1]. These findings indicated that prosodic perception in the dialects affects their impression.

However, the reason for the effect of the prosody on the impression of the dialects has not been explained quantitatively yet. Moreover, relationships and differences among the dialects have not discussed yet using their impressions.

This study aims to investigate relationships and differences among Japanese dialects and to group them based on their perceptual impressions.

In order to focus on the perception of dialect prosodies, it is necessary to prepare stimuli sounds which contain only the prosodic information of the dialects. In this study, the sounds are resynthesized from dialect speech sounds using STRAIGHT [5]. Then, the perceptual investigation and grouping of the dialects are conducted using multidimensional scaling (MDS) and multiple regression analysis in reference to a related study [6].

2. Resynthesis of prosodic stimuli

2.1 Database

Dialect speech sounds recorded in "Japan dialect speech material database (in Japanese)" from National Institute for Japanese Language and Linguistics were used. This database contains conversations by several old people. It consists of 47 kinds of dialects in total. it is thought to be effective to analyze the natural speech perception since the recorded sounds are spontaneous speeches.

In this study, the typical Japanese dialects, Aomori, Tokyo, Osaka, Fukuoka and Kagoshima dialect, were chosen. Then three sentences with duration of 4 to 7 seconds were cut out from one dialect. To eliminate their linguistic informations and preserve only their prosodic information , acoustical features which contribute to prosodic information, such as fundamental frequency (F0) and power fluctuation, speech rate and pause should be focused [8]. In this paper, F0 and power fluctuation were focused on. Then the prosodic features of dialect speeches were replaced with that of vowel /a/.

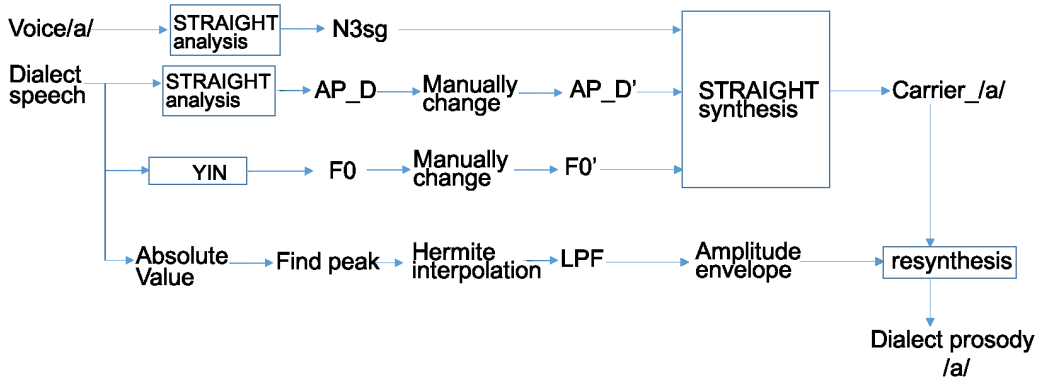

Figure 1: Schematic illustration of stimulus resynthesis

2.2 F0 and power extraction

In this study, F0 and power fluctuation were used for resynthesis.

Firstly, an utterance of vowel /a/ by a male speaker was recorded and it's duration was adjusted to the duration of dialect speech sounds.

Secondly, F0 in each dialect speech were extracted using YIN [7]. Then spectral envelope (n3sgram) and aperiodicity index (ap) in each dialect speech were also extracted using STRAIGHT [5].

Thirdly, F0, n3sgram and ap in the recorded /a/ sound were replaced with those extracted from each dialect speech.

Finally, all peaks in the absolutized waveform of each dialect speech were connected using Hermite interpolation method. The envelope from the connected peaks was low-pass-filtered. Using this envelope, one stimulus with the same temporal fluctuation as each dialect speech was resynthesized.

Figure 1 shows a schematic illustration of the resynthesis processing. In accordance to this processing, prosodic features in 3 sentences from each dialect were extracted and totally 15 stimuli were resynthesized.

2.3 Evaluation of prosody

In order to evaluate whether the stimuli have the same prosodies as the original ones, two listening experiments (experiment 1 and 2) were carried out. Experiment 1 is to ask participants to identify the prosodies among 5 dialects. Experiment 2 is to ask participants to identify the prosodies within a certain dialect.

2.3.1 Participants

Ten Japanese graduate students, aged 22 to 26, participated in the experiments. All participants had normal hearing and none had speaking disorders.

2.3.2 Apparatus

Both experiments were conducted in a soundproof room. The experimental stimuli were presented to the participants by a headphone (STAX SR-404), through an audio interface (FIREFACE UCX). The sound pressure level of all stimuli ranged from 66 to 69 dB. The sampling frequency was 44,100 Hz and the number of quantizing bits was 16.

2.3.3 Procedure

In the experiment 1, one dialect speech and 5 stimuli were set as a group for evaluation. The 5 stimuli consisted of one stimulus which was synthesized from the same dialect speech in the group, and the 4 other stimuli which were synthesized from different dialects. The stimuli were chosen as the similar length as possible.

In the experiment 2, one dialect speech and 3 stimuli were set as a group for evaluation. the 3 stimuli consisted of one stimulus which was synthesized from the same dialect speech in the group, and the 2 other stimuli which were synthesized from the different sentences in the same dialect.

Participants were asked to judge which stimulus is synthesized from the original dialect speech after listening to all sounds in one group.

2.3.4 Result

As results of both experiments, more than 70 % of the judgments were correct in almost all dialects. This indicates that the stimuli have the same prosodic information with the original dialect speech.

Table 1: Directional relationships and R of the physical feature

Physical Feature		Degree [°]	Physical Feature	R
F0 range	F0 speed range	36.9	F0 range	0.96
	F0 rhythm	52.8	F0 speed range	0.67
F0 speed range	F0 rhythm	26.9	F0 rhythm	0.72

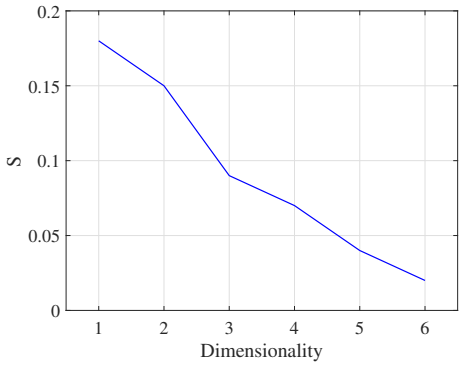

Figure 2: Relationship between stress value S and the dimensionality

3. Perceptual grouping

3.1 Experiment of similarity evaluation

In order to investigate prosodic impressions of the stimuli on the participants' perception, the following experiment (experiment 3) for evaluating the similarity among the prosodies was conducted.

3.1.1 Participants

Ten Japanese graduate students, aged 22 to 26, participated in the experiments. All participants had normal hearing and none had speaking disorders.

3.1.2 Apparatus

The apparatus was the same as those used in experiment 1 and 2.

3.1.3 Procedure

With each of the 15 dialect stimuli as a pair, 225 pairs of stimuli were prepared. Each pair of stimuli was presented followed by a gap of one second. At first, several exercise trials were conducted to accustom participants to evaluate the prosodic similarity between the pair. In this pertrial, the original speech sounds were used as stimuli. The participants were asked to evaluate each pair of speech sounds immediately on a 5-point graded scale (from -2 to 2, including 0, -2 = different, 2 = similar) every time. After that, actual trials were conducted. In this pertrial, the resynthesized stimuli were used. The participants were asked to evaluate each stimulus pair immediately on the same scale as stated above. Then, Scheffe's method of paired comparison was applied for analysis.

3.2 Building perceptual space using MDS

The data from participants' evaluations was converted into matrix which can express a dissimilarity degree or perceptual distance. The SPSS 24.0J for windows MDS ALSCAL procedure, non-metric model of Kruskal was applied using the matrix to analyze the evaluation. The relationship between stress value (S) and dimensionality was shown in Fig. 2. From S in Fig. 2, the distribution of dialect stimuli should be located in the three- dimensional perceptual distance space.

3.3 Interpretation to the distance space using MRA

To find physical features that can explain the perceptual space derived from MDS and to find how the physical features relate to each dialect stimuli, the multiple regression analysis (MRA) was used. The MRA is one statistic technique when an explanation variable is needed to predict from plural purpose variables. A physical feature Y is expressed by an explanation function of perceptual distance X_n ($n = 1, 2, 3$) of each dialect stimuli derived from MDS using the regression model as following equation:

$$Y = a_1 X_1 + a_2 X_2 + a_3 X_3 + b \quad (1)$$

where a_1, a_2 and a_3 are the partial regression coefficients and b is a constant value.

The partial regression coefficient is used to give meanings to the direction in the perceptual distance space. In this paper, as the physical features, mean value of power, range of fluctuation speed of power and F0, range of F0, fluctuation rhythm (calculate from spectrum) of power and F0 are discussed. F0 values were transformed into mel scale to consider human's pitch perception. They are analyzed by multiple regression in SPSS 24.0J for windows and then partial regression coefficients were calculated. From the values of the partial regression coefficients, directions in 3-dimension space and correlation coefficients of regression R are calculated. Table 1 shows the degrees of each pair of physical features superimposed in 3-dimensional space and R. Range of F0, range of the speed of F0, change rhythm of F0 were used as new axes to give meanings to the direction plotted in dimension-1 against dimension-2, dimension-1 against dimension-3 of

the 3-dimension space from MDS. The locations of stimuli plotted in dimension-2 against dimension-3 are inconsistent, therefore the plane of dimension-2 against dimension-3 is not shown in this paper. The direction of the arrowhead of each axis indicates that the values of physical features increases. They are shown in figure 3.

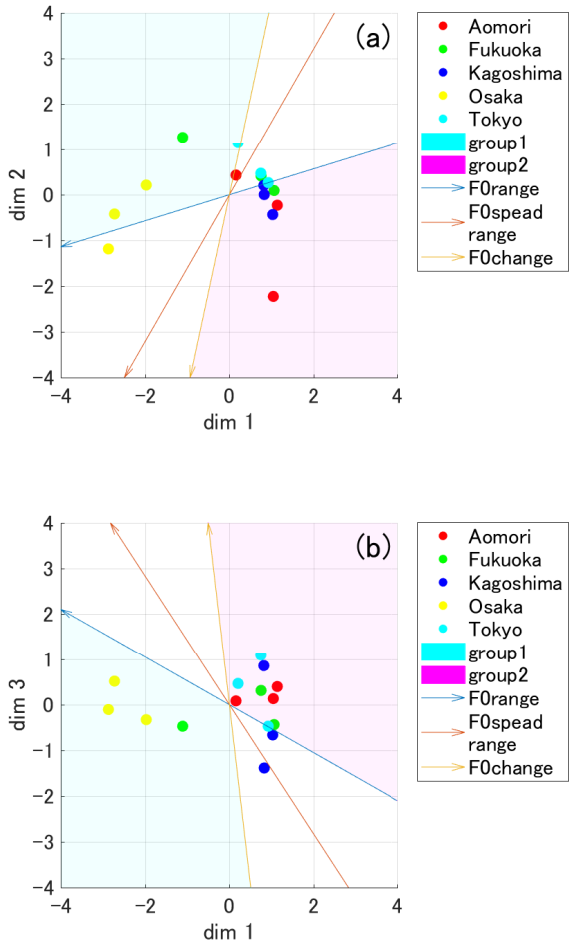

Figure 3: Directions of physical features in perceptual space, (a) dim-1 and dim-2, (b) dim-1 and dim-3, (c) dim-2 and dim-3

4. General Discussion

From the results above, it is suggested that the three physical features, range of F0, range of F0 speed and F0 fluctuation rhythm, can explain the 3-dimensional space. From the degrees and values of R in table 1, it is indicated that range of F0 and F0 fluctuation rhythm (range of the speed of F0) give two explanations to the space. Depending on the two physical features superimposed in each pair of dimensions, it is assumed that the dialects can be grouped in 2 parts which was shown in the shadows of figure 3. Group 1 is the dialects with wide ranges of F0 and relatively quick F0 fluctuation rhythm (or wide range of the speed of F0). Group 2 is the dialects with narrow ranges of F0 and relatively slow F0 fluctuation rhythm (or narrow range of the speed of F0).

5. Conclusion

In this paper, the perceptual distances of stimuli synthesized from the prosodic information of dialects were calculated in a 3-dimension space by using MDS. Then 3 physical features were superimposed by the application of a MRA into the space. As a conclusion, it is indicated that the factors among different perceptual distances of prosodies were almost depended on the difference of F0. Then the dialects can be grouped in 2 parts with the prosodic features of them.

References

[1] Y. Chan, "Study on impressions of dialects on Japanese lerners," *Prog. & Rep. for Japanese lang. & cultural study*, **28**, 70–101, 2013. (In Japanese)

[2] J. Nagase, "Formation of dialect image," *J. Senshu Japanese Lang. & Liter.*, **96**, 1–18, 2005. (In Japanese)

[3] H. Kubozono, "Accent variation of Kagoshima dialect: the breaking of composite rule," *Kobe Papers in Linguistics*, **5**, 111–123, 2005. (In Japanese)

[4] T. Moriyama, H. Ogawa and S. Tenpaku, "A method to control fundamental frequency to generate utterances of Osaka dialect," *Tech. Rep. of IEICE*, SP98-81, 1998.

[5] H. Kawahara, I. M.-Katsuse and A. de Cheveigné, "Restucturing speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, **27**, 187–207, 1999.

[6] K. Ohgushi, "Application of multidimensional scaling to perceptual analysis of sounds," *J. Acoust. Soc. Jpn.*, **67**(11) 557–558, 2011.

[7] A. de Cheveigne and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, **144**(4), 1917–1930, 2002.

[8] K. Hirose, ""Prosody and speech processing," Scientific Research of Priority Areas," *Tech. Rep. of IPSJ-SLP*, **124**, 299–302, 2003.