Building a Kansai accent dictionary using YouTube

Hiroto Noguchi (Sophia University/Tokyo Medical and Dental University)

YouTube を利用した関西方言アクセント辞書の作成

野口大斗（上智大学／東京医科歯科大学）[†]

Abstract

This paper introduces an effort to create an accent dictionary of the Osaka dialect using user-generated content on YouTube. Speech is extracted from videos, transcribed, and then forced alignment is performed on the transcribed speech. The pitch of each segment is measured, and the pitch is automatically encoded. This paper reports the results of a preliminary application of the process to a single video.

1. Introduction

There has been a remarkable development in language processing, such as TTS and language models. However, for languages with a small number of speakers, efforts to develop language resources have lagged behind. The same is true within the Japanese language. While there is a certain amount of linguistic resources for standard Japanese, this is not the case for dialects. In the case of accents, there are accent dictionaries for the Tokyo dialect, such as NHK (1998) and open-source ones (Tachibana & Katayama, 2020). On the other hand, there are accent dictionaries for Osaka dialects (Sugito, 1995), but I cannot find any large open-source ones.

With that said, it is challenging to have tens of thousands of words read by several speakers unless a researcher is blessed with informants and ample research funds. This paper presents an approach to the automatic generation of accent dictionaries using data from YouTube, a user-generated content platform.

2. Previous Studies

Sugito (1995) interviewed three speakers of each of the two generations for their accent patterns. As shown in (1), the pitch of each word is indicated by L and H. Unlike Tokyo Japanese, the Osaka dialect differs from Tokyo Japanese in whether the pitch starts high or low, depending on the word.

(1)

a. kodomo	‘child’	HHH
b. i’noti	‘life’	HLL
c. kimi’ra	‘you’	HHL
d. Suzume	‘sparrow’	LLL
e. hata’ke	‘field’	LHL

The creation of accent dictionaries has so far been mainly based on fieldwork, using written, speech-listening, and recording methods. While the accents of words needed for headwords can be obtained efficiently, the problems lie in the fact that many of them are based on lists that are not natural

[†] noguchih425@gmail.com

speech and that they are a time-consuming process.

3. Methods

3.1 Data

To eliminate the burden on the people surveyed and to obtain natural speech, data from YouTube was used. In selecting YouTubers, several were chosen based on the following conditions: the YouTuber must be from the Kansai region, must be speaking alone, and must have no background music. A video of one of them was used in this study.

3.2 Procedures

The audio was separated from the video, and the utterances were transcribed using Whisper's (Radford et al., 2023) large model. For each utterance, a text file with the same name except for the file extension was prepared with the utterance content written in *hiragana*, and segmentation was performed. The transcribed text was converted into hiragana using mecab-ipadic-NEologd (Toshinori, 2015). If a morphological analyzer supporting new words is not used, the new words will not be recognized correctly. An example of segmentation results is shown in Figure 1.

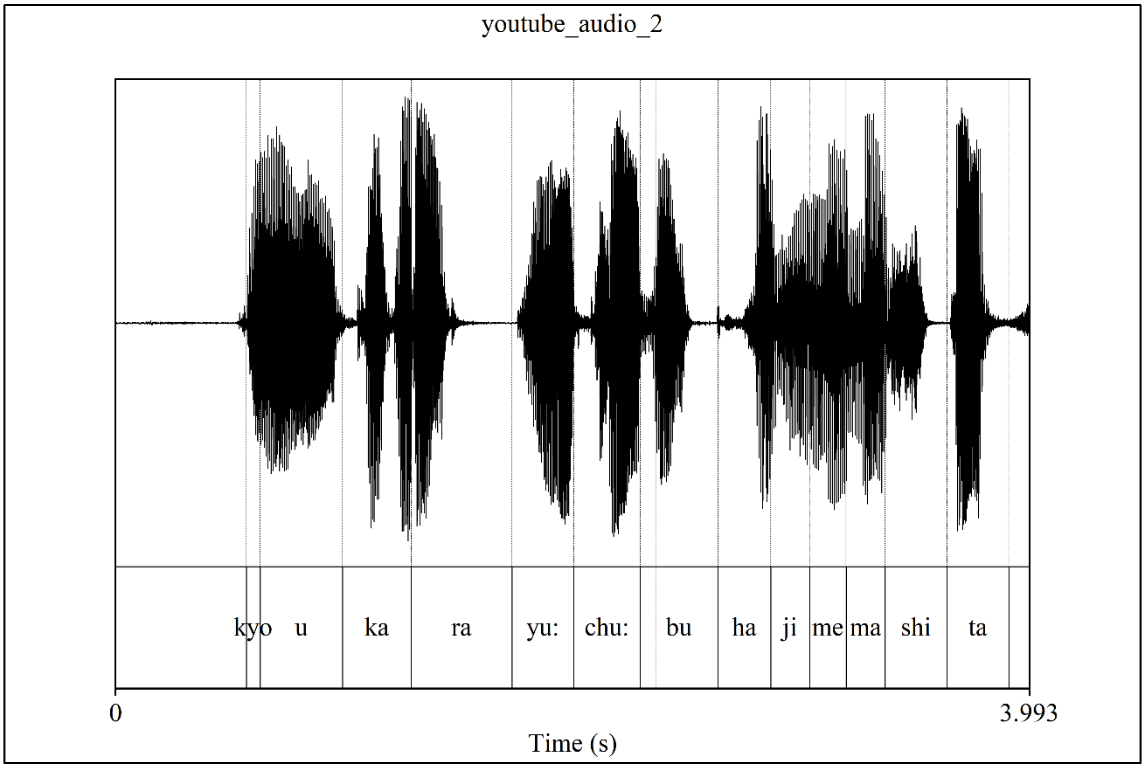

Figure 1 An example of segmentation.

For each segment, the average pitch was measured in Hertz and assigned an H and L based on whether it was higher or lower than the average for each word. Word alignment was performed using mora counts of each word obtained during morphological analysis.

4. Results

For each segment in (2), the pitch was measured as in (3). To avoid downstep effects, the average pitch for each word was compared to the segment, resulting in the pitch pattern in (4). Although there are some differences from the auditory impression, the pitch curve in Figure 2 can be transcribed to the symbols.

(2) ['kyo', 'u', 'ka', 'ra', 'yu:', 'chu:', 'bu', 'ha', 'ji', 'me', 'ma', 'shi', 'ta']
(I started my YouTube channel today.)

(3) [123.8923, 132.7712, 153.3251, 146.7026, 155.5783, 206.0359, 188.1586, 122.4416, 116.2355, 116.8804, 110.4808, 108.3529, 139.1947]

(4)

a. ['kyo', 'u']	['L', 'H']	“today”
b. ['ka', 'ra']	['H', 'L']	“from”
c. ['yu:', 'chu:', 'bu']	['L', 'H', 'H']	“YouTube”
d. ['ha', 'ji', 'me']	['H', 'L', 'L']	“start”
e. ['ma', 'shi']	['H', 'L']	(POLITE)
f. ['ta']	['L']	(PAST)

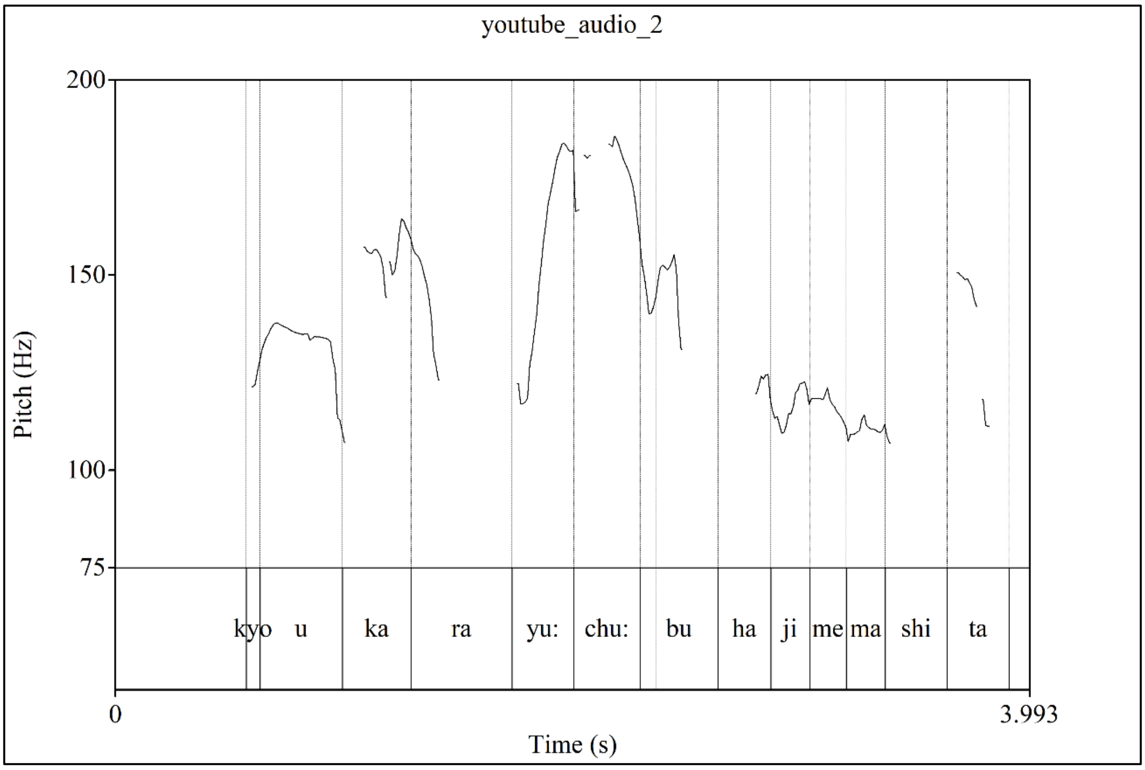

Figure 2 The pitch contour of (2).

5. Discussion

At first glance, such a pitch transcription may not seem appropriate. However, some words appear more than once. For example, (5) shows which accent pattern the word YouTube was identified within the video. The most frequently pronounced pattern, (5)a, which was pronounced ten times, is also consistent with the auditory impression.

(5)

```
['yu:', 'chu:', 'bu'] ['H', 'H', 'L'] * 10
['yu:', 'chu:', 'bu'] ['L', 'H', 'L'] * 2
['yu:', 'chu:', 'bu'] ['H', 'L', 'L'] * 1
['yu:', 'chu:', 'bu'] ['L', 'H', 'H'] * 1
```

By increasing the number of videos, it is possible to obtain a more reliable accent pattern for a larger number of words. In this 16-minute video alone, 474 types and 918 tokens appeared in terms of word count. Further investigation is needed in the future.

Acknowledgements

I would like to thank the YouTube Research Program for permission to use the data used in this paper.

References

Kenkyūjo, N. H. B. (1998). NHK nihongo hatsuon akusento jiten [NHK Accent Dictionary of the Japanese Language]. *Tokyo: Nihon Hōsō Shuppan Kyōkai.*

Radford, A., Kim, J. W., Xu, T., Brockman, G., Mcleavey, C., & Sutskever, I. (2023). Robust Speech Recognition via Large-Scale Weak Supervision. *Proceedings of the 40th International Conference on Machine Learning*, 28492–28518. https://proceedings.mlr.press/v202/radford23a.html

Sugito M. (1995). *CD-ROM accent dictionary of spoken Osaka and Tokyo Japanese*. Maruzen.

Tachibana, H., & Katayama, Y. (2020). *Accent estimation of Japanese words from their surfaces and romanizations for building large vocabulary accent dictionaries*. https://doi.org/10.1109/ICASSP40776.2020.9054081

Toshinori, S. (2015). *Neologism dictionary based on the language resources on the Web for Mecab*.