INVITED PAPER *Special Section on Award-winning Papers*

Accent Sandhi Estimation of Tokyo Dialect of Japanese Using Conditional Random Fields

Masayuki SUZUKI[†*a)], ***Member***, **Ryo KUROIWA**[†**b)], **Keisuke INNAMI**[†***c)], **Shumpei KOBAYASHI**[†****d)], **Shinya SHIMIZU**[†*****e)], ***Nonmembers***, **Nobuaki MINEMATSU**[†f)], ***Senior Member***, ***and*** **Keikichi HIROSE**[†g)], ***Fellow***

SUMMARY When synthesizing speech from Japanese text, correct assignment of accent nuclei for input text with arbitrary contents is indispensable in obtaining naturally-sounding synthetic speech. A phenomenon called accent sandhi occurs in utterances of Japanese; when a word is uttered in a sentence, its accent nucleus may change depending on the contexts of preceding/succeeding words. This paper describes a statistical method for automatically predicting the accent nucleus changes due to accent sandhi. First, as the basis of the research, a database of Japanese text was constructed with labels of accent phrase boundaries and accent nucleus positions when uttered in sentences. A single native speaker of Tokyo dialect Japanese annotated all the labels for 6,344 Japanese sentences. Then, using this database, a conditional-random-field-based method was developed using this database to predict accent phrase boundaries and accent nuclei. The proposed method predicted accent nucleus positions for accent phrases with 94.66% accuracy, clearly surpassing the 87.48% accuracy obtained using our rule-based method. A listening experiment was also conducted on synthetic speech obtained using the proposed method and that obtained using the rule-based method. The results show that our method significantly improved the naturalness of synthetic speech.

key words: *Japanese text-to-speech, accent sandhi, accent phrase boundary estimation, accent type estimation, conditional random field*

1. Introduction

Japanese text-to-speech (TTS) systems need to estimate not only phonetic symbols but also accent information from text as pre-processing [1]. TTS systems then generate speech waves given the phonetic symbols and accent information shown in Fig. 1. Because errors of accents degrade the naturalness of outputted speech, accent information from text needs to be correctly estimated.

Japanese words have their own accent nucleus position as one of their lexical attributes. However, its positions often shift when the words are read in sentences due to a phenomenon called accent sandhi. Figure 2 shows an example of accent sandhi of Tokyo dialect. The Japanese words for "Tokyo" and "university" have their own accent nucleus positions as their lexical attribute. The nucleus positions are changed when they are combined in a sentence as "The University of Tokyo" like it is in Fig. 2. Native speakers of Japanese unconsciously and correctly estimate accent sandhi. Nevertheless, making clear and perfect rules for accent sandhi is almost impossible.

Our objective is to build a high-quality system for estimating accent sandhi for Tokyo dialect of Japanese to improve the naturalness of TTS. We first built a database of 6,334 sentences that were labeled with accent phrase boundaries and an accent nucleus position of each accent phrase in sentences. Then, we then developed a conditional random fields (CRF) based system to estimate accent information from sentences. The proposed system achieved 94.66% accuracy, which is a 57% relative improvement from our rule-based baseline system. We also found that our method significantly improves the naturalness of speech synthesis.

2. Definition of Accent of Tokyo Dialect

Japanese has pitch accent at the mora level. A mora is a unit in phonology, an it generally corresponds to a Japanese Hiragana characters except Youon. An accent phrase is composed of some morphemes. Accent sandhi occur within the accent phrase. In other words, accent sandhi never occur across the boundaries of accent phrase. As for Tokyo di-

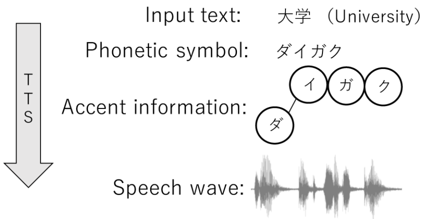


Fig. 1 Flowchart of Japanese Text-To-Speech (TTS).

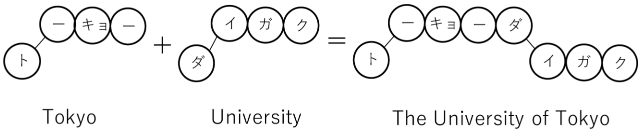


Fig. 2 Example of Japanese accent sandhi.

Manuscript received October 31, 2016.
Manuscript publicized December 8, 2016.

[†]The authors are with The University of Tokyo, Tokyo, 113–8656 Japan.
[*]Presently, with IBM Research - Tokyo.
[**]Presently, with NTT DATA Corporation.
[***]Presently, with Fujitsu Limited.
[****]Presently, with Nomura Research Institute, Ltd.
[*****]Presently, with AgIC Inc.
a) E-mail: szuk@ibm.jp.com
b) E-mail: kuroiwar@nttdata.co.jp
c) E-mail: yrpnov20@gmail.com
d) E-mail: shumpei.0601@gmail.com
e) E-mail: shinya@agic.cc
f) E-mail: mine@gavo.t.u-tokyo.ac.jp
g) E-mail: hirose@gavo.t.u-tokyo.ac.jp

DOI: 10.1587/transinf.2016AWI0004

alect, an accent phrase has at most one mora, the next mora of which has a clearly lower pitch. We call the mora the “nucleus”. With the accent nucleus position, we can define accent types for accent phrases using natural numbers; when Nth mora is an accent nucleus, we define it as accent type N. If accent phrases do not have any accent nucleus, we define it as accent type 0. In addition to the pitch drop at the accent nucleus, rapid pitch rising occurs from the first mora to the second mora except accent type 1.

These definitions are not perfect definitions to explain all the accent phenomena concerning Tokyo dialect. But we use these definitions for convenience. With the definition, the estimation of accent sandhi can be divided into two tasks: (1) a task to estimate accent phrase boundaries from morpheme sequences and (2) a task to estimate accent types for accent phrases.

3. Rule-Based Approach

3.1 Rule-Based Accent Phrase Boundary Estimation

A rule-based method to estimate accent phrase boundaries from morpheme sequences is implemented in Open JTalk [2]. Open JTalk version 1.05 uses the rules in Table 1. The rules only depend on the part-of-speech (POS) of morphemes.

3.2 Rule-Based Accent Type Estimation

Sagisuka’s rule, which is well known, estimates accent types from accent phrases [3]. The rule first annotates all the morphemes with accent modification types. Using the accent modification types, POS, and position of accent nucleus for isolated words, Sagisaka’s rule estimates the accent types for accent phrases. For example, the accent phrase of “The Univeristy of Tokyo”, which consists of Japanese two words “Tokyo / Noun” and “University / Noun / accent mode C2”, applies a rule “For two adjacent nouns, if the accent mode of latter nouns is C2, accent nucleus is on the first mora of the latter word”. This rule brings the correct accent type 5, as shown in Fig. 2.

Other rules are often applied for numeral expressions because Sagisuka’s rule does not handle such expressions. For example, Miyazaki’s rule is often used [4]. In Open JTalk 1.05, Sagisuka’s rule and their own rules for numeral expressions are implemented.

However, the rule-based approach is not perfect because many exceptions exist. Our group tried to improve these rules, but could not achieve substantial improvement [5], [6].

Table 1 Conditions to find accent phrase boundaries used in Open JTalk version 1.05. If these conditions are matched, the rule based system detects accent phrase boundaries.

POS of the previous word	POS of the word
adjective, adjectival verb, verb, or suffix	noun
verb	adjective
self-sufficient word	ancillary word
one side is an adverb, conjunction, prenominal adjective, or symbol	

4. Database

A Japanese text database with accent labels in sentences was not available, although some papers reported machine learning based systems for accent estimation outperformed rule-based systems [7]–[10]. Thus, we built a publicly available database [11], [12]. Our sentence-level accent database includes 6,334 sentences used in the JNAS corpus. We did morphological analysis with UniDic for the sentences, and manually modified errors of pronunciation estimation. Next, we labeled accent phrase boundaries and an accent nucleus position for each accent phrase in sentences.

The database has high consistency. Because the accent varies depending on dialects, individuals, and reading speed, we selected only one labeler who grew up in Tokyo and had an excellent sense of relative pitch. We got her to read sentences with a speed of 7 mora per seconds, which is widely accepted as a speed for TTS. Another person who grew up in Tokyo checked her accent labels to ensure no errors were made.

The database is available for people who have a license for JNAS or S-JNAS [13], [14]. Readers interested in using the database should contact the authors.

5. Accent Sandhi Estimation with CRF

5.1 Estimation of Accent Phrase Boundaries with CRF

We formulated a task to estimate accent phrase boundaries as binary label sequence estimation for morphology sequences, where the labels represents whether or not accent phrase boundaries are just before the morphology.

We use CRF with features in Table 2 to estimate the binary label sequence [16]. The most import feature template is POS, which is also used for rule-based estimation. Some attributes of morphemes labeled in UniDic were also used as feature templates.

Estimating accent phrase boundaries is especially difficult for adjacent nouns. For example, see Fig. 3. If we can use POS information only, we cannot estimate this kind of boundary correctly. To deal with this problem, we introduced n-gram frequency based scores as features for CRF.

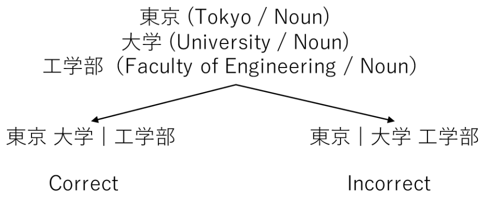


Fig. 3 Accent boundary estimation for three adjacent morphemes: Tokyo, university, and faculty-of-engineering. | represents an accent phrase boundary. Even though all the POSs are nouns, a correct boundary and an incorrect boundary were evident.

Table 2 Feature templates for CRF-based accent phrase boundary estimation. See [15] for the definition of the features. We used these features of five adjacent words (from two before to two words after) if available. We digitalized numerical numbers into 1, 2, 3, 4, and, 5 by quantiles for *j* to *m*.

Index	Feature templates
a	POS
b	Pairs of lemma, pronBase, and cType
c	cType
d	cForm
e	goshu
f	iType
g	aType
h	aModType
i	Binary label whether Bunsetsu boundary is estimated before the word
j	Bigram frequency
k	*j* divided by unigram frequency of the previous word
l	*j* divided by unigram frequency of the word
m	*j* divided by unigram frequency both of the previous word and the word

Table 3 Definitions of labels for accent relative change.

Label	definisions
Vanish	the isolated word has a nucleus, but the nucleus vanished in the sentence
Remain	the isolated word has a nucleus, and the nucleus remained in the sentence
Never	the isolated word is 0 type, and the word in the sentence also does not have a nucleus
Before	the isolated word has a nucleus, and the nucleus moved to one mora before the sentence
Last	The last mora of the word in the sentence is an accent nucleus
First	The first mora of the word in the sentence is an accent nucleus
Penultimate	The second last mora of the word in the sentence is an accent nucleus
After	the accent nucleus moved to the mora just after from the accent nucleus position of the isolated words
Second	the accent nucleus moved to the second mora after from the accent nucleus position of the isolated words
Third	the accent nucleus moved to the third mora after from the accent nucleus position of the isolated words

These scores are expected to be effective because accent boundaries do not tend to appear as boundaries between two words that often appear continuously.

We did further feature engineering. Our implementation provides more details [26].

5.2 Estimation of Accent Nucleus Positions with CRF

We formulated the task to estimate accent nucleus positions as a "label for accent relative change" in sequence estimation for the morphology sequence of accent phrases. The labels for accent relative change are defined as Table 3.

These definitions did not apply to the morphology in a few cases. In usch cases, we simply removed the data from the training data or added "Fourth" label, "Fifth" label, and so on. For the training phase, if a word accent nucleus position met multiple conditions of these definitions, the former label was applied. For the testing phase, after estimating the labels for accent for each morphemes in an accent phrase, we saw the labels from left to right and regarded the first accent nucleus as the accent nucleus for the accent phrase. With this rule, estimating labels for accent relative change for morphemes in accent phrases enables estimating accent nucleus positions.

We used CRF for modeling with features in Table 4. POS, accent type of isolated words, the number of mora, and accent modification type, which are used for rule-based systems, are also used as feature templates. In addition, we estimated the labels for accent relative change using Sagisaka's rule and Miyazaki's rule and used them as feature templates.

To further leverage Sagisaka's rule, we used this classification that is used in Sagisaka's rule [18]. Sagisaka's rule classifies morphemes into 4 types: (1) the isolated word has an accent nucleus on the last mora (2) the isolated word has an accent nucleus on the second last mora (3) the isolated word has an accent nucleus on the other mora (4) the isolated word does not have an accent nucleus. We used this classification as feature templates.

For loanwords, the accent type is known to depend on the number of mora being less than three or not, and whether it includes a heavy syllable. We added these features for loanwords [19].

For numeral expressions, the accent is known to depend on the types of counter suffixes. We used the classification of counter suffixes as Fig. 4, and used them as feature templates.

We did further feature engineering. Again, our implementation provides more details [26].

6. Experiments

6.1 Experimental Conditions

We used MeCab version 0.993 [20], CaboCha version 0.62 [21], UniDic version 1.3.12 [15] and its attached models for morphology analysis, pronunciation estimation, Bunsetsu boundary estimation, and named entity tags defined by information retrieval and extraction exercise (IREX). These

Table 4 Feature templates for estimating accent types. See [15] for a definition of the features. We used these feature of five words (from two before to two words after) if available.

Index	Feature templates
a	POS
b	aType
c	the number of mora
d	aConType for verbs
e	aConType for adverbs
f	aConType for nouns
g	aModType
h	aType after conjugation
i	label for accent relative change by rules
j	classification of Sagisaka's rule
k	orth
l	pron
m	cType
n	cForm
o	lForm
p	goshu
q	iConType
r	Binary label on whether or not the word is the first word in the accent phrase
s	the number of words in the accent phrase
t	Named entity tag defined by IREX [17]
u	Binary label on whether or not the word has exactly two mora
v	pair of *u* and goshu
w	Binary label on whether or not the word includes heavy syllables
x	the first mora of the word
y	the second first mora of the word
z	the mora just before the accent nucleus
A	the mora of the accent nucleus
B	the mora just after the accent nucleus
C	the last mora of the word
D	the second last mora of the word

a	個，位，時，分 (ふん)，時間，歳，羽，通り，斤，層，アール，センチ，キロ，ドル，度 (ど: 温度，角度)，階，球，巡，乗，週，人前，敗，着 (到着)，度目，代目，貫目，幕目，日目，球目，丁目，畳，ヶ月
b	問，台，軒，票，町，艘，代，枚，名，面，本，枚，丁
c	升
d	年 (ねん)，段 (階段)，番
e	貫，版，銭，回，点，巻
f	尺，着 (衣服)，角
g	円
h	曲，石 (こく)，匹，冊，足，拍，脚，局，発
i	合
j	度 (ど: 回数)
k	人
l	月 (がつ)，日 (にち)
m	寸

Fig. 4 Classification table for couner suffixes.

estimations had some errors, especially for pronunciation estimation. Our objective was to see the effect of accent estimation for TTS, so we removed sentences including pronunciation estimation errors in these experiments. With this processing, the number of sentences decreased from 6,334 to 4,785. We divided the 4,785 sentences into 3,786 sentences for training and 999 sentences for evaluation. The training data included 66,048 morphologies and 25,542 accent phrases. The evaluation data included 17,801 morphologies and 7,641 accent phrases.

Table 5 Results of accent phrase boundary estimation.

	Precision	Recall	F-value
Rule	89.1%	88.7%	88.9
CRF	97.4%	90.5%	**93.8**

We used CRF++ version 0.57 for implementation [22]. We used, for feature templates, the templates in Table 2 for accent phrase boundary estimation and thouse in Table 4 for accent nucleus position estimation. To calculate n-gram statistics, we used Japanese Wikipedia articles as of Apr. 10, 2012. We applied three-fold cross validation and used 0.1 for the accent boundary estimation and 0.8 for the accent nucleus position estimation for the regularization parameters. We trained the final model with the all of the training data with the regularization parameters. The implementation and trained models are publicly available [26].

6.2 Estimation of Accent Phrase Boundaries

We compared the proposed CRF-based approach and the rule-based approach. We used the rules defined in Table 1 for the rule-based system.

Table 5 shows the results. The proposed CRF-based system improved both the precision and recall of the rule-based system. Our system achieved an absolute 5 point improvement in F-values.

Table 6 shows the breakdown of the results for adja-

Table 6 Results for the case of compound nouns comprising two words.

	Precision	Recall	F-value
Rule	N/A	0%	N/A
CRF	65.2%	88.7%	**74.5**
CRF w/o *n*-gram	62.7%	85.4%	72.3

Table 7 Results of accent sandhi estimation.

Accent phrase	Accent type	Accuracy
Correct	Rule	90.30%
Correct	CRF	**97.11%**
Rule	Rule	87.48%
Rule	CRF	**94.48%**
CRF	Rule	87.61%
CRF	CRF	**94.66%**

cent nouns. The evaluation data consisted of 1,760 morphology boundaries and 606 accent phrase boundaries. To see the effect of using *n*-gram based feature templates, we prepared another CRF without *n*-gram based features. The rule-based system could not estimate these boundaries completely because no rule existed to estimate the boundary between adjacent nouns. However, CRF had much better results. The comparison of CRF and CRF without *n*-gram features yielded a 2.2 point absolute improvement after introducing *n*-gram features.

6.3 Accent Nucleus Position Estimation

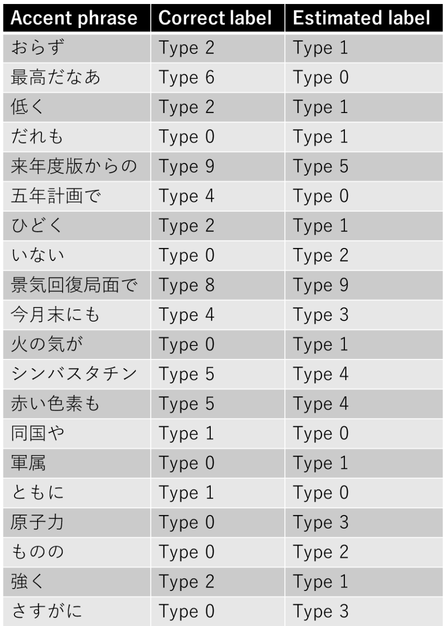
We compared our CRF-based approach and the rule-based approach. We implemented Sagisaka's rule and Miyazaki's rule for the rule-based system. We used three types for accent phrase boundaries: correct boundaries, estimated boundaries using the rule-based system, and the estimated boundaries using the CRF-based system. When we used correct accent boundaries, 7,641 accent phrases appeared in the evaluation data. Even if the estimated accent boundaries were not correct, we used the references of accent nucleus positions based on correct accent boundaries so that the final TTS system outputted natural pitch patterns while ignoring the errors of accent phrase boundaries.

Table 7 shows the results. our CRF-based system outperformed the rule-based system with all kinds of accent boundary estimation. When we applied the rule-based system both for accent phrase boundary estimation and nucleus position estimation, the correct answer rate was 87.48%, and when we applied the CRF-based system for both of them, correct error rate was 94.66%. This represents a relative 57% reduction in error.

When we used correct accent phrase boundaries, the correct answer rate reached 97.11%. Figure 5 shows 20 randomly selected of the errors. We have not done any quantitative analysis, but native Japanese speakers may feel that not only the correct accent type but also the estimated accent type are acceptable for some cases.

6.4 TTS Evaluation

We evaluated our system through TTS. We synthesized two types of speech with the rule-based accent estimation and the CRF-based accent estimation and compared them using a listening test.



Accent phrase	Correct label	Estimated label
おらず	Type 2	Type 1
最高だなあ	Type 6	Type 0
低く	Type 2	Type 1
だれも	Type 0	Type 1
来年度版からの	Type 9	Type 5
五年計画で	Type 4	Type 0
ひどく	Type 2	Type 1
いない	Type 0	Type 2
景気回復局面で	Type 8	Type 9
今月末にも	Type 4	Type 3
火の気が	Type 0	Type 1
シンバスタチン	Type 5	Type 4
赤い色素も	Type 5	Type 4
同国や	Type 1	Type 0
軍属	Type 0	Type 1
ともに	Type 1	Type 0
原子力	Type 0	Type 3
ものの	Type 0	Type 2
強く	Type 2	Type 1
さすがに	Type 0	Type 3

Fig. 5 Examples of errors of accent type estimation when correct accent phrase boundaries are given.

We used the HMM-based speech synthesis system with Open JTalk version 1.05, hts_engine API version 1.06 [24], HTS Voice Mei (Normal) version 1.1 [25] for TTS system. The sampling frequency was 48kHz, the frame period was 240 points, and the all-pass constant was 0.55. Because Open JTalk used the NAIST Japanese Dictionary [23], which was sometimes different from UniDic, we selected 873 sentences from 999 evaluation sentences where pronunciation estimation results perfectly corresponded. We selected 50 sentences from the 873 sentences randomly and used them for the listening test.

We used implementation of Open JTalk for the rule-based accent estimation system. This implementation was very similar to our implementation of the rule-based system. In our system, we only replaced the context labeling of HTS Voice for accent boundaries and the accent nucleus position using the estimated results. Other context labeling was the same.

Twelve people who had been lived in Tokyo for more than two years and usually spoke Tokyo dialect were used as listeners. The accent labeler was not included with the twelve people. The listeners used headphones and compared two voices, selecting the one that had better naturalness. We randomly changed the order of the two voices.

Figure 6 shows the results. In many sentences, our method was judged as more natural. A t-test was significant at $p < 0.01$. Thus, we can say that our system can improve TTS naturalness.

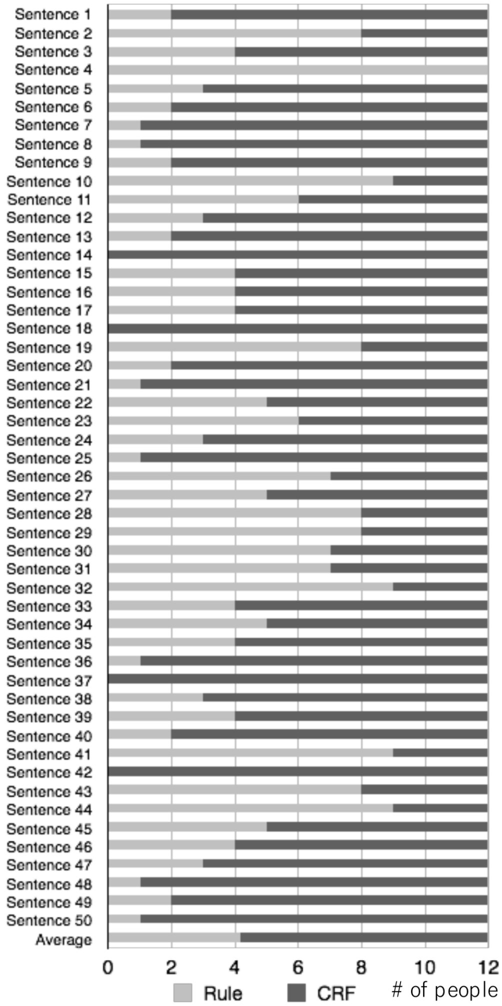


Fig. 6 Results of the paired comparison test.

7. Related Works

Nagano et al. proposed doing word segmentation, pronunciation estimation, and accent estimation simultaneously by using pairs of words, POS, pronunciation, and accent type as a token of *n*-gram [7], [8]. Our proposed system has to do them separately. One of the differences from Nagano's study is that we used a CRF-based discriminative model and did feature engineering, which is difficult for *n*-gram based modeling.

8. Conclusion

In this paper, we proposed a method to estimate accent phrase boundaries and accent types using CRF to improve the naturalness of Japanese TTS. We built a text database with accent labels in sentences. We achieved a relative 57% improvement from a rule-based system in accent estimation. The results of a listening test also show that the proposed method significantly improves the naturalness of Japanese TTS.

References

[1] S. Seto, M. Morita, T. Kagoshima, and M. Akamine, "Automatic rule generation for linguistic features analysis using inductive learning technique: Linguistic features analysis in TOS drive TTS system," Proc. 5th International Conference on Spoken Language Processing (ICSLP), pp.1059–1063, 1998.

[2] Open JTalk, http://open-jtalk.sourceforge.net/

[3] Y. Sagisaka and H. Sato, "Accentuation rules for Japanese word concatenation," IEICE Trans. Inf. & Syst. (Japanese Edition), vol.J66-D, no.7, pp.849–856, July 1983.

[4] M. Miyazaki, "Accent rules for numerical expressions for Japanese speech synthesis," Journal of Information Processing Society of Japan (IPSJ), vol.25, no.6, pp.1035–1043, 1984. [In Japanese]

[5] R. Kuroiwa, N. Minematsu, and K. Hirose, "Improvements of rules for Japanese accent sandhi using conjugative suffix," Annual Conference of Natural Language Processing in Japan, pp.995–998, 2006. [In Japanese]

[6] N. Minematsu, K. Ryuji, and K. Hirose, "Automatic estimation of accentual attribute values of words for accent sandhi rules of Japanese text-to-speech conversion," IEICE Trans. Inf. & Syst., vol.E86-D, no.3, pp.550–557, March 2003.

[7] T. Nagano, S. Mori, and M. Nishimura, "An N-gram-based approach to phoneme and accent estimation for TTS," Journal of Information Processing Society of Japan (IPSJ), vol.47, no.6, pp.1793–1801, 2006. [In Japanese]

[8] T. Nagano, R. Tachibana, and M. Nishimura, "Corpus-based text-to-speech front-end for Japanese," IEICE Trans. Inf. & Syst. (Japanese Edition), vol.J93-D, no.10, pp.2096–2106, Oct. 2010.

[9] K. Suzuki, M. Yamamoto, C. Kook, and Y. Yamashita, "Automatic accent type labeling for spoken sentences based on statistical methods using accentuation rules," Journal of Acoustic Society of Japan (ASJ), vol.66, no.10, pp.487–496, 2010. [In Japanese]

[10] M. Yamamoto, C. Kook, and Y. Yamashita, "Automatic prediction of accent phrase boundaries using linguistic and F0 information," IEICE Technical Report, SP2010-109, 2011. [In Japanese]

[11] R. Kuroiwa, Improvement of rules and statstical approach for accent sandhi estimation for Japanese speech synthesis, Master Thesis of the University of Tokyo, 2007. [In Japanese]

[12] N. Minematsu, R. Kuroiwa, K. Hirose, and M. Watanabe, "CRF-based statistical learning of Japanese accent sandhi for developing Japanese text-to-speech synthesis systems," Proc. Sixth ISCA Workshop on Speech Synthesis Workshop Proceedings, 2007.

[13] ASJ Japanese Newspaper Article Sentences Read Speech Corpus (JNAS), http://research.nii.ac.jp/src/JNAS.html

[14] Japanese Newspaper Article Sentences Read Speech Corpus of the Aged (S-JNAS), http://research.nii.ac.jp/src/S-JNAS.html

[15] Y. Den, T. Ogiso, H. Ogura, A. Yamada, N. Minematsu, K. Uchimoto, and H. Koiso, "The development of an electronic dictionary for morphological analysis and its application to Japanese corpus linguistics," Science of Japanese, vol.22, pp.101–122, 2007. [In Japanese]

[16] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random felds: Probabilistic models for segmenting and labeling sequence data," Proc. 18th International Conference on Machine Learning (ICML), pp.282–289, 2001.

[17] S. Sekine and H. Isahara, "IREX: IR and IE evaluation project in Japanese," Proc. LREC 2000.

[18] K. Innami, Error analysis and improvement for CRF based Japanese accent sandhi estimation, Master Thesis of the University of Tokyo, 2009. [In Japanese]

[19] S. Kobayashi, Improvement of CRF based accent sandhi estimation and application for Japanese-language education, Master Thesis of the University of Tokyo, 2012. [In Japanese]

[20] MeCab, http://code.google.com/p/mecab/

[21] CaboCha, http://code.google.com/p/cabocha/

[22] CRF++, https://code.google.com/p/crfpp/
[23] NAIST Japanese Dictionary, http://sourceforge.jp/projects/naist-jdic/
[24] hts_engine API, http://hts-engine.sourceforge.net/
[25] MMDAgent, http://www.mmdagent.jp/
[26] TASET, https://sites.google.com/site/suzukimasayuki/accent

Masayuki Suzuki received his B.Eng., M.Eng., Ph.D. degrees in electrical engineering and information systems from the University of Tokyo, in 2008, 2010, and 2013. Since 2013, he has been working at IBM Research - Tokyo. His research interests include speech and spoken language processing. He is a member of the Acoustical Society of Japan (ASJ), the Institute of Electronics, Information and Communications Engineers (IEICE), IEEE, and ISCA. He received the Awaya Award from the ASJ in 2013.

Ryo Kuroiwa received the Master degree in information science and technology from the University of Tokyo in 2007. He is now with NTT DATA Corporation.

Keisuke Innami received the Master degree in frontier sciences from the University of Tokyo in 2009. He is now with Fujitsu Limited.

Shumpei Kobayashi received the Master degree in information science and technology from the University of Tokyo in 2012. He is now with Nomura Research Institute, Ltd.

Shinya Shimizu recieved the Master degree in information science and technology from the University of Tokyo in 2012. He is now with AgIC Inc.

Nobuaki Minematsu received the Ph.D. degree in electronic engineering in 1995 from the University of Tokyo. In 1995, he became an assistant researcher with the Department of Information and Computer Science, Toyohashi University of Technology, and in 2000, he was an associate professor with the Graduate School of Engineering, the University of Tokyo. Since 2012, he has been a professor with the Graduate School of Engineering, the University of Tokyo. From 2002 to 2003, he was a visiting researcher at Kungl Tekniska Högskolan (KTH), Sweden. He has a wide interest in speech from science to engineering, including phonetics, phonology, language learning, speech perception, speech analysis, speech recognition, speech synthesis, and speech applications. Dr. Minematsu is a member of IEEE, ISCA, IPA, the Institute of Electronics, Information and Communication Engineering, the Acoustical Society of Japan, the Information Processing Society of Japan, the Japanese Society for Artificial Intelligence, and the Phonetic Society of Japan. He received best paper awards from Research Institute of Signal Processing in 2007 and 2013.

Keikichi Hirose received the B.E. degree in electrical engineering in 1972, and the M.E. and Ph.D. degrees in electronic engineering respectively in 1974 and 1977 from the University of Tokyo. From 1977, he was a faculty member at the University of Tokyo, and was a professor of the Department of Electronic Engineering from 1994, and a professor of the Department of Information and Communication Engineering, Graduate School of Information Science and Technology, from 2004. In 2015, he retired from the University of Tokyo, and was received Professor of Emeritus title. He is also a visiting professor of National Institute of Informatics from 2015. From March 1987 to January 1988, he was Visiting Scientist at the Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, U.S.A. He has been engaged in a wide range of research on spoken language processing, including analysis, synthesis, recognition, dialogue systems, and computer-assisted language learning. From 2000 to 2004, he was Principal Investigator of the national project "Realization of advanced spoken language information processing utilizing prosodic features." He served as the general chair for INTERSPEECH 2010, Makuhari, Japan. Since 2010, he serves as the Chair of ISCA (International Speech Communication Association) Special Interest Group on Speech Prosody (SProSIG). He became an honorary member, Polish Phonetic Association, in 2013. In 2015, he was honored as a Named Person of Merit in Science and Technology by the Mayor of Tokyo. He is a member of a number of academic societies, including ISCA (Board member), IEEE, Acoustical Society of America, Acoustical Society of Japan, Information Processing Society of Japan, and Research Institute of Signal Processing Japan (Board member).