Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 73 (2015) 47–63

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Automatic recognition of Japanese vowel length accounting for speaking rate and motivated by perception analysis

Greg Short [a,*], Keikichi Hirose [b], Mariko Kondo [a], Nobuaki Minematsu [b]

[a] *Waseda University, Tokyo, Japan*
[b] *University of Tokyo, Tokyo, Japan*

Received 28 January 2015; received in revised form 30 April 2015; accepted 8 July 2015
Available online 21 July 2015

Abstract

Automatic recognition of vowel length in Japanese has several applications in speech processing such as for computer assisted language learning (CALL) systems. Standard automatic speech recognition (ASR) systems make use of hidden Markov models (HMMs) to carry out the recognition. However, HMMs are not particularly well-suited for this problem since classification of vowel length is dependent on prosodic information, and since it is a relative feature affected by changes in the durations of surrounding sounds which vary in part due to changes in speaking rates. That being said, it is not obvious how to design an algorithm to account for these contextual dependencies, since there is still not enough known about how humans perceive the contrast. Therefore, in this paper, we conduct perceptual experiments to further understand the mechanism of human vowel length recognition. In our research, we found that the perceptual boundary is largely affected by the vowels two prior, one prior, and following the vowel for which the length is being recognized. Based on these results and the works of others, we propose an algorithm which does post-processing on alignments output by HMMs to automatically recognize vowel length. The method is primarily composed of two levels of processing dealing first with local dependencies and then long-term dependencies. We test several variations of this algorithm. The method we develop is shown to have superior recognition capabilities and be robust against speaking rate differences producing significant improvements. We test this method on three different databases: a speaking rate database, a native database, and a non-native database.

Keywords: Vowel length; Automatic recognition; Perception; Duration; Resynthesis; Stimulus continua

1. Introduction

For Japanese vowels, length is a contrastive lexical feature with each vowel having a short (S) and long (L) version. Automatic recognition of this feature has several applications in speech processing. For example, it is used in speech recognition. Also, it could be used in a computer assisted language learning (CALL) system to recognize learner errors. As many languages do not have a lexical contrast of short and long vowels, this feature of Japanese pronunciation can be difficult for nonnatives to acquire and a system that could provide feedback to the learner could be of great use (Tsurutani, 2010).

Automatic recognition of vowel length is not as straightforward as it would seem at first glance, however. Despite the contrast being primarily a result of durational differences of the vowel, the duration of the vowel is not the only variable classification depends on in recognition by natives. The duration of that vowel relative to the surrounding vowels (Hirata and Lambacher, 2004; Hirata and Whitton, 2005; Takeyasu and Giriko, 2010) as well as other features such as fundamental frequency and intensity all are said to play a role (Takiguchi et al., 2010; Matthews et al., 2011). With increased duration in the sounds

[*] Corresponding author.

http://dx.doi.org/10.1016/j.specom.2015.07.001
0167-6393/© 2015 Elsevier B.V. All rights reserved.

surrounding a particular vowel, the duration at the perceptual boundary separating short and long vowel categories increases. Hence, a robust function for classifying whether a vowel sound is short or long would be dependent not only on the duration of the vowel to recognize, but also on the durations of surrounding sounds.

For speech recognition systems, recognition is typically carried out by having separate phoneme HMMs for short and long vowels. While this approach does not require any changes to a standard speech recognition system, HMMs have the disadvantage that they by definition have little dependency on temporal variables. This results in them performing recognition in an absolute manner rather than a relative manner. Incorporating this information can be important in CALL systems since the distribution of speaking rates produced by nonnative speakers can be quite large. There have been efforts to incorporate temporal information in automatic recognition, but these efforts have not been successful in obtaining a robust function to perform this classification. They will be discussed in more detail in Section 2.1.

To incorporate temporal information into automatic vowel length recognition, a natural approach is to do post-processing on the phoneme alignments obtained by a speech recognizer as in Fig. 1. This is the general approach that we take in this paper. The major problem is that it is not clear what form the procedure coming after the forced alignment should take. Thus, this is the problem we tackle in this paper.

To guide us to a solution to this problem, we look toward the perception of vowel length in Japanese to provide a basis for how to develop the recognition algorithm. While past works in human recognition, discussed in Section 2.2, provide some clues for developing an automatic recognition method, they are still not sufficient.

After examining previous endeavors into the fields of machine and human recognition of vowel length (Section 2), we conduct further perceptual experiments in Section 3 examining how the duration at the perceptual boundary changes due to the durations of nearby sounds. Following that, we propose an algorithm based on the perceptual experiments in Section 4. Then, we conduct experiments using this algorithm discussed in Section 5. Lastly, we conclude the paper in Section 6. Since the term vowel length can mean both the time duration and phonemic length of the vowel, henceforth we use the term length to mean solely the latter and duration to refer to the time-span to avoid confusion.

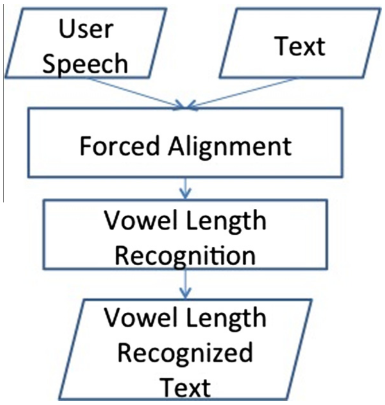

Fig. 1. A general system flow for recognizing vowel length.

2. Machine and human vowel length recognition: past research

2.1. Research on automatic recognition

There have been a few methods proposed for automatically classifying vowels as short or long other than the standard HMM-based method described in Section 1. These fall into two categories: those that attempt to recognize in an absolute manner, and those that try to incorporate speaking rate.

In Kawai's work (Kawai, 1999), Kawai created a method to recognize vowel length in an absolute manner employing resynthesized stimuli in listening tests for deriving classification equations. In these tests, the author used minimal pairs differentiated by the length of one vowel such as the pair /to:ru/ (to pass) and /toru/ (to take). The stimulus sets were created by resynthesizing a continuum of durations by lengthening and shortening the duration of the vowel that distinguishes the two words. This means, for the pair /toru/ and /to:ru/, a continuum of stimuli ranging from /toru/ to /to:ru/ was created. Then, these stimuli were played to native speakers of Japanese and the subjects selected which of the two words they heard the word as. The responses for each pair used in the selection test were fit to a logistic equation with the x-axis being the vowel duration and the y-axis the selection rate for the duration of the word-distinguishing vowel. This function was then used for classification. This method, however, suffers from the same weakness that the HMM method suffers from in that it does not account for speaking rate.

Two other methods attempting to take speaking rate into account and perform recognition in a relative manner were then proposed. One of these methods was proposed by Yamamoto and Miwa (2000). They conducted listening tests like in Kawai's method. The difference was that for each vowel duration in the continuum, they created a continuum of word durations manipulating the entire word duration. Thus, through this approach, they were able to derive an equation taking into account how the perceptual boundary which distinguishes short from long varies due to word duration. They did this for two different words.

This approach can lead to problems, though, since ideally the method should be able to recognize length for an arbitrary word. It is unclear, though, if this method would be successful at doing this. It seems unlikely that word duration would be used to normalize for speaking rate in human recognition. This would result in a probability function of the form $P(L|wd, vd)$ where wd is the word duration, vd is