Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6438–6443
Marseille, 11–16 May 2020

DNN-based Speech Synthesis Using Abundant Tags of Spontaneous Speech Corpus

Yuki Yamashita[1], Tomoki Koriyama[2], Yuki Saito[2], Shinnosuke Takamichi[2], Yusuke Ijima[3], Ryo Masumura[3], Hiroshi Saruwatari[2]
[1]Faculty of Engineering, The University of Tokyo,
[2]Graduate School of Information Science and Technology, The University of Tokyo,
[3]NTT Media Intelligence Laboratories, NTT Corporation,
[1,2] 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
yukiyama913@g.ecc.u-tokyo.ac.jp, tomoki_koriyama@ipc.i.u-tokyo.ac.jp

Abstract

In this paper, we investigate the effectiveness of using rich annotations in deep neural network (DNN)-based statistical speech synthesis. DNN-based frameworks typically use linguistic information as input features called context instead of directly using text. In such frameworks, we can synthesize not only reading-style speech but also speech with paralinguistic and nonlinguistic features by adding such information to the context. However, it is not clear what kind of information is crucial for reproducing paralinguistic and nonlinguistic features. Therefore, we investigate the effectiveness of rich tags in DNN-based speech synthesis according to the Corpus of Spontaneous Japanese (CSJ), which has a large amount of annotations on paralinguistic features such as prosody, disfluency, and morphological features. Experimental evaluation results shows that the reproducibility of paralinguistic features of synthetic speech was enhanced by adding such information as context.

Keywords: speech synthesis, context, spontaneous speech, annotation, deep neural network

1. Introduction

Speech synthesis is used in various applications such as smart speakers and robots, and situations such as announcements in public transportation and voice-overs of videos. A widely-studied speech synthesis is DNN-based one (Zen et al., 2013) in which the relationship between a text and an acoustic feature sequence is modeled by DNN. However, if we directly use text as an input of DNN, the training tends to be difficult. Instead, *context* is widely used as a model input. Context is defined as a variation factor of phonetic and prosodic features, and there are various contextual factors such as adjacent phones and stress in syllables. We encode contextual factors into a real-valued vector to enable DNN training.

In this study, we focus on the construction of context. Although general reading-style speech synthesis only uses linguistic information as a context, context is more flexible, namely, it is easy to feed various information into the input of DNN. For example, inputting a speaker vector is successful in multi-speaker modeling (Wu et al., 2015; Hojo et al., 2018). The emotion control of synthetic speech was achieved by using the degree of expressiveness as a contextual factor (An et al., 2017; Lorenzo-Trueba et al., 2018). It is also reported that local features such as emphasis can be used as the input of DNN (Wang et al., 2018).

In order to synthetize speech with a wide variety of paralinguistic features, adding such features as context is promising. However, too many input features often makes training difficult and causes overfitting problem. Therefore, in this paper, we investigate the effectiveness of rich tags in DNN-based speech synthesis according to the Corpus of Spontaneous Japanese (CSJ) (Maekawa et al., 2000). CSJ has a large amount of tags, which can be used as additional information for speech synthesis. Specifically, we choose contextual factors, such as tone, word, and disfluency, from the tags in XML files of CSJ core data, and encoded such factors into the input vectors of DNN-based speech synthesis system. By subjective evaluations, we show that it is important to use a detailed annotation in spontaneous speech synthesis.

2. Related Work

A similar study to this paper was performed for an hidden Markov model (HMM)-based speech synthesis (Koriyama et al., 2011). In this study, the tags in CSJ were used as context of HMMs. The context of HMM-based speech synthesis is a factor to construct a decision tree. Although experimental results show that tone information and phone prolongation were effective to enhance the naturalness of synthetic speech, others did not affect the performance. A possible reason is that decision trees cannot deal with complicated combination of contextual factors. Furthermore, since the size of decision tree is determined by the amount of training data, some contextual factors were scarcely used for the construction of a decision tree. We can expect to overcome this problem because DNN has an advantage to utilize complicated input features.

3. DNN-based Speech Synthesis

We describe the basics of DNN-based speech synthesis. There are two popular approaches, namely, the pipeline and end-to-end approaches. The pipeline approach consists of the models of text analysis and duration and acoustic feature models (Zen et al., 2013), whereas the end-to-end approach uses a single model, which generally predicts an acoustic feature sequence from a character sequence of text (Sotelo et al., 2017; Wang et al., 2017). In this study, we adopt the pipeline framework because the end-to-end model is difficult to train and the pipeline model still outperformed the end-to-end model in recent study (Yasuda et al., 2019).

The input features of duration and acoustic feature models are derived from contexts. Contexts represent linguistic information of the input text and are defined as the set of contextual factors that determine spectral and prosodic features of synthetic speech. For example, even for frames having the same phones, stress and unstressed phones have different sounds, and pronunciation depends on the preceding and succeeding phones. Contexts include not only categorical factors such as phone names and part-of-speech (POS) information but also numerical factors such as the length of phrases and the relative position from accented units. An example of contextual factors is shown in Sect. 4.

To utilize the contextual factors as the input of DNNs, the factors are converted into differentiable continuous values. Specifically, categorical contextual factors such as phone names are encoded to one-hot vectors. Moreover, we know that specific contextual factors are related. In such cases, we use a *question*, e.g., "Is a current phone a vowel?", and obtain binary values of 1 for "yes" and 0 for "no."

Numerical factors such as length and position can be directly used as the inputs of DNNs. However, it is reported that encoding such factors into binary vectors using questions is more effective than using the numerical factors as inputs directly (Dall et al., 2016). The questions for encoding numerical factors generally include inequality questions, e.g., "Is the factor more/less than a specific value?"

Moreover, frame position information is used to distinguish a frame from other frames that are in the same speech unit. Specifically, the time scale of the phone is normalized to $[0, 1]$ and the relative position of the frame is used as an input feature.

4. Contextual Factors for Japanese Reading-style Speech

Since the contextual factors depend on the languages and we use Japanese speech data for experiments, we explain the Japanese context set used in the demo script of HMM/DNN-based Speech Synthesis System (HTS) (Zen et al., 2007).

The Japanese context set uses five hierarchical speech units: utterance, breath group, accent phrase, mora, and phone as follows:

Utterance: The top layer, which is equivalent to a sentence in reading-style speech synthesis. Utterance includes duration information such as the numbers of breath groups, accent phrases, and moras.

Breath group: A speech unit defined by dividing an utterance by pauses. Intonation reset often occurs between breath groups. The contextual factors of breath groups are the position of the unit and the numbers of accent phrases and moras.

Accent phrase: Japanese is a pitch accent language, in which each mora is represented by a high/low pitch. A high-pitch mora that has a low pitch in the next mora is called an accent nucleus, which is important for the perception of Japanese words. One accent phrase has either one or no accent nucleus and the position of the accent nucleus mora determines the *accent type*. The contextual factors of the accent phrase include not only position information and the number of moras but also the accent type. The context of the accent phrase also contains information about the preceding and following pauses.

Mora: The mora is a commonly-used unit of Japanese, in which a word duration is controlled by the number of moras in the word. The contextual factors of a mora are composed of position information in an accent phrase and relative position information from accent nucleus moras.

Phone: The basic unit of speech. The identity of a phone is used as a contextual factor of the phone. Since some phones share features, such as voiced, vocalic, plosive, and nasal sounds, such features are used as questions.

To take the effect of adjacent units into account, we use the contextual factors of the previous and next speech units as well as those of current units. We refer to this context set as the baseline context.

5. Extended Context for Spontaneous Speech

To model spontaneous speech, which has much greater variation in utterances than reading-style speech, we extended the context set on the basis of the tags of CSJ (Maekawa et al., 2000). CSJ is designed for various purposes such as the analysis and modeling of spontaneous speech, and the *core data* in CSJ has a huge amount of manual annotation. In this section, we describe how to convert the tags into contextual factors, and we propose an extended context set derived from the tags of CSJ.

5.1. Tags included in XML files of CSJ

The XML files in CSJ includes detail tags (Maekawa et al., 2004) as shown in Fig. 1. The hierarchical structure of XML for the core CSJ is as follows:

```
<talk>
  <IPU>
    <LUW>
      <SUW>
        <TransSUW>
          <Mora> or <NonLinguisticSound>
            <Phoneme>
              <Phone>
                <XJToBILabelTone>
                <XJToBILabelWord>
                <XJToBILabelBreak>
                <XJToBILabelPrm>
```

Here, we describe how to use the elements of XML for the context set.

Talk: In CSJ, speech data is stored as a set of utterance sequences referred to as a "talk." Talks include academic presentation speech, simulated public speaking, and dialogs. The dialogs are interviews, free dialogs, and task-oriented conversations. The types of talk and speaker information can be obtained from this tag.

Figure 1: Example of XML in CSJ.

Inter-pause unit (IPU): For reading-style speech, in general, sentences are used as a unit of utterance. However, this is not appropriate for spontaneous speech because the end of the sentence is not always uttered. In CSJ, IPU is used as an utterance unit of transcriptions, in which 200 ms pauses are regarded as the boundaries of utterances.

Long-unit word (LUW), Short-unit word (SUW): Since Japanese is an agglutinative language, there is a high degree of freedom in the definition of "word" (Maekawa et al., 2014). Therefore, CSJ takes two types of unit in words: LUW and SUW. An SUW is the shortest unit defined by the dictionary UniDic (Den et al., 2008). An LUW is a longer unit representing compound words.

TransSUW: This tag is used to indicate disfluent utterances such as fillers, word fragments, and restatements.

Mora: A `<Mora>` tag has kana information. Mora tag can be used to count the number of moras in phrases.

NonLingisticSound: Nonlinguistic information such as breaths and laughing is labeled in this tag. This tag also includes a vowel-nasal filler denoted by "VN," which appears in the response utterances in dialogs.

Phoneme, Phone: `<Phone>` tags have phone-related annotation in detail, e.g., the beginning and end times and devoiced vowels. The number of phone entities used in this study is 58. The `<Phoneme>` tag is a group of `<Phone>` tags, which we ignore in this study.

XJToBILabel*: These tags are annotated at the times when X-JToBI (Maekawa et al., 2002) events appear. The tag `<XJToBILabelTone>` includes tone labels of accent (A), initial boundary tone (%L, %H), boundary pitch movement (L%, HL%, LH%, HLH%), other tags (LTBPM, PT, pointer, extender, filler).

`<XJToBILabelWord>` presents intonation information associated with words. Specifically, the perceived position of accent nucleus is annotated in this tag.

`<XJToBILabelBreak>` is used to indicate break index (BI) labels. The labels "1", "2", and "3" correspond to the boundaries of words, accent phrases, and intonation phrases, respectively. The hierarchical structure of XML is different from that of the context set described in Sect. 4. Hence, we reconstruct the structure using the break index labels of `<XJToBILabelBreak>`. Specifically, we use labels 2 and 3 as the boundaries of the accent phrase and breath group, respectively.

`<XJToBILabelPrm>` is a specific intonation label for X-JToBI. This tag is used to represent lexically irregular prominences.

5.2. Extended context

We propose the use of additional contextual factors that can be obtained from the tags of XML files based on earlier work on HMM-based speech synthesis (Koriyama et al., 2011). The tags used in this study are shown in Appendix A. Most of the tags are categorical information, we encode such tags into one-hot features. Note that we do not use all tags included in the database of CSJ because some tags appear very rarely.

5.2.1. Phone prolongation

When a speaker is thinking, surprised, or emphasizing, phones are often pronounced for longer than in ordinary situations. Since this prolongation is not lexical, additional annotation is required. The labels about phone prolongation

can be obtained from the attributes TagVLong and TagCLong in the tag <Mora>, which denote the prolongation of vowels and consonants, respectively.

5.2.2. Speaking style

When speakers utter with expressiveness such as when laughing and whispering, the speech waveform changes depending on the speaking style. This information is in the attributes of Tag{Whisper—Laughing—Uncertain—...} in the <Mora> and <NonLinguisticSound> tags.

5.2.3. Tone label

It is difficult to model pitch movements in spontaneous speech using only accent-type information because they are much more complicated than those in reading-style speech. For example, a rise-fall type of boundary pitch movement is observed in utterances including dialog acts such as turn-keeping and requesting an agreement. We utilize the labels of low, high, high-low, low-high, and high-low-high boundary pitch movements for the additional contextual factors based on the tag <XJToBILabelTone>. Moreover, we use other tone labels including "A", "pH", and "pL", and add contextual factors about these labels to not only accent phrases but also phones, because the position of tone labels is critical information for pitch contours. Irregular pitch movements annotated in <XJToBILabelPrm> are also used. In addition, we use detail boundary information of <XJToBILabelBreak>, such as "2+p", which represents the existence of pause after the phrase.

5.2.4. Disfluency

Spontaneous speech include many disfluent utterances. CSJ includes filled pauses, word fragments, and restatements as the labels of disfluency in the tag <TransSUW>. The use of these labels is expected to distinguish such disfluent utterances from normal utterances.

5.2.5. Word

In the reading-style speech synthesis described in Sect. 4., word-unit features are omitted from contextual factors because they are not important in practice (Yokomizo et al., 2010). In this study, we incorporate word-unit features into the extended context to examine the effectiveness of such features for spontaneous speech. As the contextual factors of the word unit, we use the information of the part of speech and the conjugate type and form included in the tags <LUW> and <SUW>.

5.2.6. Clause

Although an IPU is a useful unit for spontaneous speech in which the end of a sentence does not often appear explicitly, it is often too short to model sequential information. As a grammatical unit related to a sentence, we can use clauses automatically determined by the transcription texts. Attributes ClauseBoundaryLabel and CU_OperationSign in the tag <SUW> is related to the *strength* of the boundary classified into weak, strong, or absolute. The absolute boundary is equivalent to the sentence boundary of reading-style speech and it frequently becomes the utterance boundary. On the other hand, the weak boundary rarely becomes the utterance boundary. The types of clause boundary are determined according to the final word of the phrase.

Table 1: Amounts of speech data used for experiments.

Speaker	# of Talks	# of IPUs	Duration [s]
F1 (female, ID=19)	6	2144	3727
F2 (female, ID=514)	6	2231	4857
M1 (male, ID=685)	6	2213	3825
M2 (male, ID=471)	6	1650	3004

6. Experiments

6.1. Experimental conditions

We used the speech data of two males and two females included in the CSJ database. The speech data consisted of dialogs, lectures, and simulated lectures. The amounts of training data are shown in Table 1. The training data was segmented into IPUs. The context labels were created for each IPU using XML files in CSJ. We individually trained phone duration and acoustic feature models. The phone duration model used phone-level context as an input feature vector and predicted phone durations, whereas the acoustic feature model predicted frame-level acoustic features from corresponding input features.

We extracted the spectrum envelope, aperiodicity, and f_{o} using WORLD (Morise et al., 2016) from 16 kHz waveforms. We converted the WORLD features into 187-dimensional acoustic features, which consisted of the 0–59th mel-cepstrum, log f_{o}, one-dimensional code aperiodicity, their delta and delta-delta features, and voiced/unvoiced flags. For the baseline context, the dimensions of input feature vectors were 317 and 321 for the duration and acoustic feature models, respectively. For the extended context, the dimensions of input feature vectors were 730 and 734 for the duration and acoustic feature models, respectively.

The architecture of the DNN was a basic feedforward neural network. The number of hidden layers was five and each layer had 1024 hidden nodes. We used the ReLU activation functions and Adam optimizer (Kingma and Ba, 2015) with a learning rate of 0.001. To avoid the overfitting problem, we used weight decay with a coefficient of 10^{-6} and a dropout rate of 0.5. The minibatch size was 1024 and we ran 20 epochs.

For a subjective evaluation test, we merged multiple synthetic IPUs into the speech samples whose durations are approximately 5 s because some of the IPUs were too short to evaluate. The test speech samples were not included in the training data.

6.2. Subjective evaluation results

To evaluate the perceptual quality of synthetic speech, we performed subjective evaluation based on XAB tests, which are generally used to evaluate two samples and find the subtle difference of their quality. The participants on crowdsourcing service first listened to the original reference (X), and chose whether of synthetic samples A and B was similar to the reference. For each test, the number of participants was 30, and each participant evaluated randomly chosen 10 speech segments.

Table 2 shows the result where baseline and extended contexts are compared. Except for the speaker M2, it is seen

Table 2: XAB test comparing baseline context with extended one.

Speaker	Baseline	Extended	p-value
F1	44.3 %	**55.7** %	< 0.01
F2	43.0 %	**57.0** %	$< 10^{-3}$
M1	43.6 %	**56.4** %	< 0.01
M2	46.6 %	53.4 %	0.097

Table 3: XAB test on modified extended context. The speaker was F2 (ID=514).

Removed context	Removed	Extended	p-value
−Tone label	52.0 %	48.0 %	0.33
−Disfluency	49.3 %	50.7 %	0.74
−Word	47.7 %	52.3 %	0.25

that the extended context gave significantly higher scores than the baseline context.

To examine the detail of extended context, we modified extended context by removing categories in Sect. 5.2. one by one. Specifically, we removed tone label, disfluency, and word, respectively, whose tags are frequently observed in the data set. We used the data speaker F2, which has a largest amount of training data among the speakers. The result is shown in Table 3. There were no significant differences between extended and modified contexts. A possible reason is that since spontaneous speech has a large variety in paralinguistic features, removing only one category from the extended context did not affect the perceptual quality of synthetic speech.

7. Conclusions

In this paper, we investigated the effectiveness of rich tags for speech synthesis using the CSJ. We extended input feature vector by using the context according to the tags of CSJ, which includes tone label and disfluency information. Subjective evaluation shows that the use of extended context improved that the reproducibility of spontaneous speech compared with the baseline context. In future work, we should examine the combination of the categories in detail. Furthermore, we will train a multi-speaker model in order to use the tags which are rarely observed for one-speaker's data.

8. Bibliographical References

An, S., Ling, Z., and Dai, L. (2017). Emotional statistical parametric speech synthesis using LSTM-RNNs. In *Proc. APSIPA ASC*, pages 1613–1616.

Dall, R., Hashimoto, K., Oura, K., Nankaku, Y., and Tokuda, K. (2016). Redefining the linguistic context feature set for HMM and DNN TTS through position and parsing. In *Proc. INTERSPEECH*, pages 2851–2855.

Den, Y., Nakamura, J., Ogiso, T., and Ogura, H. (2008). A proper approach to Japanese morphological analysis: Dictionary, model, and evaluation. In *Proc. LREC*, pages 1019–1024.

Hojo, N., Ijima, Y., and Mizuno, H. (2018). DNN-based speech synthesis using speaker codes. *IEICE Transactions on Information and Systems*, E101.D(2):462–472.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *Proc. ICLR*.

Koriyama, T., Nose, T., and Kobayashi, T. (2011). On the use of extended context for HMM-based spontaneous conversational speech synthesis. In *Proc. INTERSPEECH*, pages 2657–2660.

Lorenzo-Trueba, J., Henter, G. E., Takaki, S., Yamagishi, J., Morino, Y., and Ochiai, Y. (2018). Investigating different representations for modeling and controlling multiple emotions in DNN-based speech synthesis. *Speech Commun.*, 99:135–143.

Maekawa, K., Kikuchi, H., Igarashi, Y., and Venditti, J. (2002). X-JToBI: an extended J-ToBI for spontaneous speech. In *Proc. 7th ICSLP*, pages 1545–1548.

Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., Koiso, H., Yamaguchi, M., Tanaka, M., and Den, Y. (2014). Balanced corpus of contemporary written Japanese. *Lang. Resour. Eval.*, 48(2):345–371.

Morise, M., Yokomori, F., and Ozawa, K. (2016). WORLD: A vocoder-based high-quality speech synthesis system for real-time applications. *IEICE Trans. Inf. & Syst.*, E99.D(7):1877–1884.

Sotelo, J., Mehri, S., Kumar, K., Santos, J. F., Kastner, K., Courville, A., and Bengio, Y. (2017). Char2wav: End-to-end speech synthesis. In *Proc. ICLR workshop*.

Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q., Agiomyrgiannakis, Y., Clark, R., and Saurous, R. A. (2017). Tacotron: Towards end-to-end speech synthesis. In *Proc. INTERSPEECH*, pages 4006–4010.

Wang, M., Wu, Z., Wu, X., Meng, H., Kang, S., Jia, J., and Cai, L. (2018). Emphatic speech synthesis and control based on characteristic transferring in end-to-end speech synthesis. In *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*, pages 1–6.

Wu, Z., Swietojanski, P., Veaux, C., Renals, S., and King, S. (2015). A study of speaker adaptation for DNN-based speech synthesis. In *Proc. INTERSPEECH*, pages 879–883.

Yasuda, Y., Wang, X., Takaki, S., and Yamagishi, J. (2019). Investigation of enhanced Tacotron text-to-speech synthesis systems with self-attention for pitch accent language. In *Proc. ICASSP*, pages 6905–6909.

Yokomizo, S., Nose, T., and Kobayashi, T. (2010). Evaluation of prosodic contextual factors for hmm-based speech synthesis. In *Proc. INTERSPEECH*, pages 430–433.

Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A. W., and Tokuda, K. (2007). The HMM-based speech synthesis system (HTS) version 2.0. In *Proc. 6th ISCA Workshop on speech synthesis (SSW6)*, pages 294–299.

Zen, H., Senior, A., and Schuster, M. (2013). Statistical parametric speech synthesis using deep neural networks. In *Proc. ICASSP*, pages 7962–7966.

9. Language Resource References

Maekawa, K., Koiso, H., Furui, S., and Isahara, H. (2000). Spontaneous speech corpus of japanese. In *Proc. LREC*, pages 947–952.

Maekawa, K., Kikuchi, H., and Tsukahara, W. (2004). Corpus of spontaneous Japanese : Design, annotation, and XML representation.

A Tags and Attributes Used in This Study

We show tags and attributes of XML files in CSJ used in this study in Tables 4 & 5 Category means the categories used for extended context 5.2..

Table 4: Tags and attributes (Talk to TransSUW).

Tag / attribute	Category
<Talk>	
TalkID	Baseline
<IPU>	Baseline
Channel	Baseline
IPUStartTime	Baseline
IPUEndTime	Baseline
<LUW>	
LUWPOS	Word
LUWConjugateForm	Word
LUWConjugateType	Word
LUWMiscPOSInfo1	Word
LUWMiscPOSInfo2	Word
LUWMiscPOSInfo3	Word
<SUW>	
SUWPOS	Word
SUWConjugateForm	Word
SUWConjugateType	Word
SUWMiscPOSInfo1	Word
SUWMiscPOSInfo2	Word
SUWMiscPOSInfo3	Word
ClauseBoundaryLabel	Clause
CU_OperationSign	Clause
<TransSUW>	
TagDisfluencyStart	Disfluency
TagDisfluencyEnd	Disfluency
TagDisfluency2Start	Disfluency
TagDisfluency2End	Disfluency
TagFillerStart	Disfluency
TagFillerEnd	Disfluency
TagFillerMidst	Disfluency
TagIncorrectStart	Disfluency
TagIncorrectEnd	Disfluency
TagIncorrectMidst	Disfluency

Table 5: Tags and attributes (Mora to XJToBILabel*).

Tag / attribute	Category
<Mora>	
MoraID	Baseline
TagVLong	Phone prolongation
TagCLong	Phone prolongation
TagWhisperStart	Speaking style
TagWhisperEnd	Speaking style
TagWhisperMidst	Speaking style
TagLaughingStart	Speaking style
TagLaughingEn	Speaking style
TagLaughingMidst	Speaking style
TagUncertainStart	Speaking style
TagUncertainEnd	Speaking style
TagUncertainMidst	Speaking style
<NonLinguisticSound>	
TagBreath	Speaking style
TagLaugh	Speaking style
TagVN	Speaking style
TagWhisperStart	Speaking style
TagWhisperEnd	Speaking style
TagWhisperMidst	Speaking style
TagLaughingStart	Speaking style
TagLaughingEnd	Speaking style
TagLaughingMidst	Speaking style
TagUncertainStart	Speaking style
TagUncertainEnd	Speaking style
TagUncertainMidst	Speaking style
<Phone>	
PhoneEntity	Baseline
Devoiced	Baseline
PhoneStartTime	Baseline
PhoneEndTime	Baseline
<XJToBILabelTone>	Tone label
ToneClass	Tone label
Divided	Tone label
<XJToBILabelWord>	
PerceivedAccPos	Baseline
<XJToBILabelBreak>	Baseline
<XJToBILabelPrm>	Tone label