DNN-based Speech Synthesis Considering Dialogue-Act Information and Its Evaluation with Respect to Illocutionary Act Naturalness

Nobukatsu Hojo1, Yusuke Iijima2, Hiroaki Sugiyama1, Noboru Miyazaki1, Takahito Kawanishi3, Kunio Kashino1

1 NTT Communication Science Laboratories
2 NTT Media Intelligence Laboratories

Abstract
This study aimed at improving synthesized speech generated by a text-to-speech (TTS) system used for a spoken dialogue system in regard to how naturally the synthesized speech conveys the system’s intention to the hearer. We call the measure of naturalness in this case “illocutionary act naturalness”. To achieve our aim, we utilized dialogue-act (DA) information as an auxiliary feature for a deep neural network (DNN)-based speech synthesis system. First, we constructed a speech database with DA tags. Second, we used the database to build the speech synthesis system. Third, we evaluated the method by comparing its performance with a DNN making use of conventional linguistic features and hidden Markov models (HMMs) supplemented with DAs. We conducted a listening test designed to evaluate illocutionary act naturalness. The results show that the proposed method improves the illocutionary act naturalness compared with the conventional method. We also found that the illocutionary act naturalness score depended on certain features of the test sentence as well as the DA and speech synthesis method. The results shows that a test set designed by considering these features will improve the reproducibility of the illocutionary act naturalness evaluation.

Index Terms: speech synthesis, spoken dialogue systems, illocutionary act naturalness

1. Introduction
This paper describes deep neural network (DNN)-based speech synthesis considering dialogue-act (DA) information in order to improve the naturalness of synthesized speech generated by a spoken dialogue system. In communication, hearers infer a speaker’s intention from every utterance [1, 2]. If a spoken dialogue system generates an utterance in an unnatural way in an attempt to express its intention, it imposes an unnecessary cognitive load for inference upon its users. Here, the intention is related to prosody [3, 4] as well as a sentence. For example, different prosodies of “Excuse me” can convey different intentions, e.g., criticism or apology. Therefore, text-to-speech (TTS) synthesis systems play a role in natural expression of intentions. On the basis of these considerations, we attempted to improve TTS in regard to how naturally the synthesized speech conveys the system’s intention to the hearer.

In the field of speech synthesis, naturalness is often defined as the quality of speech samples separated from contexts [5]. On the other hand, speech needs to be natural as a way to express intention, as described above. Based on the classification by the speech act theory [6, 7], we can rephrase the conventional naturalness, naturalness of an act to say something, as “locutionary act naturalness (LAN)”. We can also rephrase naturalness in this work, naturalness of an act to convey intentions, as “illocutionary act naturalness (IAN)”. Our study thus aims at improving the IAN of synthesized speech.

A promising approach to improving LAN is to reproduce the prosodic features of intentions. A related field of study is emotional speech synthesis [8, 9, 10, 11, 12, 13, 14]. The common point is that para-linguistic information [15] is expressed by TTS. However, emotions and intentions have different prosodic features. Emotional speech has salient features for the whole utterance. For example, “sad” speech generally has a lower F0 and a slower speech rate [16]. On the other hand, to express intentions, for example, the sentence final tone is also important in Japanese speech [17, 18, 19]. This feature is contrastive to emotional speech in that it appears locally in time. This difference shows the necessity of determining whether emotional speech synthesis methods are also effective for expressing intentions.

Previous studies have shown the effectiveness of utilizing DAs for TTS as a way of considering intentions [20, 21]. A DA is an abstract expression of a speaker’s intention [22]. However, the previous studies have not revealed two points. First, they have investigated TTS based on concatenative synthesis [20] and hidden Markov models (HMMs) [21], but not DNNs. DNNs have been shown effective for emotional speech synthesis [14], so the question here is whether they are effective at expressing intentions as well. Second, they evaluated LAN, but not IAN. We believe IAN is an important attribute of synthesized speech.

Here, we propose DNN-based TTS considering DAs and report an evaluation with respect to IAN. In particular, we used DA as an auxiliary feature for feed-forward DNNs. Although sequence modeling with neural networks is effective for emotional speech synthesis, the limited size of the speech corporuses prevented it from being used [12, 13, 14, 23]. The performance of the proposed method was compared with a DNN making use of only conventional linguistic features and hidden Markov models (HMMs) supplemented with DAs.

We also describe the design of a test set that is useful for deriving results with sufficient reproducibility. The problem is that evaluations of IAN may have low reproducibility because they depend on the selection of test-set sentences. The cause of this dependency is that IAN is likely to be affected by the sentences as well as the TTS methods, because different sentences may need different prosodic features of an intention (e.g., sentence final tones depend on sentence final particles in Japanese [17, 18, 19]). One way to alleviate this dependency is to use a larger test set, though this increases the cost of any evaluation experiment. Another approach is to make an assumption about which features of a sentence affect IAN and then design the test set by controlling the frequencies of those features in it. The evaluation experiment in this study is based on the latter approach. The validity of the design of the test set was examined by conducting a two-way analysis of variance (ANOVA) on the results of a subjective evaluation.

2. DNN-based TTS Considering DA

2.1. Speech Database with DA Tags [24]
To build a TTS system that can consider DAs, first, we constructed a speech database with DA tags. As a DA set for the experiments, we used the one proposed in [25]. This set is designed to cover a wide range of utterances of non-task oriented
open-domain conversation and consists of 30 DAs. The sentences for the speech recording were extracted from a text chat database in Japanese. This database was originally gathered by Higashinaka et al. [26] and contains 3680 conversations (with 134K sentences). The utterances have been manually tagged with DAs by two experts.

The sentences for the recording were extracted by considering the balance of the frequency of phonemes and DAs on an entropy basis [27]. For the speech recording, we used the manuscripts illustrated in Fig. 1. The manuscript shows the sentences for recording, their DAs, and several preceding utterances as context. We recorded the speech of a Japanese female professional voice actor. We instructed her to read the manuscript silently first so that she would understand the conversational context before each utterance. She spoke each utterance in a natural conversational speaking style for the context and corresponding DAs. We derived 5177 sentences from these recordings. We excluded duplications (utterances with the same sentence and DA) from them and used 3410 utterances, about 140 minutes in total, as the speech database. All of the utterances were manually annotated with phonemes, accent types, and phonetic boundary information so that they can be used as training data for speech synthesis models.

2.2. DNN-based TTS considering DA

The straightforward method to generate speech considering DA is to train a model for each DA and to switch the model according to the DA when synthesizing. However, this method needs a considerable amount of training data for each DA to derive high-quality speech, which increases costs. The proposed method utilizes DAs as input for the DNN, which enables a single DNN to model the speech of all of the DAs. Conventional HMM-based methods have succeeded in generating speech from a small amount of speech data [8, 28, 29]. However, their performance is limited because their tree structure restricts the relationship between the inputs and outputs that can be modeled. Our DNN-based method does not have such a restriction in the model architecture, so we expected that it would be able to reproduce prosodic features of DAs more precisely than HMM-based methods.

Figure 2 shows a schematic diagram of the proposed method. The architecture is the same as the one of DNN-based speech synthesis using speaker codes [30], except that the auxiliary feature z indicates a DA, not the speaker ID. We used 1-hot code as the DA code z [30].

3. Evaluation Experiments

3.1. Speech Samples

We constructed five different TTS systems for the evaluation:

- **HMM-BASELINE**: HMM making use only of conventional linguistic features [31].
- **HMM-MIXED**: HMM-based style-mixed modeling method [8].
- **HMM-ADAPT**: HMM-based average model and model adaptation method [28, 29].
- **DNN-BASELINE**: DNN making use only of conventional linguistic features [32].
- **DNN-DACODE**: The proposed DNN-based method using DA codes.

We compared the proposed method with conventional HMM-based methods, listed as HMM-MIXED, HMM-ADAPT and DNN-DACODE. We also compared DNN-BASELINE and DNN-DACODE to validate the effect of DAs derived by DNNs. We also evaluated HMM-BASELINE to validate the effect of DAs derived by HMMs.

For all five methods, we used acoustic features consisting of mel-cepstra, log F0, and band aperiodicities and their dynamic features, extracted at 5 ms intervals by using STRAIGHT [33]. For the DNN-based methods, voiced/unvoiced binary variables were added to the acoustic features. The number of dimensions of the acoustic features was 138 for the HMM-based methods and 139 for the DNN-based ones. The number of dimensions of the linguistic features for the DNNs was 486 for the duration models and 489 for the acoustic models. We used an affine-transformation post filter [34] for the mel-cepstra sequence.

For DNN-BASELINE and DNN-DACODE, the duration models had two hidden layers with 256 sigmoid units each, while the acoustic models had four hidden layers with 256 sigmoid units each. 30-dimensional DA codes were input to all the hidden layers of DNN-DACODE. The parameters were updated to minimize the mean square error by using Adam algorithm [35]-based back-propagation. The learning rate was set to $\alpha = 0.00015$ for the acoustic models and 0.001 for the duration models.

For the HMM-based methods, we used five-state left-to-right multi-space probability distribution hidden semi-Markov models (MSD-HSMMs) without skip. The MDL parameter was set to $\alpha = 1.0$. For HMM-MIXED, we used questions with respect to DAs as well as linguistic information. We used 30 questions, each of which corresponded to a DA [8]. We added 8 questions, each of which corresponded to a DA class described in sec. 3.3. The total number of questions related to DAs was 38. For HMM-ADAPT, we trained an average voice model by using the speech data of all the DAs. Then we adapted the average voice model to each DA by using the speech data of the corresponding DA. We used a combination of constrained structural maximum a posteriori linear regression (CSMmapLR) and MAP adaptation [29].

We selected 120 utterances from the speech corpus as the test set, as described in sec. 3.3. The remaining utterances were used as training data for all five methods.

3.2. Evaluation Method of IAN

We conducted a subjective evaluation of IAN. Since conversational context is important in perceiving intentions, we displayed the context in a GUI (Fig. 3). We instructed the partic-
INSTRUCTIONS:  
1. You are now chatting with a robot, Riko-san.  
   Assume that you have just finished the dialogue below.  
2. Listen to the audio and suppose it is an utterance by Riko-san  
   following the dialogue.  
3. Answer the question below.

DIALOGUE: (You are chatting in your room.)  
Riko-san: I like Epikuro rainier.  
You: Me too. How about Ichiran?

QUESTION: Do you think the audio is natural when uttered in the following mind?  
   Although there are speech samples with low voice quality,  
   please evaluate the naturalness of the way it speaks  
   (e.g. pitch, speech rate, etc.), not the voice quality.  
   Riko-san said a filler.  
   (She is thinking of what to say and wants to continue to say something.)  
   0: unnatural 1: (somewhat unnatural) 2: (somewhat natural) 3: (fair) 4: (somewhat natural) 5: (natural)

[Back] [Next]

Figure 3: Example of GUI used in the subjective evaluation experiments of IAN. The sentence of the speech sample is “so desune (Let’s see)”.

Table 1: The DA classes used in the evaluation experiments. The right column has 30 DA in total because “self-disclosure” and “question” have 8 and 7 subclasses, respectively [25].

<table>
<thead>
<tr>
<th>DA Class</th>
<th>DA</th>
<th>INFORMATION</th>
<th>sympathy, non-sympathy, appraisal</th>
</tr>
</thead>
<tbody>
<tr>
<td>QUESTION</td>
<td>question (other than “question_self”), confirmation, proposal</td>
<td></td>
<td></td>
</tr>
<tr>
<td>REPEAT-PARAPHRASE</td>
<td>repeat, paraphrase</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACKNOWLEDGEMENT</td>
<td>acknowledgement</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FILLER</td>
<td>filler</td>
<td></td>
<td></td>
</tr>
<tr>
<td>COURTESY</td>
<td>greeting, thanks, apology</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4. Results

4.1. IAN’s Dependency on Sentence Class

Fig. 4 shows the results of the subjective evaluation. First, we tested whether there were interaction effects of TTS methods and sentence classes. Concretely, for each DA class, a repeated-measures two-way ANOVA was carried out on the factors TTS methods (“METHOD”) and sentence classes (“SENTENCE”). The text boxes above the bar plots in Fig. 4 show the results of the analysis. The ANOVA revealed that “SENTENCE” and the interaction between TTS methods and sentence classes (“INTERACTION”) had a significant effect on IAN in 7 out of 8 DA classes. These results confirm that the sentence classes in this experiment actually affected the results of the evaluation of IAN.

The ANOVA results indicate that the design of the test set affects the evaluation of IAN. In other words, reproducibility will be degraded when IAN is evaluated in accordance with the conventional procedure that randomly selects utterances for the test set. For example, suppose we compare IAN of DNN-BASELINE and DNN-DACODE for “ACKNOWLEDGEMENT” in Fig. 4: sentence classes 1 and 3 have smaller differences, but sentence class 2 has a larger one. Then, the results of an evaluation using the conventional procedure may vary depending on the proportion of class 2 sentences in the test set. Therefore, the proportion should be controlled when the test set is designed.

4.2. Comparison of TTS methods

Since the interaction effects between TTS methods and sentence classes were significant, we tested simple main effects.

977
In particular, we conducted a t-test ($\alpha = 0.01$) between the TTS methods for each sentence class and DA class. The results are plotted in Fig. 4.

Comparing DNN-DACODE with HMM-MIXED and HMM-ADAPT, we see that it is superior or comparable to the conventional methods for all of the DA and sentence classes. These results show the superiority of the proposed method to the HMM-based conventional methods. Moreover, comparing DNN-DACODE and HMM-BASELINE, we see that the proposed method is superior or comparable to the conventional one for all of the DA and sentence classes. The results show the effectiveness of considering DA.

The MOS scores of the five methods were comparable for some DA and sentence classes, but different for others. This shows that DA is helpful for some sentences, but not so much for others. It would be interesting to determine why this tendency exists by comparing the MOS scores with objective and other subjective measures.

5. Conclusions

This study aimed at improving TTS in regard to how naturally the synthesized speech conveys the system’s intention, or its “illocutionary act naturalness” (IAN). For this purpose, we utilized DAs as an auxiliary feature in a DNN-based speech synthesis system. We constructed a speech database with DA tags and built five TTS systems, one of which incorporated the proposed method. We conducted a listening test that was designed to evaluate IAN. The results showed that the proposed method improved the illocutionary act naturalness compared with the conventional methods. We also found that the MOS results depend on certain features of the sentences included in the test set. Therefore, to ensure that evaluations of IAN are reproducible, we should design test sets by considering the frequencies of those features in them. Our future work will include analyzing the results by comparing them in terms of objective and other subjective measures. We will also investigate sequence models to further improve IAN.
6. References


