3 ICLICE 2016 45 Atsushi Kojima

Detection of Focus for Importance of Words: Toward Computer-Assisted Learning for Improving Presentation

Atsushi Kojima*[a], Katunobu Itou[b], Hiroshi Hanaizumi[a]
[a]Graduate School of Computer and Information Sciences,
Hosei University Tokyo, Japan
[b]Faculty of Computer and Information Sciences,
Hosei University Tokyo, Japan
*Corresponding Author: atsushi.kojima.2v@stu.hosei.ac.jp

ABSTRACT

This study introduces a presentation practice system, which enables a presenter to emphasize important words. The system indicates specific words emphasized by the user during a presentation. Based on knowledge of Japanese phonetics and language education, we focused on accent pitch, intensity, pauses, the position of a word in an utterance, and speech rate. When a presenter uses the system, they utter a sentence including a word to be emphasized in their presentation. They confirm detected results on all words in the utterance and repeat it until the word they wish to emphasize is detected as the focus. In an evaluation experiment, three students used the system until they could emphasize the focus correctly in their presentation relating to their own researches. Three other students allotted scores to each word given emphasis according to three grading levels. In the results, the mean value of the emphasis given to the chosen words by the presenters was higher by one grade after the practice system than before. This demonstrates that the feedback obtained facilitated the acquisition of speaking skills in relation to emphasizing important words in a presentation.

Keyword:, focus, prosody, focus detection, presentation, focus education, CALL (Computer Assisted Language Learning)

Introduction

In a presentation, the "focus" is used in a "place" that a speaker wants to bring attention to, so the place includes important information. For example, the place includes the importance of a word (Sridhar, 2008). An audience can know keywords in the presentation, and the presentation is easy to understand. In this paper, the focus is a factor for achieving an audience's understanding, but it is difficult for a presenter to understand if they can emphasize the focus. This is because the focus is invisible. So we proposed a method to detect an emphasized word in order to construct a presentation practice system in the Japanese language. Words targeted for detection are nouns.

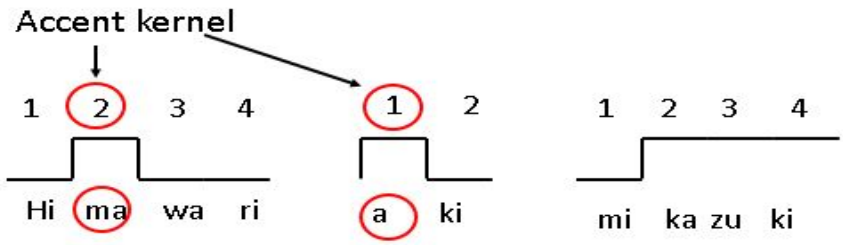
Many researchers have researched the detection of the focus for segmenting a presentation (Arons, 1994) and summarizing a presentation (Francine, 1992). Detecting the focus is researched in the English language, but detecting the focus in the Japanese language has not been researched, so we propose a method for detecting the focus based on a way to emphasize a word in a Japanese utterance.

A practice system using this method differentiates focuses or non-focuses in all words in a user's utterance, and the user confirms the detected results on all words in the utterance, then the user repeats the presentation until the word they wish to emphasize is detected as a focus. The goal is that the user acquires the focus in their presentation.

A Focus of a Word in Japanese

Previous Research in Japanese Linguistics

We reviewed the existing literature to obtain knowledge about how to emphasize a word in a sentence that uses phonetics (Nakajo, 1989) and a language education (Koori, 1989; Nakagawa, 2010). As a result, we found that the focus correlates with a pitch accent, a power in a word, a pause, and a speech rate. In particular, for pitch accent, the pitch of an accent for an emphasized word rises (Koori, 1989, p.316–339). This is because an accent in Japanese is a pitch accent (Mark, 2011). A pitch of a syllable on the word is spoken either high or low. The pitch on a word descends only once. The position of a syllable right before descending is called an accent kernel. A word has are noted at 0 place or 1 place within an accent kernel. Words that have 0 place within the accent kernel are called flat type. Words that have 1 place within the accent kernel are called falling type. A type of accent is decided by the position of an accent kernel. Figure 1 shows an example of pitch accent for words. A line above the syllable notates a pitch accent. Figure (a) shows the pitch accent of /himawari/. In this word, /ma/ is an accent kernel. The accent type of this word is a 2-type, so this word is a falling-type. Figure (b) shows the pitch accent of /aki/. In this word, /a/ is the accent kernel. The accent type of this word is a 1-type, so this word is a falling-type. Figure (c) shows the pitch accent of /mikazuki/. In this word, the number for the accent kernel is 0 position, so this word is a flat-type. When these words are emphasized, the accent type becomes clear because the pitch accent rises (Koori, 1989, p.316–339).



(a) Accent of /himawari/ (b) Accent type of/aki/ (c) Accent type of /mikazuki/

Figure1: An accent type of /himawari/ is 2-type (falling type). The accent type of /aki/ is 1 type (falling type). The accent type of /mikazuki is flat type, so the word has 0 accent kernel.

In relation to the power of a word, an emphasized word is uttered strongly (Nakajo, 1989, p.125–126).

In relation to the pause, the pause is put before and/or after the word when the word is emphasized (Koori, 1989, p.316–339).

Quantification of Focus

We quantify the focus based on the knowledge of focus. The accent pitch is estimated by calculating the maximum of the accent component (A) of each word by adapting F0 as the Fujisaki Model (Fujisaki, 1969). Additionally, maximum of F0 (F) is also used. A power of a word is estimated by calculating the maximum of the logarithmic power (P). The pause is estimated by calculating a length of time of a pause before a word (Pab) and a time length of a pause after a word (Paa).

Speech rate is quantified by two methods. The first method is a time length of a word (Wl) and the second method is a time length of per 1 syllable (Sl) in a word. In order to represent whether a word the user wishes to emphasize is located in a place close to the beginning of a sentence, we quantify the position of a word in a sentence. In this paper, we estimated that by calculating the start position of a word (Ws) and end position of a word (We) the normalized time length of the sentence is 1.

Estimating Features for Detecting the Focus

We estimate the accent component, power, length of time of a pause before a word, a length of time of a pause after a word, time length of a word, time length of a word per 1 syllable, and a position of a word in an utterance; however, a beginning time of word and an end time of word are given in advance by WaveSurfer (Sjolander, 2000), and the transcription of the utterance is given.

In order to estimate an accent command modeled by square wave, the accent command is square wave so it is decided by three parameters (beginning time of command, end time of command and amplitude). Therefore, we need to estimate those parameters (Narusawa, 2002).

In particular, we estimate F0 by a method (Morise, 2009) using low-pass filter and interpolate it using linear interpolation and smooth it with a median filter. Then we fit the cubic curve to F0 in each section and differentiate it in time direction. As a result of the differential, extra value is an amplitude, and time with extra values are beginning time of command and end time of command.

We can calculate the accent component $y_a(\mathrm{t})$ from the accent command $u_a(t)$ using

$$y_a(t) = G_a(t) * u_a(t). \quad (1)$$

$G_a(t)$ is given by

$$G_a(t) = \begin{cases} \beta^2 t \exp(-\beta t) & (t \geq 0) \\ 0 & (t < 0), \end{cases} \quad (2)$$

where t is time, * is convolution and β is the parameter that shows slope. Individually, β is small, so it can be regard as constant (20 rad/s) (Fujisaki, 1969). According to the estimated accent component, the maximum accent component from beginning time to end time is used as a feature.

To estimate the time length of pause, we adapted VAD (Voice Activity Detection) using the threshold of zero crossing and threshold of log power calculated by frames including only background noise (Rabiner, 1975), and an interval of speech is regarded as a pause. If there is a pause before the beginning time of a word or after the end time of a word, these are regarded as the pause before the word and the pause after the word, and the time length of pause is calculated by subtracting the end time of the pause from the beginning time.

Power is calculated using

$$\mathrm{E} = \sum_{m=0}^{L-1} (s[rR + m] \cdot w[m])^2 \quad (3)$$

$$r = 0,1,2\ldots.,$$

where r is the number of frames, L is the frame length, and R is the shift length. It is normalized using

$$E_{norm} = 10\log10 \frac{E}{\max(E)}. \quad (4)$$

In the series of E_{norm}, the maximum from a beginning time of word to the end time of word is a feature.

Time length of a word Wl is calculated using

$$\mathrm{Wl} = W_{etime} - W_{stime}, \quad (5)$$

where W_{etime} is the end time of a word and W_{stime} is the beginning time of a word.

Time length per 1 syllable in a word Sl is using is

$$\mathrm{Sl} = \frac{Wl}{S}, \quad (6)$$

where S is the number of syllables of a word.

The beginning position of a word in an utterance of Ws and an end position of word in an utterance of We are calculated using

$$\mathrm{Ws} = \frac{W_{stime}}{Wl}, \tag{7}$$

$$\mathrm{We} = \frac{W_{etime}}{Wl}. \tag{8}$$

Figure 2 shows the features of words being analyzed in an utterance including the emphasized word.

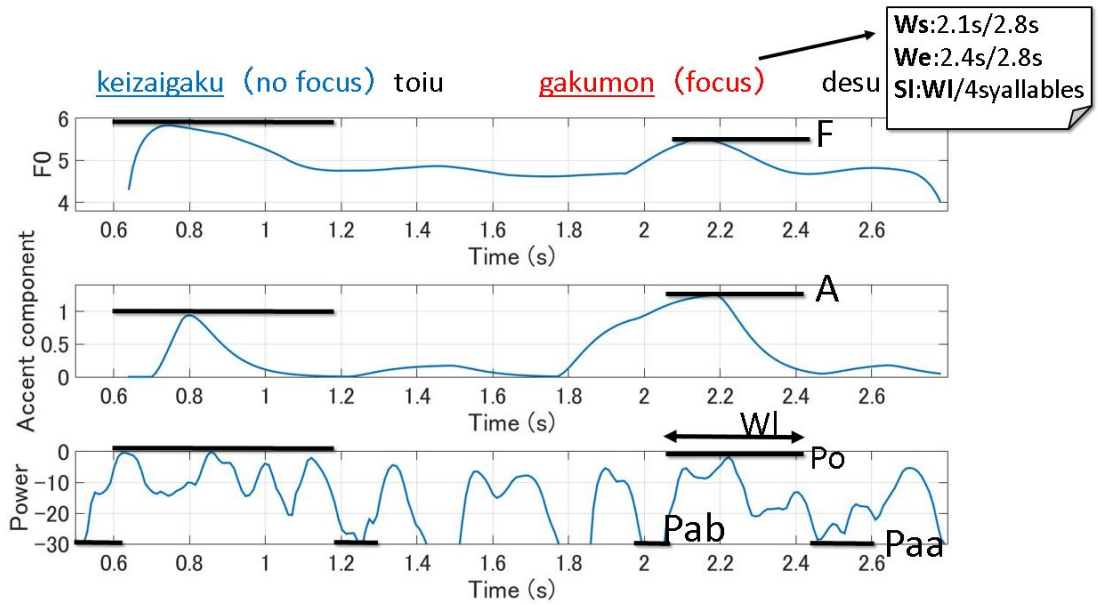


Figure 2: Example of Acoustic Features for Detecting Focus

Making a Corpus of an Emphasized Sentence

We collected utterances, including an emphasized word, in order to train and evaluate the proposed method. We tried to collect these utterances from TV and prepare them as a corpus. We prepared 303 utterances. A time length for these sentences ranged from 4.22 s ~ 17.8 s. The average time length of these sentences was 6.23 s.

This corpus is used as training data for detecting the focus in the practice system, and we evaluated the agreement rate of focus perception by the listeners. Other three listeners but a worker who gave labels listens 50 utterances selected randomly from the corpus, and make marks on emphasized words. The three listeners are not told words the worker who gave label at first listened emphatic. Number of marks is not limited, and they can listen to these utterances as many times as one likes.

The results showed that the focus labeled by the worker, was also given the same label by two or more labelers at a rate of words of 0.91. A rate of words for the same label given by all labelers was 0.50. When an emphasized word was put at beginning of a sentence, all the labelers gave the focus a label.

Evaluation of Focus Detection

We performed an experiment in order to verify how effective the proposed method is in detecting the focus. For the evaluation, 978 words were given a focus label or non-focus label. Information concerning the label was given to one worker. The Evaluation scales were based on the precision, recall, and correct classification rate. These scales are calculated by

$$\text{precision} = \frac{\text{detected number of emphasized words}}{\text{detected number of words}}, \tag{9}$$

$$\text{recall} = \frac{\text{detected number of emphasized words}}{\text{number of emphasized words}}, \tag{10}$$

$$\text{Classification rate} = \frac{\text{detected number of emphasized words+detected not emphasized words}}{\text{number of emphasized words+number of not emphasized words}}. \quad (11)$$

In an analysis condition for calculating the F0 and accent component, a frame length is 30 ms and a frame shift is 10 ms. In an analysis condition for calculating power, a frame length is 50 ms and a frame shift is 10 ms. To calculate the features, we used the Hann window. Classifier is the logistic regression model (MeCullagh, 1990). Probability was calculated by a model after training was 0.5 or more and a word was detected as a focus.

The results found that precision was 0.60, recall was 0.71, the F-measure was 0.70, and the classification rate was 0.69. For the classification rate, we used all the features at the same level as a that of previous study (Francine, 1992) with a classification rate of 0.70 that focused on detecting the focus in English.

In addition, we conducted an experiment for detecting the focus for every feature. Table 1 shows the features for detecting the focus, precision, recall, F-measure, and classification rate. According to the F-measure, the most effective feature is the accent component, and F0 and power are equally effective. A time length of a word and a beginning position of a word are more effective than the length of time of a pause before a word, and the end position of a word is as effective as a length of time of a pause after a word.

Table 1
Result of detecting focus every features

Feature	A	F	Po	Wl	Ws	Pab	We	Sl	Paa
Precision	0.62	0.59	0.65	0.61	0.53	0.64	0.52	0.51	0.49
Recall	0.71	0.61	0.56	0.51	0.55	0.44	0.53	0.49	0.21
F-measure	0.66	0.60	0.60	0.56	0.54	0.52	0.52	0.50	0.29
Classification rate	0.64	0.59	0.63	0.60	0.53	0.60	0.52	0.51	0.50

Furthermore, we evaluated the training data and the efficiency for detecting the focus. We selected training data from 978 words (focus: 489 words, non-focus: 489 words) randomly and attempted to detect the focus from all words 1000 times. We increased the training data from 100 words to 900 words every 100 words. Figure 3 shows the relation between the training data and the efficiency. The error bar shows a standard deviation. In a range from 100 words to 500 words, the precision, recall and classification rate rose, but the precision, recall, and classification rate stayed the same when we used 500 words to 900 words.

Being effective to the position of a word in an utterance is more than a time length of pause and is matched with the knowledge that a position of a word in an utterance is effective perceptually (Toyama, 2007).

Also, in not being effective at a time length per one syllable more than a time length of a word, we could verify an average time length per one syllable between emphasized words and words that were not emphasized. As a result, it became 0.45 s and 0.41 s, respectively, and there was no significant difference. Efficiency was about the same when we used 500 words for the training data being as when we used 900 words, so efficiency did not rise just by increasing the training data; therefore, we will need to select features following more research.

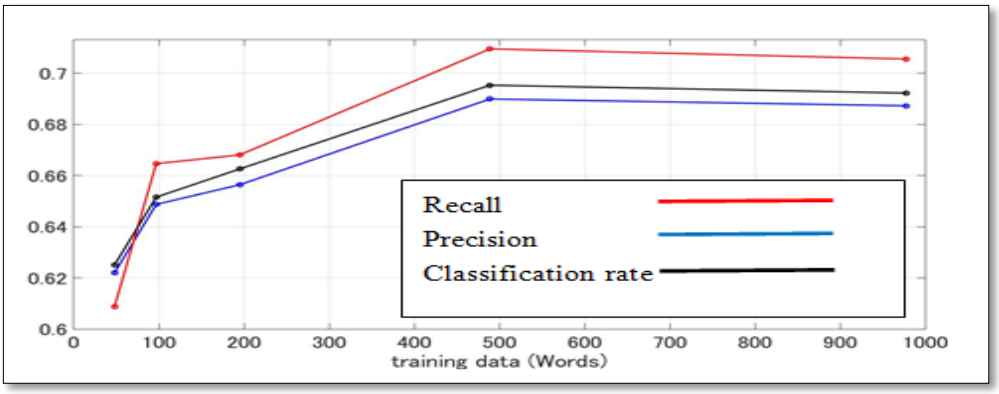


Figure 3: Training data and efficiency of detecting focus

Evaluation

Experimental Method

We define system input as an utterance by which a user tries to speak a word with emphasis. In our analysis, we calculated the features in an input utterance. Outputs are the probability of each word being given a discriminating emphasis. The value of probability ranges from 0 to 1. If the value is closer to 1, the probability of a word sounding emphatic is higher. We produced the sample utterance by manually shifting from an analyzed default value where the maximum of a word's accent component was spoken and emphasized by the user. This value may be shifted in 0.1 increments to convey emphasis without breaking into a falsetto voice. The maximum accent component in the input utterance was then converted to produce the sample utterance.

We conducted an evaluation to determine whether users could improve their ability to emphasize words that they wanted to emphasize by practicing their presentations using the proposed system. The subjects, three male students, who had never been trained in abdominal breathing, made a presentation in groups of two and three. The themes of their presentations were abstract introductions on their graduation work in computer science. The presentations were to contain five words they wanted to emphasize. The subjects chose the words they wanted to emphasize in advance, and then made their presentations. The presentation formats were realistic in their use of slides and pointers. The subjects practiced their presentations using this system. If the words chosen by the subjects appeared once or twice, the subjects were instructed to emphasize the word where the subject most wanted to emphasize the flow of the story. The subjects practiced until a word that each subject had chosen was recognized as having emphasis. After practice, the subjects made their presentation again. Additionally, the presentation texts before and after practice were not strictly identical because small details naturally changed. However, the sentence containing the word that the subjects wanted to emphasize differed very little before and after, so the results were not affected.

In the subjective evaluations, three evaluators, none of whom were presenters, listened to an entire speech and scored all the words in a transcript (1: sound not emphatic, 2:

somewhat emphatic, 3: sound emphatic).

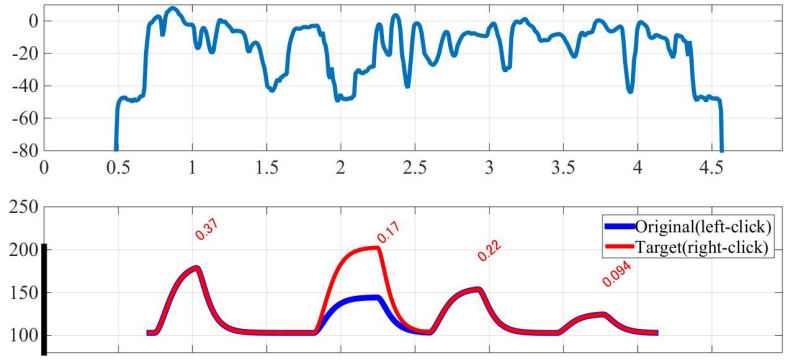


Figure 4: GUI for practicing presentations. Upper: log Power, Lower: pitch curve of sample utterance (red curve), pitch curve of input (blue curve), and probability of discriminating emphasis

The speech was streamed only once, as with a real presentation. The order of the speeches that were listened to by the evaluators, either before practice or after practice, was assigned at random. Furthermore, the evaluators were not told the words or the number of words that the subjects chose to emphasize. In the objective evaluations, we analyzed the sentences containing the words the subjects chose to emphasize, and we calculated the probability of differentiating the emphasis given to words before and after the practice.

Experimental results

All 15 words chosen by the subjects were either a proper name or a technical term relating to their research. While the number of subjects was small, the results of the subjective evaluations yielded an average word score of 1.05 before the practice and 1.10 after the practice. A t-test of the average results showed a significant level of $\alpha = 0.05$, so these average values were equal. Furthermore, the average of the word scores that the subjects wanted to emphasize before the practice was 1.20 and after the practice the average score was 2.20. A t-test of the averages showed a significant level of $\alpha = 0.05$, so they were accepted as significantly different. The average score of words that the subjects had not wanted to emphasize before the practice was 1.08 and the average score after the practice was 1.07. The t-test of this average was significant to the $\alpha = 0.05$ level, so the result was accepted as not being significantly different.

Using the proposed system, the average score of the words that the subjects wanted to emphasize was increased by 1.0. In other words, the word emphasis went up one grade level, so if a word spoken by the subjects did not sound emphatic before practice, it sounded emphatic after practice. Furthermore, the average of all the words was equal. This result demonstrated that not all the words sounded emphatic, and only words that the subjects wanted to emphasize sounded emphatic. The results of the objective evaluations showed that the average word emphasis before practice was 0.29 and after practice was 0.60. We performed a t-test of the word averages to determine whether before the practice its score was less than 0.50 and after the practice its score was more than 0.50 at a significant level of $\alpha = 0.05$ (0.50 is a boundary to discriminate a word as emphasis), and the results were accepted as significantly different. Thus, we confirmed that the words the subjects chose were emphasized on average, objectively showing that the subjects were able to acquire the skill to emphasize.

Based on comments by the subjects, they reported that they had learned to emphasize by viewing the pitch curve and the probability of demonstrating recognizable word emphasis, to help them become aware of speaking words consistently rather than emphasizing important words. According to the comments by the evaluators, in speeches before practice, auxiliary verbs or conjunctions sounded emphatic, but after practice, nouns sounded emphatic. Therefore, differentiating words for emphasis was effective for acquiring the ability to give desired words emphasis.

However, the emphasis of one specific word—MPEG4—was not improved. In subjective evaluations, its score increased, but in objective evaluations, the word was scored as having no emphasis. We then confirmed by actual listening that the word did not sound emphatic. The reason is that MPEG4 is a flat-accent-type word. Flat-accent-type words are apt to be difficult to speak in a high pitch (Koori, 1989).

Conclusion

This study proposes a system to support individuals' ability to learn to emphasize important words presentations. The system determines whether a user is emphasizing the words they want to emphasize. When a presenter uses the system, they utter a sentence including a word to be emphasized in their presentation. They confirm the detected results on all the words in the utterance, and they repeat the utterance until the word they wish to emphasize is detected as the focus.

To propose a method to detecting the focus in the system, we researched ways to emphasize a word in Japanese utterances based on Japanese phonetics and Japanese education. Therefore, we found that the focus correlates with a pitch accent, power in a word, a pause, and speech rate; we quantified the focus based the knowledge. In particular, pitch accent was quantified as a maximum of the accent component in a word. A pause was quantified as a time length of the pause before the word and a time length of the pause after the word. Power was quantified as a maximum of log power in the word. The speech rate was quantified as a time length of the word and a time length of 1 syllable in the word. The closeness of distance between the word and the beginning of a sentence was quantified as a start position of a word and an end position of a word that was normalized by time length of the sentence is 1. This method can detect the focus with a 0.69 classification rate.

We performed subjective evaluations to determine whether the user was emphasizing words they wanted to emphasize by using the proposed system. Our results indicated that the average score of all the words increased to the 1.0 level with respect to emphasis perceived, indicating that this system is effective for the acquisition of emphasis in presentations. In future studies, we aim to increase the number of subjects to improve the system evaluation.

References

Arons, B. 1994. Pitch-based emphasis detection for segmenting speech recordings. *Proceedings of International Conference on Spoken Language Processing vol.4*, (pp. 1931–1934).

C. Nakagawa and N. Nakamura, Shokyuubungata de dekiru nihongohatsuonakuthibithi, (Japanese), *asuku shuppan,* 2010.

Francine, R. et al. 1992. The use of emphasis to automatically summarize a spoken discourse. *in Proc. ICASSP*, (pp. 229–232).

Fujisaki, H. et al. 1969. A model for synthesis of pitch contours of connected speech, *Annual Report, Engg. Res. Inst.,University of Tokyo, vol. 28*, (pp. 53–60).

Fujisaki, H., et al. 1984. Analysis of voice fundamental frequency contours for declarative sentences of Japanese. *The Journal Of The Acoustical Society Of Japan*, 5 (4), 233–242.

McCullagh, P., & Nelder, J. A. 1990. *Generalized linear models*. New York: Chapman Hall.

Morise, M., et al. 2009. Fast and reliable F0 estimation method based on the period extraction of vocal fold vibration of singing voice and speech. *In AES 35th International Conference, CD-ROM, London, UK*.

Narusawa, S., et al. 2002. A method for automatic extraction of model parameters from fundamental frequency contours of speech. *In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 509–512).

O, Nakajo, Nihongo no onin to akusento, (Japanese), *keiso shoboo*, 1989.

Rabiner, et al. 1975. An algorithm for determining the endpoints of isolated utterances, *The Bell System Technical Journal*, 54 (2), 297–315.

S. Koori, Kouza nihongo to nihongo kyouiku2 nihongo no onsei onin, (Japanese), *Meiji shoin*, 1989.

Sjolander and Beskow, WaveSurfer? An open source speech tool, *in Proc. ISCSLP*, 2000, (pp. 464–467).

Sridhar V, K, R, et al. 2008. Detecting prominence in conversational speech: pitch accent, givenness and focus. In P. Barbosa, S. Madureira, & C. Reis (Eds.), *Proceedings of the Speech Prosody Conference* (pp. 453–456).

Toyama, A. tsutawaru hanashikata notameno 10no ruuru, (Japanese),

van Oostendorp, M. et al. 2011. *The Blackwell Companion to Phonology*. Wiley-Blackwell.