A Study on How Human Annotations Benefit the TTS Voice

Min Chu, Yining Chen, Yong Zhao, Yusheng Li and Frank Soong

Microsoft Research Asia, Beijing, China
{minchu, ynchen, yzhao, yushli, frankkps}@microsoft.com

Abstract

When we built the unit inventory from the Blizzard corpus, three types of manual works were performed. All these works took about 12 working days of our labelers. In order to see how much benefit these manual works bring us, we performed several perceptual experiments to compare the speech generated with/without manual works. The results show that although the manual proofreading identified more than 500 word-errors, no improvement is observed in our experiment. Both manual checking of segmental boundaries and manual prosody annotations make the synthesized speech better. And the later one brings more benefit. The preference rate between the final version of the synthetic speech with limited manual works and the fully automatically processed version is 68% to 32%.

1. Introduction

In a concatenative text-to-speech system where appropriate waveform segments in a large speech database are selected to generate natural sounding speech, the naturalness of synthetic speech, to a great extent, depends on the quality of the unit inventory. The whole process of database collection and annotation is rather complicated and contains plenty minutiae that should be handled carefully. In the Blizzard Challenge, a high quality speech corpus is provided. Therefore, we need not care about the script generation and the speech recording. Then, the tasks that are remained for building a new voice include obtaining the phonetic transcriptions, the segmental boundaries, and the prosodic labels. There are several ways to generate such information fully automatically, yet, we believe that limited human interferences such as manually checking or labeling are helpful. In this paper, we present what kind of human interferences we performed and how much benefit we gained from them.

Altogether, three types of manual works are performed when we processed the speech corpus, including error detection of recording script, checking of segmental boundary and prosodic annotation.

The TTS corpus is normally recorded with carefully monitoring. Yet, when generating the phonetic transcription for the speech corpus, we still found some mismatches between the recorded speech and the script. These mismatches are caused by reading errors, text-normalization errors, letter-to-sound errors or idiosyncratic pronunciations. These errors can degrade the TTS speech quality when a problematic unit is selected for synthesis. In our experience of the Blizzard Challenge 2006, we found that about 1% words not conform to their speech waveforms in either orthography or phone layer. Therefore, we identify some problematic sentences and checked them manually.

To make a speech corpus usable to a concatenative TTS, the phonetic transcriptions have to be aligned with the corresponding speech waveforms. HMM based forced alignment has been widely adopted for automatically boundary alignment [1]. Yet, despite its universal maximum likelihood and relatively consistent segmentation output, such a method can not guarantee the automatic boundaries are optimal for concatenation-based synthesis. Thus, post-refinement of boundaries is often performed to adjust the boundaries for optimal speech synthesis [2, 3]. In our previous study, we have proposed to use context-dependent boundary models [4] to fine tune the locations of segmental boundaries. This approach needs a small amount of manually labeled boundary references to train the refining model.

In order to achieve high quality synthetic speech, prosody annotation is often performed on the speech corpus, either manually or automatically. In most TTS systems, there is a prosody prediction module that predicts either categorical prosodic features, such as phrase boundary locations, boundary tone and pitch accent locations and types, or numerical features such as pitch, duration and intensity. Such prediction modules can be used to generate the prosody annotation for a speech corpus. However, the prediction based upon text may not match well the actual acoustic realizations. In [5], we have proposed a multi-classifier framework for automatic prosody annotation, in which the appearance of a prosodic event is jointly decided by an acoustic classifier, a linguistic classifier and a combined classifier. To train such a prediction model, a set of easy manipulated prosodic events have been defined and labeled manually.

In order to verify the validity of performing such manually checking and labeling, in this paper, several perceptual experiments are carried out to compare the speech generated with/without without manual works.

The paper is organized as follows. In Section 2, the framework of our TTS system is introduced. The details on generating the three types of annotations with/without human interferences are described in Section 3. In Section 4, perceptual experiments are introduced to investigate the benefits from different human interferences. The final conclusion is drawn in Section 5.

2. System overview of Mulan TTS system

Our TTS system Mulan [6] is a phone based concatenative speech synthesis system, in which prosody is modeled under a soft prediction strategy [7]. Unlike the traditional deterministic way to predict prosodic targets by maximizing the likelihood of the training tokens with respect to the model parameters, the soft prediction generates acceptable regions by minimizing the probability of violating the invariant property in prosody. The

output of the soft-prediction prosody model is not the best path (or the most likely path) in the feature space. Instead, acceptable regions are marked by eliminating paths which violate the invariant property. With such a soft prediction strategy, the categorical targets for pitch and duration instead of the numerical targets are predicted first. And, such prosodic constraints are imposed with the highest priority in unit selection in order to get the right prosody, i.e. a prosodic-constrained unit selection algorithm is used.

In this unit selection approach, the stylized invariance of prosody is captured by clustering all tokens of a base unit with a CART (Classification And Regression Tree), wherein querying only their prosodic constraints, such as the stress level, break level, and position in phrase and word, etc. The splitting criterion for CART is to maximize reduction of the weighted squared error of three features: average f_0 (fundamental frequency), dynamic range of f_0 and duration. Such a clustering is quite similar to that used in the CART based prosody prediction model of a traditional TTS system. All units on the same leaf node share common prosodic constraints. What different in our approach is that the mean value of a leaf node in the CART is used as a reference instead of the prosody target of the cluster of tokens. A token which is away from the reference by more than a pre-specified distance threshold is rejected. All tokens within the distance threshold are remained and considered prosodically equivalent in unit selection. For a pre-defined base unit set, such a tree is built for each base unit and served as the index for prosodic characteristics of all tokens of the base unit.

During speech synthesis, a cluster of prosodically equivalent tokens is first selected for each base unit by querying the CART with the target prosodic constraints. All tokens on all selected leaf nodes form a segment lattice. If the speech database is large enough that covers all types of variations represented by the prosodic constraints, all tokens on the same leaf node will have the same prosodic constraints and there will always be a leaf node that matches the target prosodic constraints exactly available for each target unit. Then, only segmental constraints need to be considered in calculating the target cost. However, due to data sparse issue, the CART will cluster instances with similar constraints into the same leaf nodes. Therefore, prosodic constraints are still used in calculating target cost to rank the candidates. The target cost is defined as the weighted sum of the source-target distances of all prosodic constraints and segmental constraints, as illustrated in equation (1) and (2)

$$C_T = \sum_{i=1}^{I} w_{pi} C_{Pi} + \sum_{j=1}^{J} w_{Sj} C_{Sj} \quad (1)$$

$$\sum_{i=1}^{I} w_{pi} + \sum_{j=1}^{J} w_{Sj} = 1 \quad (2)$$

Where, C_{Pi} and C_{sj} are the source-target distances of the *i-th* prosodic constraint and the *j-th* segmental constraints, respectively; w_{pi} and w_{Sj} are the weights corresponding; I and J are the total numbers of prosodic constraints and segmental constraints used in unit selection.

For defining the transition cost between two adjacent tokens, the continuity for splicing two segments is quantized into four levels: 1) continuous — if the two tokens are continuous segments in the unit inventory, the splicing of them will be very natural, therefore the target cost is set to 0; 2) semi-continuous — though the two tokens are not continuous segments in the unit inventory, the discontinuity at their boundary are not often perceptible, for example, the splicing of two voiceless segments (such as /s/+/t/) belongs to this level, a small cost is assigned; 3) weakly discontinuous — discontinuity across the concatenation boundary is often perceptible, yet not very strong, for example, the splicing between a voiced segment and an unvoiced segment (such as /s/+/ a:/) or vice versa belongs to this level, a moderate cost is used; 4) strongly discontinuous — the discontinuity across the splicing boundary is perceptible and annoying, for example, the splicing between voiced segments belongs to this level, a large cost is assigned. The first two types of splicing are preferred in unit selection and the 4th type should be avoided as much as possible. The overall cost for a path in the unit lattice is then defined as the sum of target costs of all tokens along the path plus the sum of the transition costs between two adjacent tokens.

3. Data processing with limited manual works

Since the schedule for Blizzard Challenge is very tight, we only arrange limited manual works in proofreading problematic sentences, checking the segmental boundaries and labeling prosodic events. The details are introduced below.

3.1. Phonetic transcription

First, we used our Mulan front-end to generate the phonetic transcription of the speech corpus fully automatically. We found that there are some mismatches between the speech and the transcription, which are caused by reading errors, text-normalization errors or letter-to-sound errors. However, it is not realistic to proofread all scripts. We identified several small groups of problematic sentences with different focuses for manually reviewing.

The first group focuses on checking the pronunciation of polyphonic words. We have developed an interactive tool, with which the human labeler can listen to all instances of a word in the corpus and he/she can change the phone strings when necessary. The second group is to check the pronunciation of out-of-vocabulary words, abbreviations, acronyms and words with multiple capital letters.

It took us about three working days to finish the proofreading and 534 words are corrected. Thus, we obtained a fully automatically processed phonetic transcription and a transcription with limited manual checks.

3.2. Phonetic segmentation

In order to align the phonetic transcriptions with the corresponding speech waveforms, HMM-based forced-alignment were applied to the whole speech corpus first. Then, we have the Arctic part of the corpus checked manually. It took the labeler about 5 working days.

We used 20,000 hand-labeled boundaries to train the context-dependent boundary models [4] and refined the boundary locations in the remaining data with them. The goodness of the boundary models are tested in 10000 manually labeled boundaries. The boundary accuracy (if the distance from an auto-boundary to its manually labeled reference is smaller than 20ms, it is counted as a correct one) is 90.6% after refining. The accuracy of forced-aligned boundaries is only 77.6%. A rather significant improvement is achieved.

3.3. Prosody annotation

Two types of prosodic events are normally labeled in a TTS speech corpus, the phrase boundary (w/o boundary type) and the pitch accent (w/o accent type). ToBI [8] is a widely adopted prosodic representation. It is first proposed for English and has been extended in many languages. However, annotating a speech corpus with ToBI is a very difficult task even for professionals. It will take even experienced labelers from 100 to 200 times real time [9]. The across personal agreement ratio for accent, edge tone and boundary indices are rather low (reported as 71%, 86%, and 74%, respectively, in [10]) and the agreement ratio on the presence and absence of accent and edge tone are much higher (92% and 93%, respectively). Therefore, a simple version prosody representation, ToBI lite [11], is proposed recently. However, we think ToBI lite is too much compressed. The pitch movements at phrase boundaries play an important role in unit selection. Therefore, we designed a set of prosodic events with complexity between ToBI and ToBI lite. It includes two-level boundary strengths (correspond to the minor phrase and the major phrase boundaries), five boundary types (full rise, minor rise, full fall, minor fall, and flat, corresponding to the perceptual pitch movement before the boundary) and two-level accents (accented or not). All these prosodic events have perceivable cues so that a well trained human annotator can achieve good self-consistency. In our experiment in English, the same annotator labeled the same sentences twice in a four-week time span. The agreement ratio on presence or absence of accent is 95%, on boundary strength is 93.5% and on boundary strength plus boundary type is 90%. After the training section, labeling all these prosodic events with our tool takes about 5-10 times real time. It took our labelers four working days to label the whole Blizzard corpus.

4. Perceptual study of benefits from manual works

During the data processing, we have corrected more than 500 word-errors in the phonetic transcription, increased the boundary accuracy from 77.6% to 90.6% and labeled prosody events in the corpus with about 12 days of manual works from our labelers. We want to know how much benefit we have achieved from these works. Therefore, perceptual experiments are performed to compare the speech synthesized with/without these manual labels. All together, 5 unit inventories are built. Details are given in the Table 1.

The naturalness testing set of the Blizzard Challenge (50 sentences from Novel, Conversation and News, respectively) is first synthesized with all the five unit inventories. The instances generated with the unit inventory A, B, C and D are compared with the those generated with the unit inventory E separately. Totally, 16 subjects participated in the experiments. They listened to 50-150 pairs of utterances that randomly selected from all the comparing pairs and were forced to make a choice that either the first or the second sentence in each pair sounds more natural. Finally, we got at least 200 votes for each comparing group. The results of all the comparing groups are given below.

Table 1. Five unit inventories built with different configurations

Unit inventories		A	B	C	D	E
Phonetic transcription	Auto-generated	√	√			
	Manually checked			√	√	√
Unit Segmentation	Forced-aligned	√		√		
	Post refined		√		√	√
Prosody annotation	Auto-generated	√			√	
	manually annotated		√	√		√

4.1. With/without human proofreading

Figure 1 gives the user preference between sentences synthesized with unit inventory E and B. It is interesting that although we corrected more than 500 word-errors, no benefit is observed from the experiment. A possible reason is that few units used in the two versions are from the words with errors.

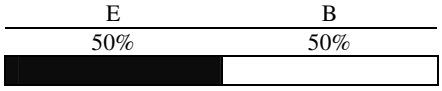

Figure 1. *Preference result for with/without manually proofreading (E—with; B—without)*

4.2. With/without boundary post-refining

Figure 2 shows the user preference between sentences synthesized with unit inventory E and C. Some improvements are observed. And the improvements are statistically significant ($p<0.01$). We can conclude that more precise segmental boundaries benefit the synthetic voice quality.

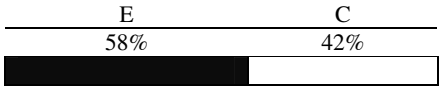

Figure 2. *Preference result for segmental boundaries with/without post-refining (E—with; C—without)*

4.3. With/without manually prosody labeling

The preference result between automatically generated prosody labels and the manually created prosody labels is given in Figure 3. Larger improvement is achieved ($p<0.0005$). This implies that accurately labeling the prosody events is very important from achieving high naturalness in synthetic speech in a concatenation-based TTS system.

Figure 3. *Preference result for auto vs. manual prosody annotations (E—manual; D—auto)*

4.4. Fully automatic version vs. the final version

Figure 4 illustrated the preference rate between the fully automatically processed version and the final version with some manually works. Significantly improvements are observed. Therefore, we can conclude that with limited manual works, the voice quality of out TTS system is significantly improved.

Figure 4. *Preference results for data process with/without manual works (E—with; A—without)*

5. Conclusion

In this paper, the fully automatically processed unit inventory is compared with the unit inventory that was processed with limited manual works. The perceptual results show that more accurate segmental boundaries and more precise prosody annotations can improve the naturalness of synthesized speech. With proper learning algorithms, only limited manual labeling can improve the final results significantly.

Although the experiment result shows that by correcting the 1% errors in script, no measurable improvement can be found. We still think it is necessary to get rid of these errors. According to our results, proper labeling the prosodic events brings the most significant improvements. In current stage, the whole corpus is labeled manually. In next step, we will work on improving the prediction model with fewer manual data. The results also confirm the effectiveness of the prosodic event set we designed.

6. Acknowledgements

The authors would like to thank XiaoLin Song and Jay Waltmunson for arranging the listening test and all participants of the experiments.

7. References

[1] Toledano, D. T., Gómez, A. H.: Automatic Phonetic Segmentation, IEEE Trans. Speech and Audio Processing, Vol. 11 (2003) 617-625

[2] Kominek, J., Bennet, C., Black, A. W.: Evaluating and Correcting Phoneme Segmentation for Unit Selection Synthesis, Proc. Eurospeech-2003, Geneva, 313-316

[3] Adell J., and Bonafonte, A.: Towards Phone Segmentation for Concatenative Speech Synthesis, Proc. 5th ISCA Speech Synthesis Workshop, (2004)

[4] Wang, L.J., Zhao, Y., Chu, M., Soong, F. K., Zhou, J. L., Cao, Z. G.: Context-Dependent Boundary Model for Refining Boundaries Segmentation of TTS Units, IEICE Transactions on Information and System, Vol. E89-D, N0. 3 (2006) 1082-1091

[5] Chen, Y. N., Lai, M., Chu, M., Soong, F. K., Zhao, Y., Hu, F. Y.: Automatic Accent Annotation with Limited Manually Labeled Data, Proc. Speech Prosody 2006, Dresden

[6] Chu, M, Peng, H., Zhao, Y., Niu, Z and Chang, E.: Microsoft Mulan — a bilingual TTS systems, *Proc. ICASSP2003*, Hong Kong, 2003

[7] Chu, M., Zhao, Y. and Chang, E.: Modeling Stylized Invariance and Local Variability of Prosody in Text-to-Speech Synthesis, Speech Communication, Vol. 48, Issue 6, 2006, pp.716-726

[8] Beckman, M. and Ayers Elam, G., Guidelines for ToBI Labeling, Version 3, 1997

[9] Syrdal, A.K., Hirschberg, J., McGory J., Bechman, M.: Automatic ToBI Prediction and Alignment to Speed Manual Labeling of Prosody, Speech Communication, Vol. 33, No. 1, (2001) 135-151

[10] Syrdal A. K. and McGory, J.: Inter-Transcriber Reliability of ToBI Prosodic Labeling, Proc. ICSLP-2000, Beijing

[11] Wightman, C. W., Syrdal, A. K., Stemmer, G., Conkie A., Beutnagel, M.: Perceptually Based Automatic Prosody Labeling and Prosodically Enriched Unit Selection Improve Concatenative Text-to-Speech Synthesis, Proc. ICSLP-2000, Beijing.