8th ISCA Speech Synthesis Workshop • August 31 – September 2, 2013 • Barcelona, Spain

Context labels based on "bunsetsu" for HMM-based speech synthesis of Japanese

Hiroya Hashimoto[1], *Keikichi Hirose*[1], *Nobuaki Minematsu*[2]

[1]Department of Information and Communication Engineering
[2]Department of Electrical Engineering and Information Systems,
the University of Tokyo, Japan
{hiroya, hirose, mine}@gavo.t.u-tokyo.ac.jp

Abstract

A new set of context labels was developed for HMM-based speech synthesis of Japanese. The conventional labels include those directly related to sentence length, such as number of "mora" and order of breath group in a sentence. When reading a sentence, it is unlikely that we count its total length before utterance. Also a set of increased number of labels is required to handle sentences with various lengths, resulting in a less efficient clustering process. Furthermore, labels related to prosody are mostly designed based on the unit "accent phrase," whose definition is somewhat unclear; it is not uniquely defined for a given sentence, but also is affected by other factors such as speaker identity, speaking rate, and utterance style. Accent phrase boundaries may be labeled differently for utterances of the same content, and this situation affects other labels, because of numerical labeling scheme counted from the sentence/breath-group initial. In the proposed labels, "bunsetsu" is used instead. Also, we only view its relations with preceding and following "bunsetsu's." Thus labels not related to the sentence lengths are obtained, with easier automatic prediction only from sentence representations. Validity of the proposed labels was shown through speech synthesis experiments.
Index Terms: speech synthesis, context labels, linguistic information

1. Introduction

Recently, statistical framework, such as hidden Markov modeling, has been successfully introduced to analysis-synthesis-based speech synthesis systems[1]. Although there still are some degradations in speech quality as compared to waveform concatenation methods, HMM-based speech synthesis is now widely used, since it can generate speech in various voice qualities and speaking styles from a very limited speech corpus through adaptation/conversion techniques[2, 3]. Although HMM's are commonly used in speech recognition, they are differently organized in speech synthesis. In the case of speech recognition, since the aim is to recognize phonemes, one HMM is trained for each phoneme separately for surrounding phonemes. Other factors affecting phoneme features, such as positions in an utterance, accent types, etc. are not counted. However, in the case of speech synthesis, variations of phoneme features need to be realized as correctly as possible. Therefore, various contextual factors are taken into account, and a number of context labels are prepared to represent these factors. Since combinations of these labels are huge, we faced to the problem of data sparseness, if we try to train an HMM for each combination. Grouping of conditions are commonly done before HMM training to solve the problem. Also several methods have been developed to reduce context numbers, including one to select important labels by finding relation of labels using Bayesian networks[4], and one to use F0 digitized for each phoneme instead of accent types as prosody contexts[5]. Resulting synthetic speech, however, includes some unnaturalness. One possible reason for this situation resides in the design of the context labels.

The context labels widely used for Japanese speech synthesis are those used in HTS[6], a well-known HMM-based speech synthesizer. They include following two problems. First, labels related to prosody are designed based on "accent phrase," whose definition has an ambiguity and cannot be decided only from text. It may be subject to change by utterance speeds and styles. The second problem is the sequential numbering from the sentence/breath-group initial adopted in some labels. In order to cope with sentences with arbitrary lengths, a large number of labels are required. Moreover, labels can be differently assigned for the same phrases (with similar prosodic features), but in different sentences. This label ambiguity may also happen even for the same sentences, when they have different accent phrase boundaries.

In order to solve this situation, we newly developed a set of context labels based on "bunsetsu" instead of "accent phrase." "Bunsetsu" is defined as a basic unit of Japanese syntax and pronunciation consisting of content word(s) followed or not followed by particles. "Bunsetsu" boundaries can be predicted only from text with high performance. Furthermore, we avoided to use positions in sentence/breath-group.

The rest of the paper is organized as follows: Newly proposed context labels are presented and compared with conventional context labels in section 2. In section 3, the proposed context labels are evaluated through listening test of synthetic speech. Section 4 concludes the paper with some discussions.

2. Context labels

2.1. Context labels for HTS (conventional context labels)

Table 1 shows context labels adopted in HTS Japanese speech synthesizer. Labels related to prosody are designed based on "accent phrase," which is defined as an utterance unit with a pair of rise and fall of fundamental frequency (F0) contour (an accent component). Mora with F0 fall is crucial for perception of Japanese lexical accent and is called an accent nucleus. The context labels are designed assuming one accent component in each accent phrase, though an accent phrase can have a minor accent component, called secondary accent. Labeling of accent

Utterance 1: yamano/ueno/mukooni/kireina/hanaga. . .
1 2 3 4 5
Utterance 2: yamanoueno/mukooni/kireina/hanaga. . .
1 2 3 4
(Over the hill top, there are beautiful floors. . .)

Figure 1: An example of accent phrase labeling

phrases for speech corpus of HMM training usually conducted manually by labelers referring to texts and speech sounds. Certain inconsistencies are unavoidable in the labeling process. Especially, it is often difficult to tell an minor F0 movement being as "reduced" accent by de-focusing, secondary accent, or no accent component. Although there have been several attempts for automatic extraction of accent phrases, their performances are not high enough. To begin with, there are number of cases hard to exactly tell whether a (linguistic) phrase consists of one accent phrase or two (or more) accent phrases. The cases may increase when we handle spontaneous speech. When a sentence is uttered in a different speaking style or by a different speaker, the accent phrases may change, because they are units of "utterance." Regardless of these accent phrase ambiguities, accent phrases are predicted only from texts in HMM-based speech synthesis. This situation may not cause a serious problem when handling a speech corpus carefully (and thus consistently) uttered in reading style by a professional speaker. Because of the cost of labeling, however the same accent labeling is often used for a new speech corpus by a different speaker or in a different speaking style. This may increase errors in accent phrase labeling.

Since accent phrases are sequentially numbered in a breath group, a difference in accent phrase labeling may spread to other parts as shown in Fig 1. Symbol "/" indicates accent phrase boundary. Context label "position of the current phrase in the current breath group" (in Table 1) is totally differently labeled, though the difference between two utterances is the existence of accent phrase boundary after "yamano" in utterance 1. Also since the context labels include those on lengths of sentence, breath group, and accent phrase, which are counted by number of morae, unnecessarily large numbers need to be prepared to cope with various sentences and speaking styles. Although these labels are usually discarded (or summarized) through context clustering, some of these labels sometimes degrade synthetic speech quality.

2.2. Proposed context labels

In order to solve the problems listed in the previous section, a new set of context labels is constructed as shown in Table 2. Two labels ID1 and ID2 are prepared to represent POS (part-of-speech) and S-POS (supplemental POS), respectively, based on the Unidic Japanese dictionary for morpheme analysis[7]. ID1 includes following parts-of-speech: verb, noun, adjective, adjectival verb, adnominal, adverb, pronoun, interjection, particle, auxiliary verb, prefix, suffix, sentence initial, short pause, and sentence end. The last three items are included, since pauses largely affect other prosodic features. ID2 is supplemental to ID1 and indicates the role of the word. It includes: "can be used as content word," "can be used as particle," general, common noun, numeral, proper noun, noun like, verb like, adjective like, adjectival verb like, nominative particle, "particle that attaches to a phrase and acts on the whole phrase," adverbial particle, conjunctive particle, binding particle, sentence-end particle, stem of auxiliary verb, "tari" conjugation, and filler.

Table 1: Context labels adopted in Japanese HTS

Previous phoneme identity
Current phoneme identity
Next phoneme identity
Position of the current mora in the current accent phrase
Difference between accent type and position of the current mora
POS of the previous word
Inflected form of the previous word
Conjugation type of the previous word
POS of the current word
Inflected form of the current word
Conjugation type of the current word
POS of the next word
Inflected form of the next word
Conjugation type of the next word
Number of morae of the previous accent phrase
Accent type of the previous accent phrase
Connection intensity between the previous accent phrase and the current accent phrase
Pause existence between the previous accent phrase and the current accent phrase
Number of morae in the current accent phrase
Accent type in the current accent phrase
Connection intensity between the previous accent phrase and the next accent phrase
Position of the current accent phrase in the current breath group
Interrogative sentence or not
Number of morae of the next accent phrase
Accent type of the next accent phrase
Connection intensity between the next accent phrase and the current accent phrase
Pause existence between the next accent phrase and the current accent phrase
Number of morae of the previous breath group
Number of morae of the current breath group
Position of the current breath group in the sentence
Number of morae of the next breath group
Number of morae of the sentence

The new labels have following three major differences from those of HTS.

i) "Bunsetsu" is used instead of "accent phrase." Since "bunsetsu" is a grammatically defined unit, it can be identified uniquely from text. Also "very long" samples found in accent phrases do not occur, and maximum number can be set small for "bunsetsu" length counted in mora unit.

ii) High (1) or Low (0) is assigned to each moraic F0 pattern instead of accent types. Japanese word accent types are often stylized with high and low patterns of F0 contours in mora unit. High-low assignment can be automatically done for each mora when accent types are given. Due to accent concatenation, Japanese word accent in continuous speech may change from that of isolated utterance. When two content words concatenates, they are uttered together in one accent type. However, when we emphasize one of two words, concatenation may not happen; two words are uttered with their original accent types. If

Table 2: Context labels of the proposed method

Previous phoneme identity
Current phoneme identity
Next phoneme identity
F0 level of the previous mora (0:Low, 1:High)
F0 level of the current mora (0:Low, 1:High)
F0 level of the next mora (0:Low, 1:High)
Position of the current mora in the current word (counted from word initial)
Position of the current mora in the current word (counted from word end)
Position of the current mora in the current "bunsetsu" (counted from "bunsetsu" initial)
Position of the current mora in the current "bunsetsu" (counted from "bunsetsu" end)
Number of morae of the previous word
Number of morae of the current word
Number of morae of the next word
Number of morae of the previous "bunsetsu"
Number of morae of the current "bunsetsu"
Number of morae of the next "bunsetsu"
POS ID1 of the previous word
POS ID1 of the current word
POS ID1 of the next word
POS ID1 of the content word in the previous "bunsetsu"
POS ID1 of the content word in the current "bunsetsu"
POS ID1 of the content word in the next "bunsetsu"
S-POS ID2 of the previous word
S-POS ID2 of the current word
S-POS ID2 of the next word
S-POS ID2 of the content word in the previous "bunsetsu"
S-POS ID2 of the content word in the current "bunsetsu"
S-POS ID2 of the content word in the next "bunsetsu"
Whether the current mora consisting of only one short vowel or not
Whether the current mora containing long vowel or not

we use "accent type," these phenomena are explained in complex, but they can be simplified with high-low pattern representations. Secondary accents can also be labeled easily.

iii) Only relative positions are adopted. Absolute positions, such as position of breath group in sentence, and position of accent phrase in breath group, are not used. Total lengths of sentence and breath group are not used. These can reduce the total number of labels and can prevent labeling ambiguities affecting other parts.

There are minor differences in the labeling: a label identifying long vowel from singleton is included, and a label identifying interrogative sentence from declarative sentence is deleted. The second change is done, because no interrogative sentences are included in the corpus used in the experiment. As for the first one, we can also include long vowels in the phoneme set instead. (In HTS, a long vowel is represented by two short vowels. This sometimes causes confusions with two short vowels, which are not merged to a long vowel.)

Since high/low F0 level of a mora is tightly related to those of preceding and following morae, in the context clustering process, combinations of labels of "F0 levels of previous, current, and next morae" are included in the question sets. For instance, "previous (0) + current (1) + next (0)" is included.

3. Speech synthesis experiment

HMM-based speech synthesis is conducted for two cases, using HTS context labels and using proposed labels. Synthetic speech from the two cases is compared in their naturalness through a listening test.

3.1. Method

From ATR continuous speech corpus B set[8], utterances by male speaker (MMI) and female speaker (FTY) are selected for the synthesis experiment. (Speech syntheses for speaker MMI and speaker FTY are conducted.) Each speaker uttered 503 sentences, and 450 sentences are used for HMM training, with rest 53 sentences for testing. Speech signals are sampled at 16 kHz sampling rate, and STRAIGHT analysis[9] is used to extract spectral envelope, F0, and aperiodicity with 5-ms frame shift. Minimum and maximum values for F0 extraction are set to 120 Hz and 400 Hz for the female speaker, and 60 Hz and 250 Hz for the male speaker. The spectral envelope is converted to mel-cepstral coefficients using a recursion formula. The feature vector is 138 dimensional, consisting of 40 mel-cepstral coefficients including the 0th coefficient, the logarithm of fundamental frequency, 5 band-aperiodicity (0–1 kHz, 1–2 kHz, 2–4 kHz, 4–6 kHz, 6–8 kHz) and their delta and delta-delta coefficients. HMM with five-state left-to-right model topology is used. Output from each state is represented by a single Gaussian with diagonal covariance matrix. Context clustering is conducted using binary decision trees with MDL stop criterion. HMM training is conducted by HTS-2.1.

Context labels related to lexical accents are manually labeled, while those of part-of-speech are automatically labeled using open source Japanese parser "mecab"[10] with manual correction.

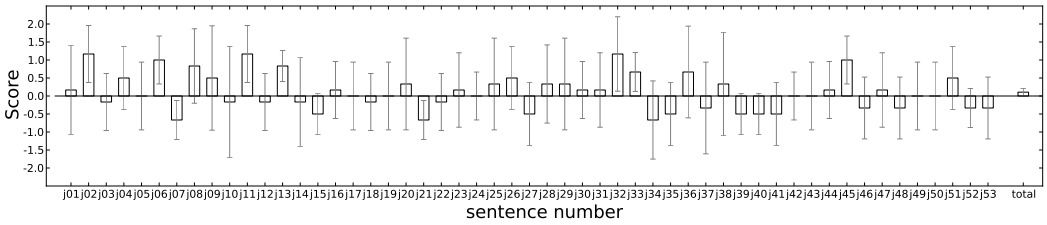
6 native speakers of Japanese evaluated the synthetic speech. They are asked to listen pairs of synthetic speech (one is by conventional labels and the other is by proposed labels), and select one on 5-scale scoring (2: one by proposed labels is clearly better than one by conventional labels, 1: one by proposed labels is better than one by conventional labels, 0: same, -1: one by conventional labels is better than one by proposed labels, -2: one by conventional labels is clearly better than one by proposed labels).

3.2. Result

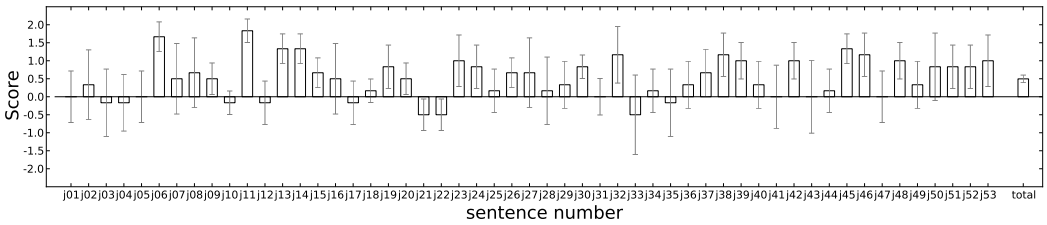
Results are shown in Fig. 2 with a confidence interval of 95%. The average score over the 53 test sentences is 0.109 with 0.106 confidence interval in significance level of 5% for FTY. and 0.497 with 0.105 for MMI. From the results, it is found that the proposed method improves the quality of the synthetic speech.

Fig. 3 compares generated F0 contours by the two sets of context labels. Some unnatural movements in the F0 contour by conventional context labels are settled by proposed context labels, indicating that the lexical accents can be well represented only by high-low F0 labeling of each mora. One of the major reasons of the degradation by the proposed labeling is the duration control. Inspection of decision trees for durations indicates that context labels on "number of morae" and on "position of mora" often appear near the top nodes for the proposed context labels, while they do not appear for the conventional context labels. Further studies on the labels are necessary from this viewpoint.

Combinations of "F0 levels of previous, current, and next morae" appear near the top node of the decision trees for funda-

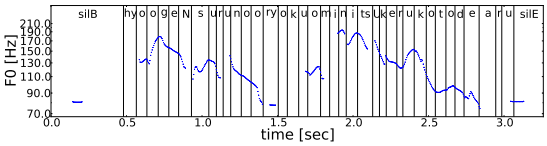


(a) FTY (female speaker)

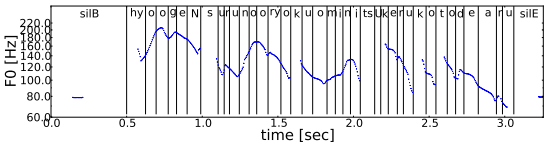


(b) MMI (male speaker)

Figure 2: Result of subjective test



(a) conventional



(b) proposed

Figure 3: Comparison of F0 contours for a Japanese sentence: hyoogeNsurunooryokuo minitsukerukotodearu (Is is to obtain an ability of expressing.). From top to bottom, F0 contour generated with conventional context labels, that generated with the proposed context labels. (Speaker MMI)

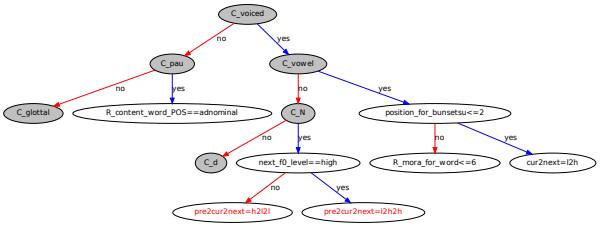


Figure 4: An example of decision tree for F0 (Speaker MMI, 2nd state). The nodes in red are combinations of "F0 levels of previous, current, and next morae."

mental frequencies as shown in Fig 4. This result indicates the correlations of labels. Further experiments are necessary to find the best combinations.

4. Conclusion and disucussion

A new set of context labels was constructed for HMM-based speech synthesis. Since the new labels are not using absolute positions of units in utterances, efficient and compact labeling is possible for speech corpus with various lengths. The labels also adopts "bunsetsu" instead of "accent phrase," enabling consistent labeling only from text. These features facilitate the HMM training process, and thus improve the synthetic speech quality, which is proved through a listening test of synthetic speech. The effect of the new labels may come clearer when handling long sentences, which are not included in the current speech corpus. (Maximum length of the sentence included in the current speech corpus is around 60 morae.)

Although, in the current experiment, the new labels and those of HTS are assigned manually, automatic labeling will be easier for the new labels. We already have conducted a preliminary speech synthesis experiment using speech corpus with automatically assigned labels, and confirmed that no apparent degradation observable for the new labels. We are now further improving the labels so that they can well handle various styles of speech.

The labeling scheme should be also beneficial to languages other than Japanese: for instance, in English HTS, the context labels include ones such as "position of the current syllable in the current word," "position of the current syllable in the current phrase," and "number of syllables in the utterance." These labels may cause problems similar to Japanese.

5. References

[1] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Speaker interpolation in HMM-based speech synthesis system," Proc. EUROSPEECH, pp. 2523–2526, 1997.

[2] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis," IEICE Trans. Inf. & Syst., vol. E88-D, no. 3, pp. 503–509, 2005.

[3] T. Nose, Y. Kato, and T. Kobayashi, "A speaker adaptation technique for MRHSMM-based style control of synthetic speech," Proc. ICASSP, pp. 833–836, 2007.

[4] Heng Lu, and Simon King, "Bayesian Networks to find relevant context features for HMM-based speech synthesis," Proc. INTERSPEECH, 2012.

[5] T. Nose, K. Ooki, T. Kobayashi, "HMM-based speech synthesis with unsupervised labeling of accentual context based on F0 quantization and average voice model," Proc. ICASSP, pp. 4622–4625, 2010.

[6] HTS, http://hts.sp.nitech.ac.jp/

[7] Unidic, http://sourceforge.jp/projects/unidic/

[8] A. Kurematsu, Kazuya Takeda, Yoshinori Sagisaka, Shigeru Katagiri, Hisao Kuwabara, and Kiyohiro Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," Speech Communication, vol. 9, pp. 357–363, 1990.

[9] H. Kawahara, I. Matsuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds," Speech Communication, vol. 27, no. 3–4, pp. 187–207, 1999.

[10] Mecab, http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html