Concept-to-Speech Generation by Integrating Syntagmatic Features into HMM-Based Speech Synthesis

Xin Wang, Zhen-Hua Ling, Li-Rong Dai

National Engineering Laboratory for Speech and Language Information Processing, University of Science and Technology of China, Hefei, China

Abstract

In conventional concept-to-speech (CTS) methods, a common step is predicting abstract prosodic descriptions, such as the locations of accents and phrase boundaries, from the linguistic information provided by the text generation module. But the prediction results always contain errors, and unacceptable prosodic prediction may ruin the synthesized speech. In addition, linguistic information, which can be given conveniently and accurately by text generation, has not been directly utilized in the acoustic modelling and speech generation of CTS. This paper displays a CTS method utilizing HMM-based speech synthesis (HTS) and a text generation module called Komet-Penman multiligual (KPML). In this method, syntagmatic features derived from the linguistic information given by KPML is directly added to the context features for context-dependent HMM modelling. Further, prosodic features are discarded during acoustic modelling to avoid costly prosodic annotation on the training waveforms and inaccurate prosodic prediction at synthesis time. Experiments show that the proposed method performs no worse than the conventional method with automatic prosodic prediction. When manual prosodic annotation on the training corpus is unavailable, the proposed method performs better.

Index Terms: concept-to-speech, HMM-based speech synthesis, natural language generation

1. Introduction

Speech production is a vital task in human-machine interaction. Among different modes of speech production, text-to-speech (TTS) is aimed at converting text into speech; however, when the input is encoded in semantic or conceptual representation, concept-to-speech (CTS) is necessary.

Although CTS could be realized by appending TTS to a text generation module as Figure 1(a) shows, certain couplings of the two components are assumed to yield better performance [1]. Typical method is enriching the text with prosodic descriptions which are inferred from the rich linguistic information given by the text generation module [2]. This method is shown in Figure 1(b). The prosodic descriptions usually include abstract symbols such as accents [3] and phrase boundaries [4].

In both the conventional CTS methods show in Figure 1(a) and Figure 1(b), to build a prosodic model which predicts the prosodic descriptions based on plain texts or rich linguistic features is demanding. First, manual annotation for training data collection is laborious. Even with a well annotated corpus, both automatically and manually acquired prosodic model may be imperfect. The one best hypothesis given by a prosodic model may be unacceptable and finally worsen the synthesized speech. Another imperfection of many conventional CTS methods is employing TTS as a black box. The linguistic features from text generator could have been used for tuning the speech synthesizer in TTS. Some methods open the black box and eliminate the redundant text analyser in TTS [5], but the speech synthesizer still shares no knowledge with the text generator.

This paper proposes a Chinese CTS method based on HMM-based speech synthesis (HTS) [6] and a text generation module called Komet-Penman multiligual (KPML) [7]. Specific linguistic features from KPML, namely syntagmatic features, are added into the model context of HMMs. In this way, the missing link between the linguistic information and acoustic modelling can be examined. To avoid the possible damage from inaccurate prosodic prediction to the synthesized speech, prosodic features in the original context are eliminated after slight modification on the structure of HMMs. The loss caused by the elimination could be compensated by syntagmatic features because of their potential for describing the supragrammatical property of speech. Besides, syntagmatic features are derived from the linguistic information given by KPML, thus to predict syntagmatic features at synthesis time is needless.

In the rest of the paper, Section 2 will show the proposed method and the usage of syntagmatic features for HMM-based acoustic modelling. Experiment and analysis will be given in Section 3. The conclusion will be drawn in Section 4.

2. The Proposed Method

2.1. General structure

Figure 1(c) shows the proposed method: language generation and speech synthesizer are based on KPML and HTS respectively; the interface converts all the linguistic information from KPML into abstract structure compatible to HTS. For Chinese CTS, the linguistic information contains syntagmatic features and the sequence of phoneme and tone; no plain text is involved.

Different from conventional methods, the proposed method directly utilizes syntagmatic features from KPML for acoustic modelling in HTS. Besides, prosodic features and the prosodic model are eliminated to avoid the inaccurate prosodic predictions at synthesis time and the necessity of manual prosodic annotation on the corpus. Sections below will display the details.

2.2. Text generation based on KPML

There are commonly two steps in a fully-fledged text generation module. Inputs are first mapped into text specification which encodes semantic items; then text specification is converted into text [8]. Because generating text specifications from data is task-dependent, the text generation module in our proposed method is only concerned with the second step.
That balloon is rising in English.

Figure 2: Syntagmatic tree for “那支气球正在上升”， which means that balloon is rising in English.

Chinese spoken language is assumed to be hierarchically organized based on prosodic words, phrases, and other constituents [12]. For example, “那支气球正在上升” may be read as “那支气球正在上升”. In the utterance, “那支气球正在上升” indicates the boundary between prosodic words such as “那支” and “气球”; 一个 separates two prosodic phrases. Prosodic features derived from the hierarchy are essential to realize the timing, intonation, and other prosodic subtleties in synthesized speech.

For the process of acoustic parameter generation shown in (2), prosodic features must be predicted either by the prosodic module within TTS in Figure 1(a) or the stand-alone prosodic model in Figure 1(b). Both approaches may predict inaccurate prosodic features, which directly affects the predicted \( \hat{\alpha} \). The possible discrepancy between prosodic features and acoustic data \( \hat{O} \) may be used to improve the poor acoustic model. As the result, the prediction of \( \alpha \) at synthesis time will be influenced.

A simple solution is eliminating prosodic features. Thus prosodic modelling can be avoided and corpus annotation is unnecessary. As substitute for prosodic features, linguistic features from the text generation module could be exploited.

---

**Table 1: Syntagmatic features of “支” zhi1 in Figure 2**

<table>
<thead>
<tr>
<th>Feature name</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>( FF_F, FF_C, FF_F )</td>
<td>Deitic.Quantity-marker, Thing</td>
</tr>
<tr>
<td>( HD_F, HD_F )</td>
<td>2, 2</td>
</tr>
<tr>
<td>( PostF )</td>
<td>0, 1, 1</td>
</tr>
<tr>
<td>( NumF )</td>
<td>1</td>
</tr>
</tbody>
</table>

---

**Figure 1: Diagrams of the conventional CTS methods (a), (b) and the proposed CTS method (c).**

Among existing implementations, KPML, a text generation module based on systemic functional grammar [9], is chosen because of its support to Chinese text generation [7]. Given grammar resources and input text specification in sentence planning language (SPL) [10], KPML generates the surface text and corresponding syntagmatic tree. In this paper, we use syntagmatic tree to denote the grammatical structure specified by the functional grammar resources. The syntagmatic tree for sentence “那支气球正在上升”, which means That balloon is rising in English, is shown in Figure 2. The leaves of the tree contain Pinyin and Chinese characters. The non-terminal node bears the function of a unit on certain rank scale [9]. For example, the father node of “支” zhi1 is a quantity-marker for balloon in Chinese. From the syntagmatic tree, phoneme, tone, and syntagmatic features can be extracted for speech synthesis.

---

**2.3. Speech synthesizer based on HTS**

The speech synthesizer should support direct utilization of syntagmatic features in acoustic modelling. Among possible solutions, HMM-based speech synthesis (HTS) [6] is selected. The core of HTS is modelling the distribution of acoustic parameters \( O \) given corresponding context \( W \) and predicting acoustic parameters \( \alpha \) given target context \( w \). The two steps are shown in (1) and (2) where \( \lambda \) is the model parameter set of HMM.

\[
\hat{\lambda} = \arg \max_{\lambda} p(O|W, \lambda) \tag{1}
\]

\[
\hat{\alpha} = \arg \max_{\alpha} p(\alpha|w, \hat{\lambda}) \tag{2}
\]

In HTS, the model contexts are composed of various types of linguistic features. The model contexts in the proposed method include tone, phoneme, and syntagmatic features.

---

**2.4. Integrating syntagmatic features into HTS**

**2.4.1. Extracting syntagmatic features**

Phoneme and tone from the Pinyin sequence are well defined, but syntagmatic features require careful design. Basic syntagmatic features for “支” zhi1 are shown in Table 1. Here, \( FF_F \), \( FF_C \), and \( FF_F \) encode the functional tags for preceding, current, and following word, namely Deitic for “支”, Quantity-marker for “气球”.

---

**Figure 2: Syntagmatic tree for “那支气球正在上升”， which means That balloon is rising in English.**

“支”， and Thing for “气球”. HD describes the difference of hierarchical level between “支” and neighbouring words. PostF describes relative, forward, and back features governed by the father node Quantity-marker. NumF depicts the length of father node. Then FF, PostF, and NumF will be extended to the grandfather and grand-grandfather nodes. Finally syntagmatic features of 23 dimensions will be formed for every leaf of the syntagmatic tree.

Our definition of syntagmatic features are similar to the syntactic features used in [11], but the syntagmatic features are oriented to language function. For example, the word “正在”， which means now in English, bears the function of tense-marker. In syntactic parsing, it may be labelled as adverb which denotes the part-of-speech. What’s more, Subject in Figure 2 is composed of three units “气球”, “支”, and “气球”. In syntactic structure, the latter two units may form a small tree and then a large tree with the first unit. In other words, the syntagmatic tree is flatter than the syntactic tree[9].

---

**2.4.2. Replacing prosodic features with syntagmatic features**

In English, bears the function of tension-frame. For example, “Tense-marker. In syntactic parsing, it may be labelled as “adverb” which denotes the part-of-speech. What’s more, Subject in Figure 2 is composed of three units “气球”, “支”, and “气球”. In syntactic structure, the latter two units may form a small tree and then a large tree with the first unit. In other words, the syntagmatic tree is flatter than the syntactic tree[9].
our proposed method, syntagmatic features will play the role. Through incorporating syntagmatic features into the model context, direct link between the linguistic information and acoustic model can also be established. Grammatical structure of the sentence is documented to be related with prosodic structure [13][14]. We assume this correlation also exists between syntagmatic structure and prosodic structure.

But the substitution can’t be realized directly. A key barrier is modelling silent pause. In HTS for Chinese, the location of silent segment in training set is specified explicitly so that it can be modelled by a silent pause (SP) model. During prediction, prosodic features specify the location of SP model; then silent segment can be predicted. Unfortunately, syntagmatic features don’t specify silent pause explicitly. One solution may be modelling the location of SP based on syntagmatic features; however, this method is nothing but simplified prosodic modelling.

Another solution is to model the silent pause implicitly within the acoustical model. To implement the idea, we assume that all the silent segments are modelled by the last state of the HMM which describes the phone ahead of the SP. Based on the assumption, a transitional arc is added to skip over the last emitting state. Figure 3 shows the idea. For model training, the Baum-Welch algorithm still works [15]. With the additional transition, the last state can be skipped both in model training and acoustic parameter predicting.

3. Experiments

3.1. Corpus preparation

For CTS generation integrating syntagmatic or syntactic features, strict match between those linguistic features and utterances in the corpus is required. Thus we followed the three steps in [2] to prepare the corpus. First, we prepared a set of input concept based on which KPMIL generated the 378 sentences with syntagmatic features. All the sentences simulated the responses of a customer service representative to the clients. In the second step, a female speaker read all the sentences in a simulated conversational environment. At last, prosodic features were manually annotated on the recorded waveform. The total duration of the utterances is about 35 minutes.

For HTS, an additional step is extracting acoustic parameters. Base on the waveform, 40th order frequency warped Line Spectral Pairs (LSP) and an extra gain dimension were derived from the spectral envelop extracted by STRAIGHT [16]. Discontinuous F0 trajectory was also extracted by STRAIGHT. LSP, F0, and their first order and second order dynamic components formed the acoustic parameter set. For experiments, 10% of the corpus were randomly selected as test set.

3.2. Prosodic models for comparison

One aim of experiments is to show the negative influence that conventional CTS framework received from two aspects: the inaccurate prosodic annotation on training corpus; the inaccurate prosodic prediction on test set during speech synthesis.

To simulate the influence, we utilize PM2, an embedded prosodic model in a fully developed text analyzer provided by iFlyTek. Incorporating PM2 at synthesis time simulates the practical situation where prosodic features for the test set must be predicted automatically. Applying PM2 to training data simulates the case when prosodic features for the training set were annotated based on text analysis instead of human perception.

Because PM2 was not trained using corpus from the same domain mentioned in Section 3.1, we built a decision-tree-based prosodic model PM1 using syntagmatic and prosodic features from both training and test sets mentioned in Section 3.1. Thus PM1 simulates a prosodic model with performance approximating the possible upper bound. The performance of the two prosodic models on training and test data is shown in Table 2.

3.3. System construction

Experimental groups for conventional methods with prosodic features are group Aa to Ae listed in Table 3. They all followed the standard configuration for HTS, but prosodic features for training and test sets came from different sources. Among the set A* (brief notation for Aa to Ae), Aa simulated the ideal situation where prosodic features on both training and test sets were manually annotated; Ab and Ac simulated practical situations where prosodic features on test set must be predicted based on text or other linguistic information; Ad and Ae were more practical cases when hand-annotation on the training set was unavailable. Compared with counterpart in A*, group from B* included syntagmatic features into the model context without eliminating prosodic features.

All groups from C* discarded prosodic features but they varied in the structure of HMM. Cc utilized the standard left-to-right HMM; Cc and Cb contained the additional transition shown in Figure 3. Because the last state of HMM for C* would be used for describing silent pause, the number of HMM states for non-silent phonemes may decline. For a fair comparison, HMM in the proposed system Cc had 6 emitting states.

3.4. Results and analysis

For objective evaluation, durations for the test set were obtained through force alignment. In subjective evaluation, they were predicted by the duration model of each system. The results of objective evaluation are shown in Figure 4. Among set A*, Aa achieves the lowest RMSE on both LSP and F0 streams. If prosodic features for either training or test set are not manually annotated, the values of RMSE rise.

First, comparison among Aa, Ab, and Ac in Figure 4(b) and 4(c) shows that Ab and Ac yield worse results than Aa. Although prosodic features on training set for the three systems are manually annotated, prosodic features for test set in Ab and Ac are predicted using PM1 and PM2. Inspection shows that predicted spectral parameters deviates from natural ones more severely when the corresponding segment in Ab and Ac bears fallacious prosodic features. For F0 stream, fallaciously predicted prosodic features also affect the predicted F0 contour in adjacent segments. Thus inaccurate prosodic prediction on test set affects the predicted acoustic parameters negatively.

On the influence of inaccurate prosodic annotation on train-
The proposed system doesn’t perform worse than the conventional methods, even with imperfect prosodic annotation and prediction. This result is caused by the discrepancy between prosodic features and acoustic data in training set of Ad and Ae. With the inferior acoustic model, Ad and Ae give higher RMSE than their counterparts Ab and Ac which were trained on manually annotated prosodic features.

Thus the imperfect prosodic annotation and prediction affects the performance of conventional CTS. The above results are also manifest on set B*. Although Ba using both syntagmatic and prosodic features yields the lowest RMSE on both spectral and F0 streams among all groups, the inaccurate prosodic features still disturb the performance of any group from Bb to Be as the values of RMSE on spectral and F0 streams in Figure 4(b) and 4(c) show.

Compared with the conventional systems, RMSE for spectral stream of proposed system Ca is only lower than those from Ac and Be; however, Ca performs better on F0 stream than group from Ac to Ae and Be to Be. The objective results show that the proposed system Ca doesn’t perform worse than the conventional methods when prosodic features for either training or test set are not manually annotated. Among other system without prosodic features, Cb achieves higher likelihood on the training set and lower RMSE on test set than Ca. Thus comparison between Cb and Ca supports the proposed modification on the structure of HMM. Comparison between Ca and Cb shows that one more state leads to better results both on training and test sets. Thus no over-fitting occurs after adding the additional state and transition for the proposed Ca.

Subjective evaluation between Ca and groups from set A* was conducted. The result is shown in Figure 5. The quality of synthesized speech from Ca is inferior to that from Aa with manually annotated prosodic features on both training and test sets. But the quality of synthesized speech from conventional method drops when prosodic features for test set are inferred from text or syntagmatic features. The deterioration is more evident when training set is based on imperfect annotation, which is shown by the tests between Ca and Ad, Ca and Ac and Ae. Another series of evaluation between Ca and groups from B* also shows similar results. In addition, the difference between group from B* and counterpart from A* is trivial.

The results above indicate that the proposed methods is inferior to conventional methods when both training and test set acquire manually annotated prosodic features; however, in practical environment where prosodic features for test set is not guaranteed to be totally acceptable, the proposed method perform comparably to the conventional method. When manual annotation on the corpus is unavailable, the conventional methods yield worse results than the proposed method. Even based on prosodic model PM1 with high performance, the deterioration is evident. Although the results may be partially due to the idiosyncrasies of the corpus, for practical use of CTS, the challenge from prosodic modeling always exists.

4. Conclusions

In this paper, we have proposed a CTS method based on KPM-L and HTS. Then we mainly discussed using syntagmatic features as substitutes for prosodic features. Experiments involving simulated conventional systems under various conditions were conducted. The results support the idea that syntagmatic features could achieve comparable or better results when prosodic annotation and modelling are imperfect.

Our work on CTS currently only focuses on neutral speech synthesis. More exploration will be launched into the potential of linguistic information for expressive speech synthesis.

5. Acknowledgements

This work was partially funded by the National Nature Science Foundation of China (Grant No. 61273032).
6. References


