Analysis of Chinese Interrogative Intonation and its Synthesis in HMM-Based Synthesis System

Yongxin Wang, Jia Jia and Lianhong Cai
Key Laboratory of Pervasive Computing, Ministry of Education
Tsinghua National Laboratory for Information Science and Technology(TNList)
Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
wangbb33@126.com

***Abstract*— In the training process of HMM-based speech synthesis systems, the states of the HMM models are clustered based on the context features labeled on the training corpus. A good feature set would enable the synthesis system to generate natural prosody. In this paper, in order to synthesize interrogative intonation, we analyzed the acoustic features of the interrogative intonation and the context features that are related to the interrogative intonation in Chinese. We then expanded the feature set used conventional HMM-based speech synthesis system with interrogative intonation related features. A new speech synthesis system is trained with the proposed feature set. The analysis of train and synthesis results show that the proposed feature set is able to discriminate declarative and interrogative intonations. Perception listening test also shows the synthesis system using the expanded feature set can generate interrogative intonation.**

***Index Terms*— HMM-base Synthesis, Intonation, Interrogative Sentence, Context Feature**

I. Introduction

The interrogative intonation is the intonation used in questions. It is reported that the boundary tone is a very important difference between Chinese interrogative intonation and declarative intonation [1]. It is also noticed that the overall pitch level may be raised in interrogative sentences [2]. For question with question words or final particles, the interrogative intonation would have more variations [3]. In speech synthesis systems, it is necessary to cache all these variations to generate natural interrogative sentences.

HMM-based speech synthesis is flexible in generating different kinds of prosody features. The HMM based speech synthesis systems use decision tree to cluster HMM parameters with context features [4]. The design of this context feature set would affect the ability of the synthesis system to generate different variations of prosodic feature.

In HMM-based speech synthesis, the context feature set generally includes phonetic features and prosodic features [5]. ToBI endtone of phrase is also introduced into the context feature set in some synthesis systems [4]. However, using ToBI endtone requires that the ToBI indexes are labeled in training corpus, and a separate ToBI index predictor must be developed.

In this paper, we proposed an expanded context feature set with purely textual features to capture the difference between variations of interrogative intonation and declarative intonation. We first analyzed the intonation of interrogative sentences from a speech corpus and the context features that may affect the interrogative intonation, then expanded the context feature set to include more context features that are related to the interrogative intonation. Train result is analyzed to check the ability of the proposed feature set to discriminate interrogative and declarative intonations. Perception test is carried out to verify the ability of synthesis system trained with the proposed context feature set to generate interrogative intonation.

II. Analysis of Interrogative Intonation

A. Typical Interrogative Intonation

In questions with no question words or final particles, the interrogative intonation becomes the only means to transfer the interrogative information. Typical interrogative intonation can be found in this kind of questions. It is reported that the rising boundary tone is important in transferring interrogative information [1]. To get a better picture of the boundary tone for interrogative sentences, a short sentence corpus is developed in our lab with contrastive recording of interrogative and declarative sentences.

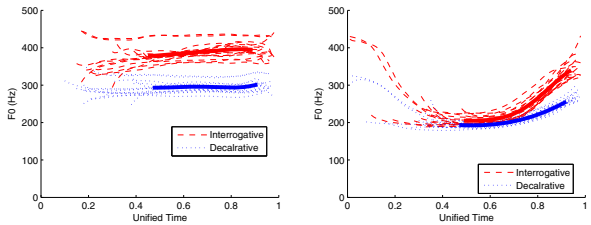
The texts for the corpus are short sentences in order to focus on the final boundary tone. All different tone combinations for the last two syllables are included in the corpus. The sentences in the corpus do not have any question words or final particles. Each sentence is performed by a female speaker using both declarative and interrogative intonations. When reading with interrogative intonation, instructions about what to be asked are given to the speaker, but not how to realize the intonation. A total of 128 pairs of contrastive recording are obtained.

The final syllable is the main carrier of the boundary tone in questions without a final particle. Fig. 1 shows the pitch contours of the final syllables with different lexical tones from interrogative and declarative sentences. The time is

*This work is supported by National Natural Science Foundation of China (61003094, 60805008, 90820304, 60928005, 60910130)

978-0-7695-4539-4/11 $26.00 © 2011 IEEE
DOI 10.1109/ICICIS.2011.89

unif ed to be 1 for each syllable. The front part of the pitch contour varies because of the effect of co-articulation. The f nal part of the syllable, or the core of the tone contour, converges in each intonation. The thick line in the f gure is the average of the core part of the pitch contour.



(a) 1st tone (b) 2nd tone

(c) 3rd tone (d) 4th tone

Fig. 1: Comparison of the pitch contours of the f nal syllables between interrogative and declarative sentences. The thick line in the f gure shows the average of the core part of the pitch contour.

Fig. 1 also shows that the rising boundary tone caused the top points of the pitch contour to rise, but not the bottom points. This can be clearly seen for the 2nd and the 4th tones. The 1st tone is a high f at tone. The pitch level of the f nal 1st tone rises in interrogative sentences, as it has no bottom point. The 2nd tone gets to a relatively low bottom point f rst with similar pitch values in both intonations, and then rises to the top point which is higher in interrogative intonation. The 3rd tone is a low tone, and it sometimes gets creaky in the middle of the syllable. So it is hard to get a proper pitch contour in the middle part of the 3rd tone. But as the tone f nally rises up, we can still see the tone rises to a higher value in interrogative sentences than in declarative sentences. The 4th tone is a falling tone and it gets to a higher top point in interrogative sentences before the core part, and falls down to a relatively low bottom point. This result illustrates the typical boundary tone in interrogative sentences.

B. Variations of Interrogative Intonation

The short sentence contrastive corpus aims to cache the characteristics of typical interrogative intonation. However, in real speech, interrogative intonation has many variations. Different types of questions, the existence of question words or f nal particles may all cause the question to be using a different variation of interrogative intonation. To get coverage of the different variations of interrogative intonation, the TH-CoSS corpus is used [6]. The TH-CoSS corpus contains recordings of declarative and interrogative sentences read by a female announcer. There is no contrastive recording in this corpus. In the interrogative part of the corpus, different question types, questions with or without question words or f nal particles can all be found. Fig. 2 shows the pitch contours of the f nal syllable from general questions and rhetorical questions which end with a 4th tone. It shows more variations than the short sentence corpus.

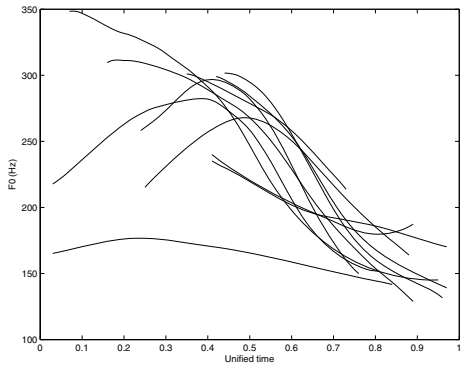


Fig. 2: Pitch contours of the f nal syllable from general questions and rhetoric questions which end with a 4th tone.

The interrogative f nal particle is also an important way to transfer interrogative information in Chinese questions. In the TH-CoSS corpus, there are sentences end with different f nal particles, like “ne”, “ba”, “ma”, etc. The f nal particles are all of the neutral tone. These f nal particles also affect the intonation of the sentence. Fig. 3 shows the pitch contours of three mostly used f nal particles in interrogative sentences from the corpus. The f gure shows that the particle “ba” has a lower ending pitch than others. Using average pitch as an indication of the pitch level of the f nal particles, we found the “ba” has an average pitch of 155.1 Hz. This is signif cant lower than the average pitch of “ma” (178.3 Hz) and “ne” (182.2 Hz) based on Student’s t-test, with a signif cance level of 0.01. As the f nal particle is part of the f nal boundary tone, this shows the effect of the f nal particle on the interrogative intonation.

The typical interrogative intonation has the rising boundary tone. But the interrogative intonation also has many variations depending on the sentence type and sentence structure.

III. Synthesis of Interrogation in HMM-based Synthesis Systems

We choose HMM-based speech synthesis to synthesize interrogative intonation because of its f exibility to control prosody features. With a well-designed context feature set used in training, an explicit prosody model is not needed in

the synthesis system, but trained inexplicitly in the HMM models. In this paper, we added some additional features to the context feature set used in Chinese HMM synthesis system to enable the synthesis system to generate interrogative intonations.

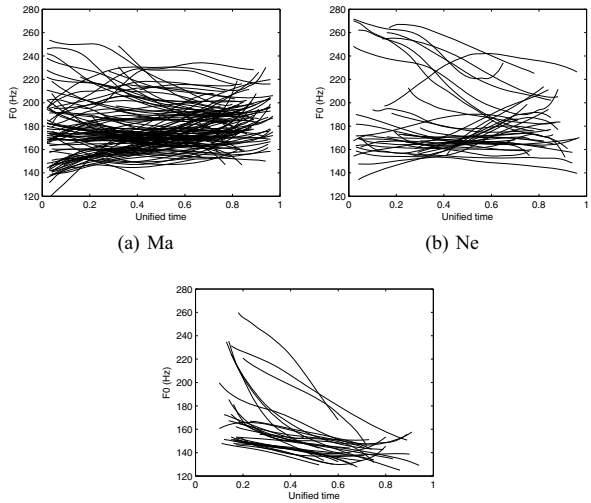
A. Context Feature Set Design

The conventional context feature set used for synthesizing Chinese sentences includes phonetic information (pinyin, initial, initial type, final, final type, tone, information about previous and next syllable, etc.) and prosody information (forward and backward position of the current prosodic unit in any higher prosodic units, prosodic unit length, etc.) [7].

For generating questions along with declarative sentences, the first feature that should be added to the context feature set is whether the sentence is a question, and what type of question it is. This feature is called sentence type in this experiment and it may take the values of declarative sentence, general question, rhetoric question, wh-question, alternative question and affirmative-negative question.

As the final boundary tone is very important in interrogative intonation, special feature is given to the last four syllables of interrogative sentences. This feature is called the distance from the end of interrogative sentence. Syllables farther than four syllables from the end of interrogative sentences and all syllables from declarative sentences are labeled with infinity. This feature combines prosody features with sentence type, and focuses on the final boundary tone.

The interrogative intonation may have many variations. There may be many factors that contribute to these variations. In this experiment, we took in only the effect of question words and final particles.



(a) Ma (b) Ne

(c) Ba

Fig. 3: Pitch contours of final particles “ma”, “ne” and “ba”.

The question word is usually the question focus, and may also affect the whole sentence intonation, including the boundary tone. Two kinds of features are added relating to the question word, which are the position of the question word and the syllable distance from the question word. The position is simply divided into sentence front, middle and end and is labeled to every syllable in a sentence. It is used to capture the effect of the question word on the whole sentence intonation. The syllable distance from a question word is from -3 to 3, while 0 is labeled to the question word itself. Farther syllables are labeled with infinity. This is used to capture the effect of the question word on the intonation as a question focus.

Final particles would also affect the interrogative intonation, and different final particle may behave differently. So the existence and kind of the final particle is label to the whole sentence.

The addition context features added to the original context feature set are: sentence type, distance from the end of interrogative sentence, question word position, distance from question word and final particle type.

B. Training Corpus

The training corpus used for generating the interrogative sentences is the TH-CoSS corpus [6]. The corpus contains 5,000 declarative sentences, along with 500 interrogative sentences. Phonetic and prosodic features are labeled in the corpus, along with the syllable boundaries. Other acoustic features are extracted from the training corpus automatically.

Among the interrogative sentences, several question types are included, e.g. general questions, wh-questions, rhetorical questions, alternative questions, affirmative-negative questions, etc. It also contains sentences with or without wh-words or final particles. In other words, it covers the full spectrum of different types of questions.

All the additional features are labeled manually. Question related features are given special values when they are not effective or in the declarative sentences.

C. Training and Synthesis

Five hundred declarative sentences are selected to be trained with the five hundred interrogative sentences to construct our synthesis system. One synthesis system using the original context feature set and one using the expanded context feature set are constructed.

Fig. 4 shows the pitch contour of the sentence “Zhe4ge4 wen4ti2 shi4bu5shi4 yi3jing1 jie3jue2le5?” (Is this problem already solved?) synthesized with the original context feature set and the expanded context feature set. The final part of the sentence synthesized with the expanded context feature set is much higher than the one synthesized with the original context feature set. The rise in pitch at the final part of the sentence is consistent with what we observed from the contrastive short sentence corpus. This shows that synthesis

system using the proposed expanded context feature set is able to generate interrogative sentences.

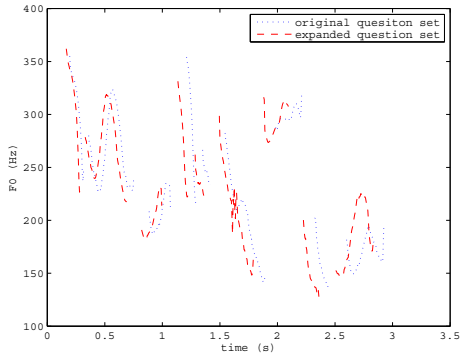


Fig. 4: Comparison of the pitch contours of the sentence "Zhe4ge4 wen4ti2 shi4bu5shi4 yi3jing1 jie3jue2le5?" (Is this problem already solved?) synthesized with the original context feature set and the expanded context feature set

To check how well the proposed context feature set is used in the training process, we also did some analysis on the decision trees built in the synthesis system. As the pitch contour is the main feature of intonation, we mainly looked at the decision trees created for logarithmic fundamental frequency (lf0). For lf0, in the ten decision trees built for each state in the HMM model, there are totally 6,901 leaves. At every split from the root to the leaf in the decision tree, one of the features is considered. The collection of the features considered from the root to a leaf node is the feature set used to separate this leaf from other leaves in the decision tree. Among the 6,901 leaves, 2,815 leaves (40.8%) took account of our additional features. Every kind of our additional features is used in the decision trees. This shows our additional features are able to capture the differences between declarative and interrogative sentences.

The most frequently used feature in the added features is the sentence type. 2,187 leaves (31.7%) used this information. This just shows that questions are using a different intonation as declarative sentences.

Wh-word related features come next, 434 leaves (6.3%) used this information. Wh-words are usually the question focus of the sentence. As no other question focuses are labeled besides wh-word, this means that we can put more effort on question focuses to improve the synthesis result.

The next is the feature about the distances from the end of interrogative sentences. 276 leaves (4.0%) used this information. This feature is overlapping with some of the existing prosody features, so they did not appear in more leaves though the final boundary tone is important in interrogative intonation.

The last is the feature about final particles. Only 152 leaves (2.2%) used this information. The different acoustic features of different final particles themselves would be captured by the phonetic features of the final particle. This result shows that the effect of the final particles on the whole sentence intonation is limited.

IV. Perception Test

A perception test is carried out on the interrogative sentences synthesized with the general context feature set and the expanded context feature set. Ten sentences are synthesized with both context feature sets, generating ten contrastive sentence pairs. These ten pairs of sentences are presented to 13 listeners at random order. The listeners are all college students with normal hearing ability. The listeners are asked which one of the two are more like a question, using a 5-point scale from -2 to 2.

The final average score comparing synthesis results using the expanded context feature set with the original one is 0.68. The positive score shows the expanded context feature set containing only textual features for questions has captured the difference between interrogative intonation and declarative intonation. This result is statistical significant with a significance level of 0.01.

V. Conclusion

In this paper, we analyzed interrogative corpora to get a general view of how Chinese question intonation is realized. We then proposed an expanded context feature set including only textual features to include question related context features to enable HMM-based speech synthesizer to generate interrogative intonation. Analysis on the train result shows that the additional features are effective in discriminating different intonations. Perception test has shown that using our proposed expanded context feature set, HMM-based speech synthesizer is able to generate proper interrogative intonation.

References

[1] M. Lin, "On production and perception of boundary tone in chinese intonation," in *the International Symposium on Tonal Aspects of Languages With Emphasis on Tone Languages*, Beijing, China, Mar. 2004, pp. 125–130.

[2] J. Yuan, C. Shih, and G. P. Kochanski, "Comparison of declarative and interrogative intonation in chinese," in *Speech Prosody 2002*, Aix-en-Provence, France, Apr. 2002.

[3] Y. Wang, "Phonetic representation of interrogative tone in chinese mandarin," in *The 8th Phonetics Conference of China and the International Symposium on Phonetic Frontiers*, Beijing, China, 2008.

[4] K. Tokuda, H. Zen, and A. Black, "An HMM-based speech synthesis system applied to english," in *Proceedings of 2002 IEEE Workshop on Speech Synthesis, 2002.*, Santa Monica, California, Sep. 2002, pp. 227–230.

[5] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, Nov. 2009.

[6] L. Cai, D. Cui, and R. Cai, "TH-CoSS, a Mandarin Speech Corpus for TTS," *Journal of Chinese Information Processing*, vol. 2007, no. 02, 2007.

[7] Q. Duan, S. Kang, Z. Shuang, Z. Wu, L. Cai, and Y. Qin, "Comparison of Syllable/Phone HMM based mandarin TTS," in *20th International Conference on Pattern Recognition*, Istanbul, Turkey, Aug. 2010.