MICROSOFT MULAN — A BILINGUAL TTS SYSTEM

Min Chu, Hu Peng, Yong Zhao, Zhengyu Niu and Eric Chang

Microsoft Research Asia, Beijing, China, 100080

ABSTRACT

This paper describes a bilingual text-to-speech (TTS) system, Microsoft Mulan, which switches between Mandarin and English smoothly and which maintains the sentence level intonation even for mixed-lingual texts. Mulan is constructed on the basis of the *Soft Prediction Only* prosodic strategy and the *Prosodic-Constraint Orient* unit-selection strategy. The unit-selection module of Mulan is shared across languages. It is insensitive to language identity, even though the syllable is used as the smallest unit in Mandarin, and the phoneme in English. Mulan has a unique module, the language-dispatching module, which dispatches texts to the language-specific front-ends and merges the outputs of the two front-ends together. The mixed texts are "uttered" out with the same voice. According to our informal listening test, the speech synthesized with Mulan sounds quite natural. Sample waves can be heard at: http://research.microsoft.com/~echang/projects/tts/mulan.htm.

1. INTRODUCTION

There is a long tradition of research on speech synthesis technology. With applications such as spoken dialog systems, call center services, voiced-enabled web and email services that are being introduced, an increasing emphasis has been put on generating naturally-sounding speech in recent years. There have been clear improvements in TTS by all quality measures in the past few years. However, most TTS systems can only handle a single language, which is often not adequate since many applications need to deal with multiple languages. Therefore, multilingual TTS systems are in great demand. The few multilingual TTS systems developed [1][2][3], however, can only deal with a single language at each call. For those systems, switching between languages requires switching between TTS engines.

In our usability study of Mandarin TTS, the lack of ability to handle English words and phrases embedded in Chinese text deters the adoption of TTS technology, since much Chinese content, especially IT-related articles or emails, contain English words, phrases, or names. Like the multilingual systems mentioned above, we could solve this problem by switching between two TTS engines. The main drawback of this approach is that voices coming out of the two engines usually sound different. Users are always annoyed when hearing such two-voice utterances. Furthermore, switching between two engines will destroy the overall sentence intonation. For example, if the sentence "我用 OfficeXP 写文章" ("I write articles with OfficeXP") is sent to a Mandarin TTS engine and an English one, respectively, the output will sound like three independent sentences which are "我用 (I use)", "OfficeXP" and "写文章 (write articles)". In this paper, we present a bilingual TTS system, Microsoft Mulan, which can switch between the two languages freely and smoothly without fragmenting the sentence level intonation. The two languages are spoken out with the same voice and sound like having been spoken by a bilingual person.

The organization of this paper is as follows. The unified strategies for prosody and unit selection are presented in Sections 2 and 3 respectively. The architecture of Mulan and its other main components are described in Section 4. Section 5 provides the final discussion.

2. PROSODY STRATEGY — SOFT PREDICTION ONLY (SPO)

Conventionally, TTS systems have a prosody model that takes some high-level prosodic constraints, such as part-of-speech (POS), phrasing, accent and emphasis etc., as input and makes hard predictions on pitch and duration, i.e. generating deterministic values for them. Such a prosody model can be realized with a set of rules [4][5], a statistical model [6][7], or a neural network [8]. However, when investigating the human prosodic behavior over a large speech corpus, we found that there are many important variations in prosody that are difficult to address with such a hard prediction model. We conducted a study on syllable duration over a very large Mandarin speech corpus containing 190,000 syllables, in which five duration-related features are considered. All these features take category values and result in 1000 possible combinations, i.e. there are 1000 cells in the feature space. Since all units in the same cell are indistinguishable by their features, they will be represented by the same value no matter what kind of prediction scheme is used. To get the minimum RMSE (Root Mean Square Error), the best representative of each cell should be the mean of the cell. The RMSE calculated in this situation is 41 mili-seconds in our investigation, and this is the bottom limit for any prediction model using the five features as input. Comparing with the average syllable duration over the whole corpus, which is 245 milliseconds, this means that the best hard prediction model will still result in 10-20 percent prediction errors in duration with the current feature set. This phenomenon can be viewed from two sides. On one hand, the RMSE can be reduced by considering more features or using more categories for each feature. However, when the feature space is partitioned into smaller and smaller cells, the generality of the prediction model will become poorer and poorer. It's very difficult to implement a very precise predictor. On the other hand, we can imagine that if a time scaling algorithm is used to adjust the duration of all these syllables to the mean of their cells, the modified version of these utterances must sound worse than the original ones. In fact, the 10-20 percent variation in duration is crucial for naturalness. The same is true for pitch. If sentences similar in structure are all generated with the same pitch pattern, they will sound

monotonous and boring. Humans always generate some variations in prosody to make their speech expressive, while most TTS systems cannot.

We believe that the lack of proper variation in prosody is one of the main reasons why most state-of-the-art synthetic speech sound mechanical and not human-like. A very natural TTS system does not need a precise predictor. In fact, the pitch and duration of each segment in the synthesized utterance have a range of reasonable values. Thus, we proposed a Soft Prediction Only strategy for prosody, i.e. prosodic features such pitch and duration of most units in the same prosodic cell, i.e. they share the same high-level prosodic constraints, are reasonable and should not be adjusted to the core of the cell. With this strategy, synthesized utterances will achieve richer intonation by inheriting the prosodic habit of the original voice talent. The SPO strategy has successfully guided us to construct a very natural Mandarin TTS system [9] and it is extended to English speech synthesis in this paper. Even though pitch accent is believed to play a more important role in English, a stress language, than in Mandarin, a tonal language, the SPO strategy has demonstrated promising results in the bilingual Mulan.

Another benefit of the SPO strategy is less artificial sounds. With traditional hard prediction prosody models, the targets of prosodic features are realized by adjusting the pitch and duration of selected units with scaling algorithms, such as PSOLA [10] or HNM [11]. Although these systems have the advantages of flexibility in controlling of prosody, they often suffer from significant quality decrease in timbre. Mechanical or reverberant sounds are two typical distortions that are heard in speech synthesized this way. In our approach, since no scaling is performed, no artificial sound is induced. The synthesized speech sounds just like the voice talent of the speech corpus.

It should be pointed out that all the benefits of SPO strategy will be achieved only when a good set of prosody-related features is used (these features can be derived from raw texts), when a prosodically enriched speech corpus is available, and when a powerful unit selection algorithm is developed. Although ample details on how to select the most suitable units from the unit inventory were presented in our previous paper [12], in the next section, the unit selection strategy is addressed again by answering the question of why it should be prosodic-constraint oriented under the SPO framework. A brief introduction of features used in unit selection is also included in Section 3. The speech corpus used is to be described in Section 4.

3. UNIT SELECTION STRATEGY — PROSODIC-CONSTRAINT ORIENTED (PCO)

In concatenative TTS systems, the unnatural sound in synthetic utterances originates from three main sources. First, the finite ability of prosody prediction models leads to irregular or flattening intonation, Second, the pitch and time scale algorithms result in buzzing or mechanical sound. And finally, the unsmooth splicing of units causes a cracking sound. In conventional systems, since the pitch and duration are adjusted toward their target values with scaling algorithms, the acoustic features of units and their phonetic context are considered as the most important factors that will affect the quality of synthesized speech. Thus, instances of a unit are first clustered by their phonetic contexts [13][14], and then they are pruned by their distances from the core of the cluster or by their HMM scores. In those systems, prosodic features are used only to select one from several instances within the same cluster. When the prosodic features of the selected unit do not match their predicted target, they will be scaled with signal processing methods. However, under our SPO prosody strategy, no pitch and duration scaling are performed. Prosodic constraints become the only restriction for getting natural prosody. Thus, the unit selection scheme should be prosodic-constraint oriented. In our approach, all instances of a unit are clustered first by their prosodic constraints such as the stress level, break level, and position in phrase and word etc. These features are used to be the input for hard-prediction prosody models in conventional systems, yet, they are used to predict a cluster of instances for a unit in our approach. All instances in the same cluster are considered to have reasonable prosody features after pruning some exceptional cases. The most suitable one is then picked out by considering the continuity of concatenations.

The degree of continuity for the splicing of two segments can be classified into 3 categories: 1) If the two segments are continuous segments in the unit inventory, they have very natural splicing. 2) Though the two segments are not continuous, if the spectral distance across the splicing boundary is small, no audible distortion occurs at the splicing boundary. This is a comfortable splicing. 3) The spectral distance across the splicing boundary is large. However, not all large distances are perceptible. We found that although the spectral distance for an unvoiced to unvoiced splicing is often large, fewer discontinuous sounds are audible. If a voiced segment is followed by an unvoiced segment or vice versa, the large spectral distance across the boundary is often imperceptible. Most of the concatenations in these situations are still comfortable. However, the chance for generating an annoying clicking sound increases for a voiced-voiced splicing. The natural splicing and the comfortable splicing are the kinds of concatenation we prefer. The unit selecting algorithm should pick out a series of segments from the prosodically reasonable pools of candidates to achieve the natural or comfortable splicing as much as possible.

The design of our unit selection algorithm is based on all considerations discussed above. The syllable is the smallest unit for Mandarin, and the phoneme for English. A total of 7 prosodic constraints are considered. They are: position in phrase; position in word; position in syllable; left tone; right tone; accent level in word; and emphasis level in phrase. Among them, position in syllable and accent level in word are effective only for English, and right/left tone are effective only for Mandarin. All instances for a base unit are clustered using a CART (Classification And Regression Tree) by querying about the prosodic constraints. The splitting criterion for CART is to maximize the reduction in the weighted sum of the MSEs (Mean Squared Error) of the three features: the average f_0, the dynamic range of f_0, and the duration. The MSE of each feature is defined as the mean of the squared distances from the feature values of all instances to the mean value of their host leaves. After the trees are grown, instances on the same leaf node have similar prosodic features. Two phonetic constraints, the left and right phonetic contexts and a smoothness cost, are used to assure the continuity of the concatenation between units. A concatenative cost is defined as the weighted sum of the source-target distances of the 7 prosodic constraints, the 2 phonetic constraints and the smoothness cost. The distance table for each prosodic/phonetic constraint and the weights for all components are first assigned

manually and then tuned automatically with the method presented in [15]. When synthesizing an utterance, prosodic constraints are first used to find a cluster of instances (a leaf node in the CART tree) for each unit, then, Viterbi search is used to find the best instance for each unit that will generate the smallest overall concatenative cost. The selected segments are then concatenated one by one to form a synthetic utterance.

Our Mandarin TTS system based on the SPO strategy and PCO strategy has generated very natural sounding speech [16]. The two strategies are extended to English in this paper. Although the two languages adopt units in different size, they share the same unit selection algorithm and the same set of features for units. Therefore, the back-end (or the unit selection algorithm) in Mulan is language independent, and it can process unit sequences in a single language or a mixture of the two languages. However, there are still many front-end processes that are language specific. The architecture of Mulan is given in the next section.

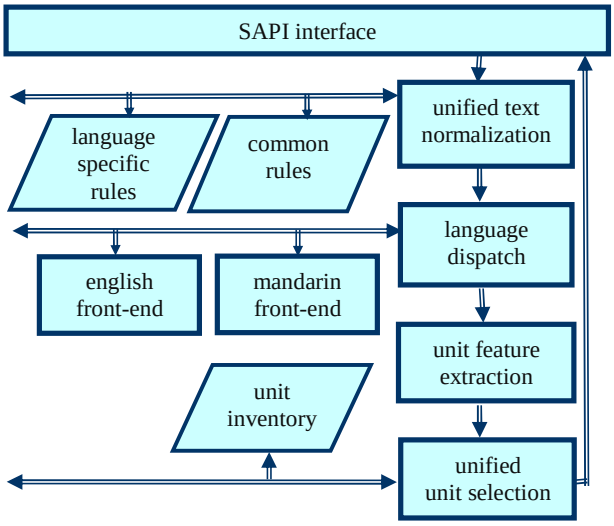

Figure 1. Architecture of Microsoft Mulan

4. ARCHITECTURE OF MULAN

Sproat [3] defined the ideal "multilingual" system as that all language-specific information should be stored in data tables and all algorithms should be shared by all languages. However, this ideal state is hard to achieve in practice since languages in different phyla, such as English and Mandarin, face quite different problems that are not easy to solve with the same algorithm. Thus, most multilingual systems have a common framework for all languages; yet, language-specific modules are still used. Our bilingual Mulan faces the same challenge. Furthermore, besides achieving the maximum sharing of components across languages, flexibility and extensibility are also very important features that should be considered during architecture designing. Although only two languages are to be handled in our current plan, it should not be difficult to add new languages in the future. With all these guidelines in mind, the architecture of the bilingual Mulan is designed as shown in Figure 1. It's a Microsoft SAPI5.0/5.1 compatible [17] TTS engine. Raw texts or XML tagged texts are passed into the engine through the SAPI interface. The input text is first processed by the text-normalization module, which changes numbers (date, time and money, etc.) and symbols into readable text strings. This module is a language independent rule interpreter, although rules are normally language-specific. The output of the text-normalization module is then passed to the language-dispatching module, which assigns language IDs to sentences and words. The Mandarin front-end and the English front-end deal with all language-specific processes, such as phrasing and grapheme-to-phoneme conversion for both languages, word segmentation for Mandarin and abbreviation expansion for English. They are developed separately by experts of each language. The outputs from the two front-ends are merged together according to their original sequence, and then are converted into unit features by the feature-extraction module. It is very important that the unit features are calculated after they are merged back into the same sentence. Only at this stage, the sentence level prosodic constraints can be obtained. At last, the unit sequence and their features are sent to the unit selection module to find the most suitable series of segments for concatenating.

In section 3, we have introduced the unit features and unit selection approach that is adopted. The remaining modules in our system are described below.

4.1 The unified text-normalization module

The text-normalization module is a language independent rule interpreter. It has two components. One is a pattern identifier. The other is a pattern interpreter which converts a matching pattern into a readable text string according to the rules. Each rule has two parts too. One is the definition of a pattern, and the other is the converting rule for the pattern. The definition part can either be shared by both languages or be specified to one of them. The converting rules are always language specific. Table 1 gives an example of a rule for date. The pattern in this rule has 5 components. Item1 and Item3 are integers between 1 to 12 and 1 to 31 respectively. Item5 is a four-digit integer. The three numbers are delimited by two '/'. For example, 2/16/1998. The Chinese interpretation for the pattern is that Item5 (explained as number one by one) followed by "年(year)", then followed by item1 (as cardinal number), then by "月(month)", then by item3 (as cardinal number) and then by "日(day)" at the end. The English interpretation for the pattern is that item1 (as a name for month) followed by item3 (as an ordinal number), then by item5 (as a number for years). The example above is converted to "一九九八年二月十六日" in Chinese and "February sixteenth nineteen ninety eight" in English, respectively. If a new language is to be added, the rule interpreting module does not need to be changed. Only new rules for the new language should be added.

Table 1: An example of a rule for date

Rule name: date07_2
Rule pattern: Item1 (Pp) TOKEN_INT (F) isMonth; Item2 (Pp) TOKEN_EM_DASH Item3 (Pp) TOKEN_INT (F) isDay; Item4 (Pp) TOKEN_EM_DASH Item5 (Pp) TOKEN_INT (L) =4;
Chinese interpretation:

Item5 (Pnt) IdenNum1 && Item0 (Pnc) 年
Item1 (Pnt) Cardinal && Item0 (Pnc) 月
Item3 (Pnt) Cardinal && Item0 (Pnc) 日

English interpretation:
Item1 (Pnt) Month
Item3 (Pnt) Ordinal
Item5 (Pnt) YearNum

4.2 The language-dispatch module and unit feature extraction module

The language dispatching module is very unique in our bilingual Mulan. It assigns language IDs to sentences or words. On sentence level, three modes are processed currently. They are pure English, pure Chinese, and Chinese with embedded English. English with Chinese embedded is not considered now. In the Chinese with English embedded mode, each word in the sentence is assigned an ID (either Chinese or English). Sentences or words are then dispatched to the corresponding front-end. The outputs from the two front-ends are merged back according to their original sequence in a sentence. Thus, all sentence level information will be kept. The 7 prosodic features and 2 phonetic features of all units (Chinese syllables or English phonemes) in the sentence are then extracted by the feature extraction module. They will be used during unit selection.

4.3 The speech corpus

A bilingual speech corpus collected at MSR Asia is used. It contains approximately 15,000 Mandarin utterances and 10,000 English utterances. The whole corpus is read by a professional female voice talent who is a native Chinese speaker and can speak English fluently. Forced alignment is performed to label the syllable boundaries in Mandarin and phoneme boundaries in English. Each unit in the corpus is indexed by a PCO CART.

The English utterances, the Mandarin utterances, and the mixed utterances generated by Mulan sound quite natural according to our informal listening test. Sample waves can be heard at:
http://research.microsoft.com/~echang/projects/tts/mulan.htm.

5. DISCUSSION

The SPO prosody strategy and PCO unit selection strategy were first proposed for Mandarin TTS in our previous study and were extended to English in this paper. Very natural utterances have been generated in both languages. Since Mandarin is representative of tonal languages and English represents the stress languages, the successful implementation of the SPO and PCO strategy in both languages shows great potential of extensibility of the two strategies. They will be applied to more languages in our future studies.

The two strategies have some limitations, though. First, their success depends heavily on the quality of the speech corpus including the prosodic and phonetic coverage of units and the consistency of recording environment. Thus corpus design and collection are very important for constructing a natural TTS system. Defects in speech corpus are difficult to recover with post-processing. Secondly, the speaking style of the speech corpus sets an upper limit for the synthesized speech. It is almost impossible to synthesize speech beyond the style of the original speech corpus under the no scaling framework. In our future study, generating speech in different styles will be one of the topics.

6. REFERENCES

[1] P. B. Mareüil and B. Soulage, "Input/output normalization and linguistic analysis for a multilingual text-to-speech Synthesis System", *Proc. of 4th ISCA workshop on speech synthesis*, Scotland, 2001.

[2] http://www.research.att.com/projects/tts/.

[3] R. Sproat, (Editor), *Multilingual text-to-speech synthesis: the Bell Labs approach*, Kluwer Academic Publisher, 1998.

[4] D. H. Klatt, "The Klattalk text-to-speech conversion system", *Proc. of ICASSP'82*, pp. 1589-1592, 1982.

[5] H. Fujisaki, K. Hirose, N. Takahashi and H. Morikawa, "Acoustic characteristics and the underlying rules of intonation of the common Japanese used by radio and TV announcers", *Proc. of ICASSP'86*, pp. 2039-2042, 1986.

[6] K. N. Ross and M. Ostendorf, "A dynamical system model for generating fundamental frequency for speech synthesis", *IEEE transactions on speech and audio processing*, Vol.7, No. 3, pp. 295-309, 1999.

[7] J. R. Bellegarda, K. Silverman, K. Lenzo, and V. Anderson, "Statistical prosodic modeling: from corpus design to parameter estimation", *IEEE transactions on speech and audio processing*, Vol.9. No.1, pp. 52-66, 2001.

[8] S. Chen, S. Hwang and Y. Wang, "An RNN-based prosodic information synthesizer for Mandarin text-to-speech", *IEEE transactions on speech and audio processing*, Vol.6, No.3, pp. 226-239, 1998.

[9] M. Chu, H. Peng and E. Chang, "A concatenative Mandarin TTS system without prosody model and prosody modification", *Proceedings of 4th ISCA workshop on speech synthesis*, Scotland, 2001.

[10] E. Moulines and F. Charpentier, "Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones", *Speech Communication* Vol. 9, pp. 453-467, 1990.

[11] Y. Stylianou, T. Dutoit, and J. Schroeter, "Diphone concatenation using a harmonic plus noise model of speech", *Proc. of Eurospeech '97*, pp. 613-616, Rhodes, 1997.

[12] M. Chu, H. Peng, H. Yang and E. Chang, "Selecting non-uniform units from a very large corpus for concatenative speech synthesizer", *Proc. of ICASSP'2001*, Salt Lake City, 2001.

[13] X. D. Huang, A. Acero, J. Adcock, *et al*, "Whistler: a trainable text-to-speech system", *Proc. of ICSLP'96*, Philadelphia, 1996.

[14] R. E. Donovan and E. M. Eide, "The IBM trainable speech synthesis system", *Proc. of ICSLP'98*, Sydney, 1998.

[15] H. Peng, Y. Zhao and M. Chu, "Perpetually optimizing the cost function for unit selection in a TTS system with one single run of MOS evaluation", *Proc. of ICSLP'2002*, Denver, 2002.

[16] M. Chu and H. Peng, "An objective measure for estimating MOS of synthesized speech", *Proc. of Eurospeech'2001*, Aalborg, 2001.

[17] http://www.microsoft.com/speech/techinfo/compliance/