ATLANTIS PRESS

Advances in Intelligent Systems Research, volume 168

International Conference on Modeling, Analysis, Simulation Technologies and Applications (MASTA 2019)

A Novel Unit Selection and Unit Smoothing Method for Chinese Concatenation Speech

Xiao-kang YANG[1,*], Zhi-cheng LIU[2], Qi-long SUN[3] and Hao-yuan WANG[1]

[1]China Mobile Online Services Co., Ltd

[2]School of Computer Software, Tianjin University, Weijin Road, Tianjin, 300072, China

[3]Qinghai Nationalities University

*Corresponding author

Keywords: Statistical parametric speech synthesis, LSTM, DTW, Concatenation smoothing.

Abstract. This paper introduces a new approach to unit selection and unit concatenation, in which Chinese character is the smallest unit in speech corpus and at concatenation stage, speech segments are not only concatenated in phase, but also in amplitude. A conventional hybrid system is used in this paper. Firstly, LSTM were adopted for acoustic model and duration model, and prosody is predicted by Conditional Random Fields (CRFs). Secondly, without considering continuously-valued cost, we use Dynamic Time Warping (DTW) directly to select units with acoustic features such as mel-cepstrum and Fundamental Frequency (F0). At last, an improved cross-fade method taking amplitude into account is adopted in waveform concatenation to improve smoothing and very natural speech is synthesized.

Introduction

The recent rise of deep neural networks (DNNs) has brought an increase in performance in both automatic speech recognition (ASR) statistical text-to-speech (TTS) technology [1]. Parameter synthesis system and unit concatenation synthesis system as two mainstream speech synthesis systems also have been well-developed because of DNNs. Not like parameter synthesis method using vocoder such as WORLD to synthesize speech, concatenation synthesis method select proper units from multiple instances to achieve better flexibility quality in prosody and timbre [2].

Tonal syllable is considered to be the basic units in synthesis system for Mandarin, because of very strong co-articulation between phonemes in a same syllable and much less concatenation points between syllables [2]. Some researchers find the juncture between phonemes, between syllables, between rhythm units and between stops in Chinese, between phonemes are the strongest, the one between syllables is next stronger and the others are weaker [3,4] trains a group of syllable classifiers, which takes not continuously-valued cost into each candidate unit. [5,6] also suggest that eliminating the unacceptable units is crucial to synthesize a natural speech.

LSTM guided unit selection synthesis system have achieved state-of-the-art performance in statistical parametric speech synthesis (SPSS) system due to its deep architecture and capacities to long-term dependencies across the linguistic features, which HMM doesn't possess [7,8] confirms the objective result that LSTM can better model acoustic features than DNN. So, LSTM is adopted for acoustic modeling and duration modeling. And, the current mainstream hybrid system is applied to synthesize speech, in which CART decision trees take part in the unit pre-selection and DTW is being used for selecting optimal unit. At last a new fade-in/out method taking into account amplitude is adopted in waveforms concatenation to improve smoothing.

The paper is organized as follows. Section 2 discusses the preprocessing techniques for the speech corpus that is used in our method. Section 3 introduces the approach about the LSTM and CRF-based parametric synthesis system and describes an improved algorithm of unit concatenation smoothing. Objective tests and evaluation is presented in section 4. Section 5 is conclusion.

Preparation of Speech Corpus

The quality of unit concatenation speech depends to a large speech database including the variability and availability of pronunciation unit [2]. It is crucial to construct a corpus that covers all speech units for starting with synthesis speech. In our speech corpus, 8000 sentences, about 10 hours, are recorded by a professional female announcer.

Basic Concatenation Unit

Phone-sized units as the basic units are selected in many concatenation synthesis systems [7,9], because of easily collecting and rich prosody varieties. But, smaller unit means much more points to concatenate and more distortions caused. In order to preserve the naturalness of speech, tonal syllables are chosen as the basic units in some systems [2,4], and some achievements have been made. Yet, the same tonal syllables may correspond to different Chinese characters, so it is waste computation when in the stage of unit pre-selection. In this paper we try to use Chinese characters as the basic units. In our text corpus, there are 3665 Chinese characters, which is enough to meet 3500 Chinese characters in common use.

The multi-dimensional and CART decision tree [2] is used here to solve the problem of units pre-selection.

Front-End

In this paper, a hybrid synthesis system is adopted and we use Merlin speech synthesis toolkit [8] to realize our SPSS system. The SPSS system takes linguistic features as input, and employs neural networks to predict acoustic parameters. So linguistic features and acoustic parameters are needed, before the training of acoustic model. In the front-end stage, Chinese phonetic annotations for the corpus are performed manually. First, Prosody is perceptually labeled according to the C-ToBI [10], but in our annotations rule, only four parallel tiers are labeled.

(1) Syntactic function tier: four sentence types are annotated for interrogative, imperative, declaratives and exclamation sentences

(2) Pin Yin tier: Pinyin and tone for each syllable are labeled, e.g. 1,2,3,4 and 5 for different tones.

(3) Initial and final tier: the initials including zero initials and finals of each syllable are annotated.

(4) Break index tier: Break is perceptually labeled for each syllable, in which break indices=1-4.

break index=1: the minimum break between syllables, usually breaks within a prosodic word;

break index=2: for prosodic word boundary;

break index=3: for minor prosodic phrase boundary

break index=4: for prosodic group boundary, usually breaks between sentences.

Certainly, phones duration and part-of-speech(POS) are needed , and they are labeled fine-tuned manually by our internal tool.

Method

In this section, we will introduce the brief outline of our proposed hybrid synthesis system. Same as the popular hybrid framework, our framework, as showed in Fig 1, consist of training stage and synthesis stage.

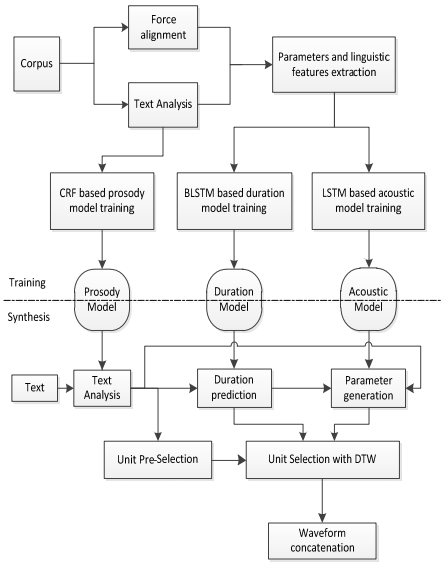

Figure 1. An overview of our system

Training Stage

There are there models that need training, including prosody model, duration model and acoustic model. Break annotations manually fine-tuned have been got, before training prosody model. CRFs are a type of discriminative undirected probabilistic graphical model. They are used to encode known relationships between observations and construct consistent interpretations and are often used for labeling or parsing of sequential data. Specifically, CRFs find applications in POS Tagging and Break index. Here, we use CRF++ as the prosody model training tool, in which text corpus is directly used as the input and the output is prosody annotations. The objective results of F1-mesure using CRF are presented in Table 1.

Table 1. Results of CRF

Break index	#1	#2	#3	#4
F1	0.8655	0.8033	0.9154	0.9998

In this paper, Merlin toolkit is used for features extracted [8]. Merlin is not a complete, so we build ourselves front-end and questions-set to get HTS-style labels. The Merlin toolkit converts the labels into binary vectors and continuous features for duration model and acoustic model training stage input.

At duration model training stage, phone-level duration got with our internal force alignment tools is used as the output and BLSTM is trained. There are 3 hidden bidirectional layers with 512 nodes each layer in the structure of BLSTM. We also use DNN model with 3 feedforward hidden layers and 2014 units in each hidden layer as a contrast. In Table 2, the objective results of BLSTM and DNN are presented, which confirms that BLSTM can produce better effect.

Table 2: Comparison of objective results between DNN and BLSTM. RMSE is calculated on a frame number

model	RMSE(frame)
DNN	6.21
BLSTM	4.38

At acoustic model training stage, WORLD[WORLD: a vocoder-based high-quality speech synthesis system for real-time applications] is used to extract the output acoustic features, including 60-dimensional Mel-Cepstral Coefficients(MCCs), 1-dimensional band aperiodicities(BAPs), and fundamental frequency on log scale(log F0) at 5 ms frame intervals. Taking into account dynamic factors, the output acoustic features also consist of deltas and delta-deltas feature, plus a voiced/unvoiced binary feature. At last, all acoustic features are normalized using min-max to the range [0.01, 0.09]. At this stage, we get three models here, including BLSTM and DNN using in

duration model, and LSTM stacked by 3 hidden layers with 512 units in each layer. The objective results of the 3 models are presented in Table3. The results in Table 3 prove LSTM and BLSTM can achieve less distortion than DNN. But considering of the size of the model, the amount of calculation and the little difference of effects between LSTM and BLSTM, we use LSTM structure to train acoustic model.

Table 3: MCD: Mel-Cepstral Distortion. BAP: distortion of band aperiodicities. F0 RMSE is calculated on a linear scale. V/UV: voiced/unvoiced error

model	MCD(dB)	BAP(dB)	RMSE(Hz)	V/UV
DNN	5.223	0.391	34.890	7.03%
LSTM	5.043	0.299	33.514	6.35%
BLSTM	5.041	0.301	33.501	6.33%

Synthesis Stage

Selection units step. At the synthesis stage, firstly, some toolkits such as the tool of word segmentation and the tool of Chinese characters to pinyin are used in the text analysis phase. Simultaneously, prosody is predicted using the CRF prosody model. Then phone duration is predicted by the duration model. Next, concatenate linguistic features from the text analysis phase and phone duration, which is used to predict the phone-level acoustic parameters.

In our system, Chinese character is the basic concatenation unit, so we need to concatenate phone-level acoustic parameters to get character-level acoustic parameters. DTW can calculate the similarity between time series with different time lengths. Therefore, we can use DTW to calculate the cost between target unit and candidate unit from pre-selection units. Every Chinese character will get their optimal candidate unit, and the waveform fragments of optimal units are concatenated. Before getting the whole final waveform, some smoothing processing is needed between the waveform fragments.

Improved smoothing algorithm. Cross-fade method has been used in many synthesis systems to smooth the phase discontinuity [9,11] at concatenation step. With cross-fade method, speech segments are concatenated in phase with fade-in/out of previous/post-segment ends. In conventional cross-fade, we only find the concatenation points with the smallest distortion in phase, without taking amplitude into account. But smallest distortion in phase not means the smallest distortion in amplitude. Here we present an improved cross-fade method.

It is crucial to find the concatenation points at the previous and post segment for getting natural speech. In order to get the optimal concatenation points, a minimum error cost criterion is proposed and several formulas are defined as follows:

$$\mathbf{C = \arg\min(w_v * D_v + w_r * D_r)} \quad (1)$$

$$\mathbf{D_v = | V_a - V_b |} \quad (2)$$

$$\mathbf{D_r = | k_a - k_b |} \quad (3)$$

Where **C** is the minimum error cost, $\mathbf{V_a}$ and $\mathbf{V_b}$ stand for the concatenation point amplitude respectively in the previous segments **a** and post segments **b**. $\mathbf{k_a}$ and $\mathbf{k_b}$ are the slope in the concatenation points. $\mathbf{w_v}$ and $\mathbf{w_r}$ are weights that are set manually.

There are 3 steps to be taken for getting the minimum error cost **C**:

1) The last 10ms waveform of the previous concatenation unit is selected as **a** and the waveform starting in 10 ms of the post concatenation unit is selected as **b**. Our speech materials are recorded in 16 kHz, so segment **a** and **b** length are both 160 points.

2) Fix the points in segments in turn and go through the points in segment **b** to get a sequence value of $\mathbf{C^*}$:

$$\mathbf{k_a = a_t - a_{t-1}} \quad (4)$$

$$k_b = b_t - b_{t-1} \quad (5)$$

3) After the second step, every point in segment **a** can match their optimal concatenation point in segment **b**. So the optimal concatenation point in both segment can be obtained when we take out the minimum error cost value **C** from $\mathbf{C}^*$.

After repeated test, we find the top optimization weight in formula (1):

$$w_v = 0.2 \quad (6)$$

$$w_r = 0.8 \quad (7)$$

At last, speech units are concatenated with cross-fade method introduced in [11].

Experiment

To demonstrate the performance of our hybrid system, we conduct an internal listening test. First, CRF and LSTM based SPSS is adopted for the baseline system. Therefore, three systems are compared:

A: baseline system.

B: hybrid system without improved smoothing algorithm.

C: hybrid system with improved smoothing algorithm.

30 sentences are synthesized by each system and 10 listeners, all of them are native Mandarin speakers, take part in the test. We experiment with two different sizes of corpus 5h and 10h to obtain the impact of synthesis quality. The results are shown in Fig 2 and Fig 3. It proves that the improved smoothing algorithm is very effective in the larger corpus and has verified the fact that concatenation synthesis needs a large amount of corpus.

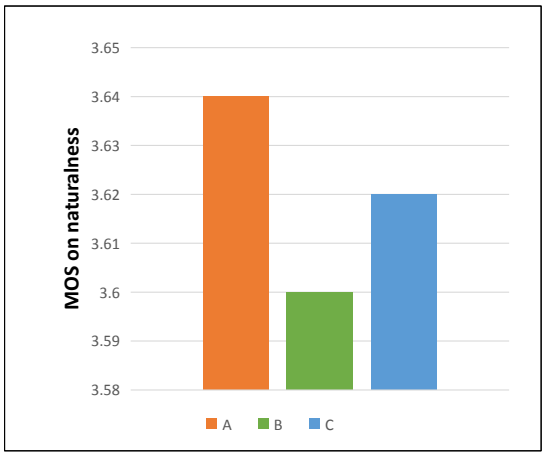

Figure 2. MOSs of three systems on the basis of 5h corpus

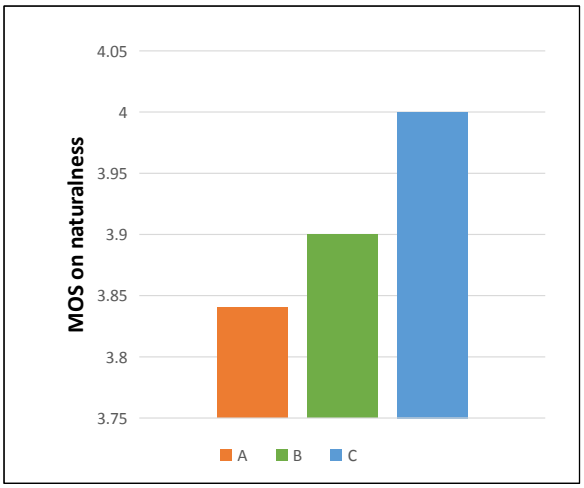

Figure 3: MOSs of three systems on the basis of 10h corpus

Conclusion

This paper presents a new unit selection and concatenation smoothing method which uses Chinese characters as basic unit and improved cross-fade technique. Our hybrid system takes non-continuously-valued cost into account while selecting directly the optimal units to obtain the final waveform. The experiment results show that the proposed system with improved cross-fade method can achieve better performance. The future work will focus on the appropriate expanding of corpus scale and prosody control technique.

Acknowledgement

The authors would like to thank the anonymous reviewers and text annotate for their helpful comments and suggestions.

References

[1] Henter, G. E., Ronanki, S., Watts, O., Wester, M., Wu, Z., & King, S. (2016, March). Robust TTS duration modelling using DNNS. In ICASSP (pp. 5130-5134).

[2] Chu, M., Peng, H., Yang, H. Y., & Chang, E. (2001, May). Selecting non-uniform units from a very large corpus for concatenative speech synthesizer. In icassp (pp. 785-788). IEEE.

[3] CHU, M., TANG, D., SI, H., TIAN, X., LU, S., & KONG, J. (1997). Research on perception of juncture between syllables in Chinese. Acta Acustica, 2, 001.

[4] Zhang, R., Tao, J., Li, Y., & Wen, Z. (2013, November). A novel unit selection method for concatenation speech system using similarity measure. In Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), 2013 International Conference (pp. 1-5). IEEE.

[5] Bellegarda, J. R. (2010). A dynamic cost weighting framework for unit selection text–to–speech synthesis. IEEE Transactions on Audio, Speech, and Language Processing, 18(6), 1455-1463.

[6] Bellegarda, J. R. (2010). A dynamic cost weighting framework for unit selection text–to–speech synthesis. IEEE Transactions on Audio, Speech, and Language Processing, 18(6), 1455-1463.

[7] Tao, J., Zheng, Y., Wen, Z., Li, Y., & Liu, B. (2016). A BLSTM Guided Unit Selection Synthesis System for Blizzard Challenge 2016.

[8] Wu, Z., Watts, O., & King, S. (2016). Merlin: An open source neural network speech synthesis system. Proc. SSW, Sunnyvale, USA.

[9] Ling, Z. H., Qin, L., Lu, H., Gao, Y., Dai, L. R., Wang, R. H., ... & Hu, G. P. (2007, August). The USTC and iFlytek speech synthesis systems for Blizzard Challenge 2007. In Blizzard Challenge Workshop.

[10] Li, A. (2002). Chinese prosody and prosodic labeling of spontaneous speech. In Speech Prosody 2002, International Conference.

[11] Hirai, T., & Tenpaku, S. (2004). Using 5 ms segments in concatenative speech synthesis. In Fifth ISCA workshop on speech synthesis.