YouTubeを利用した関西方言アクセント辞書の作成

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 presents an attempt to create an accent dictionary of the Osaka dialect using user-generated content on YouTube. Speech data were extracted from videos, transcribed, and force-aligned based on the transcribed speech. The pitch of each segment was measured, and the pitch patterns were automatically detected. This paper discusses the preliminary application of the process for a single video.

1. Introduction

Language processing has undergone significant advancements, particularly in text-to-speech systems and language models. However, for languages spoken by smaller populations, linguistic resources remain underdeveloped. This disparity is evident within dialects in the Japanese language. While substantial linguistic resources exist for standard Japanese, regional dialects are often overlooked. Regarding accents, well-established accent dictionaries exist for the Tokyo dialect, such as those by NHK (1998) and open-source contributions by Tachibana and Katayama (2020).

Conversely, although accent dictionaries exist for the Osaka dialect, such as Sugito (1995), comprehensive open-source versions do not exist. Open-source accent dictionaries are valuable for linguistics and language processing. However, unless researchers have a rich network of informants and generous funding, generating extensive accent data, including thousands of words read by multiple speakers, can be a tremendous task. This paper presents a unique method for automatically generating accent dictionaries using user-generated content on YouTube.

2. Previous Studies

Sugito (1995) conducted in-depth interviews with three informants from two generations to determine accent patterns. As presented in Example (1), the pitch of each word is marked with an “L” for low and “H” for high. The Osaka dialect is characterized by the fact that each word begins with a lexical high or low.

(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 development of accent dictionaries has traditionally been grounded in rigorous fieldwork, employing written documentation, auditory analysis, and direct recording. Although these approaches

[†] noguchih425@gmail.com

enable efficient acquisition of word accents essential for headwords, they have limitations. Many of these dictionaries primarily rely on lists that do not reflect natural speech. Furthermore, these methods are time-consuming, posing challenges to comprehensive and swift data collection.

3. Methods

3.1 Data

Data from YouTube were used to eliminate the burden on informants and capture authentic speech patterns. YouTubers were chosen based on the following conditions: the YouTuber (1) must be from the Kansai region, (2) must be speaking alone, and (3) must not play background music. For this study, a video from one such qualified YouTuber was selected.

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. The transcribed text was converted into hiragana using mecab-ipadic-NEologd (Toshinori, 2015) because the original version of MeCab does not correctly recognize new words like "YouTuber." Afterward, forced alignment was performed using Julius (Kawahara, 2015). The average pitch for each segment was measured in hertz and assigned an H or L based on whether it was higher or lower than the average pitch for each word. Figure 1 presents an example of the segmentation results.

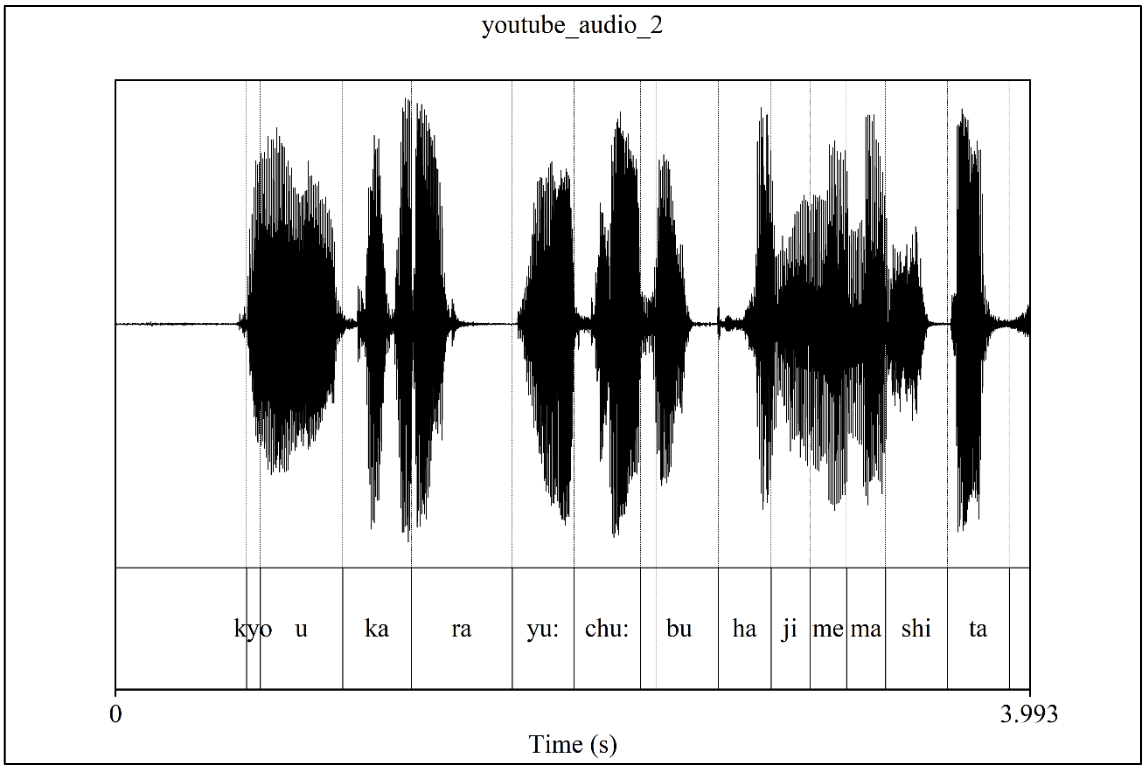

Figure 1: Segmentation example.

4. Results

The pitch was measured for each segment in Example (2). The average pitch for each word was compared to the segments to avoid down-step effects. Although some differences occur from the auditory impression, the pitch curve in Figure 2 can be transcribed to the symbols in Example (3). Due to the specifications in Julius, long vowels are not divided into mora units, and pitch transcription within long vowels does not reflect the actual pitch contour.

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

(3)

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)

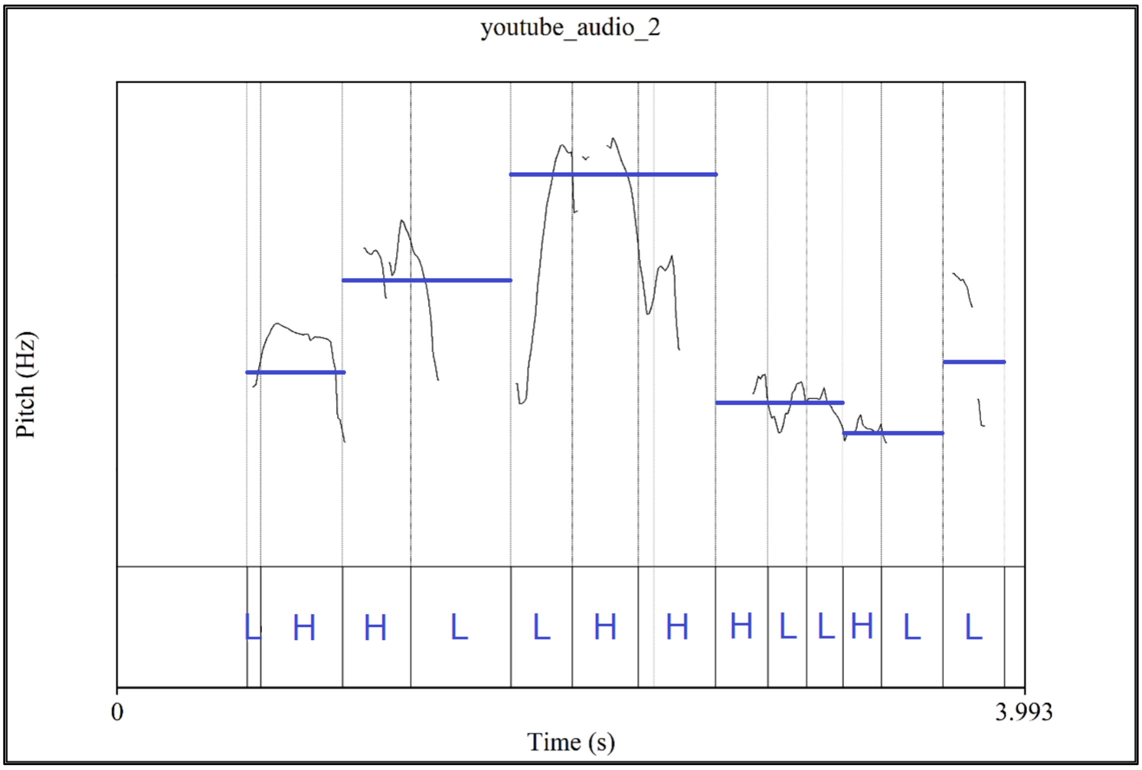

Figure 2: Pitch contour for Example (2).

5. Discussion

The pitch transcription method may not initially seem appropriate. However, some words appear more than once. For example, Example (4) indicates the accent pattern for the word “YouTube” was identified as in the video. The most frequently pronounced pattern, Example (4)a, which was pronounced 10 times, is also consistent with the auditory impression. Increasing the number of videos is expected to obtain more reliable accent patterns for a more extensive vocabulary. In this 16-min video alone, 474 types and 918 tokens appeared in terms of word count. Further investigation is needed in the future.

(4)
a. [‘yu:,’ ‘chu:,’ ‘bu’] [‘H,’ ‘H,’ ‘L’] * 10
b. [‘yu:,’ ‘chu:,’ ‘bu’] [‘L,’ ‘H,’ ‘L’] * 2
c. [‘yu:,’ ‘chu:,’ ‘bu’] [‘H,’ ‘L,’ ‘L’] * 1
d. [‘yu:,’ ‘chu:,’ ‘bu’] [‘L,’ ‘H,’ ‘H’] * 1

Acknowledgments

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

References

Kawahara, T. (2005). Open-source speech recognition software Julius. *JSAI*, 20(1), 41-49.

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. Tokyo: Maruzen.

Tachibana, H., & Katayama, Y. (2020). Accent estimation of Japanese words from their surfaces and romanizations for building large vocabulary accent dictionaries. *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. 8059-8063. https://doi.org/10.1109/ICASSP40776.2020.9054081

Toshinori, S. (2015). Neologism dictionary based on the language resources on the Web for Mecab. https://github. com/neologd/mecab-ipadic-neologd