Cross-lingual speech synthesis by phone concatenation

SCHM course project report

Yifan Li
2010011115
EE department of Tsinghua University
thueeliyifan@gmail.com

I. Introduction

People communicates with each other through speaking activities a lot in their daily life. As the technology keeps evolving, people are dreaming that maybe one day the computers could also talk like normal human beings. Speech synthesis, the artificial production of human speech, is the area where scientists are studying this important topic. Once this technology becomes true, it represents that people have made a big step toward creating a real human-like system. Nowadays, speech synthesizer has been utilized in a lot of areas like phone services, human and computer interactions and special music effects. However, most of the systems are designed targeted to the mainstream language like English and Spanish on the world. The fact is that the number of languages are estimated around 7,000. It’s time-consuming to create one synthesizer for each language. In this report, a method that uses English data and acoustic models to synthesize Chinese speech is proposed. Two Chinese sentences are tested through this method. The poor performance of phoneme recognizer on another language is the biggest problem. And the artificial adjustment shows the possibility of synthesizing of Chinese through English phonemes. However, the tone and missing of some special phonemes would weaken the final effect a lot.

In the next sections, I would talk about the basic principles of Text-to-speech systems, HMM based speech recognizer and concatenative synthesis. In system design and implementation section, the system block would be introduced with the exact implementation of each block. In results section, two sentences experiments are displayed and compared. In the discussion section, the analysis of the final results would be showed.

Ⅱ.Basic Principles and Methodologies

TEXT-TO-SPEECH SYSTEMS

Text-To-Speech system is the systems that transform the text into human speech. It could be also viewed as a speech coding system with an extremely high compression ratio together with a high degree of flexibility in choosing style, voice, rate, pitch range, and other playback effects.

The process of converting text into speech may seem quite simple at the first sight. However, we would quickly find the high complexity of this process after we have studied the activity when human are speaking texts aloud. First, the words need to be converted into speakable forms like phonemes. Second, the system needs to convey the information of the sentence and process it in a lot of efforts to sound natural. This is the most challenging part during the process. At last, the sound is generated. A more specific system architecture figure is displayed in figure 1.

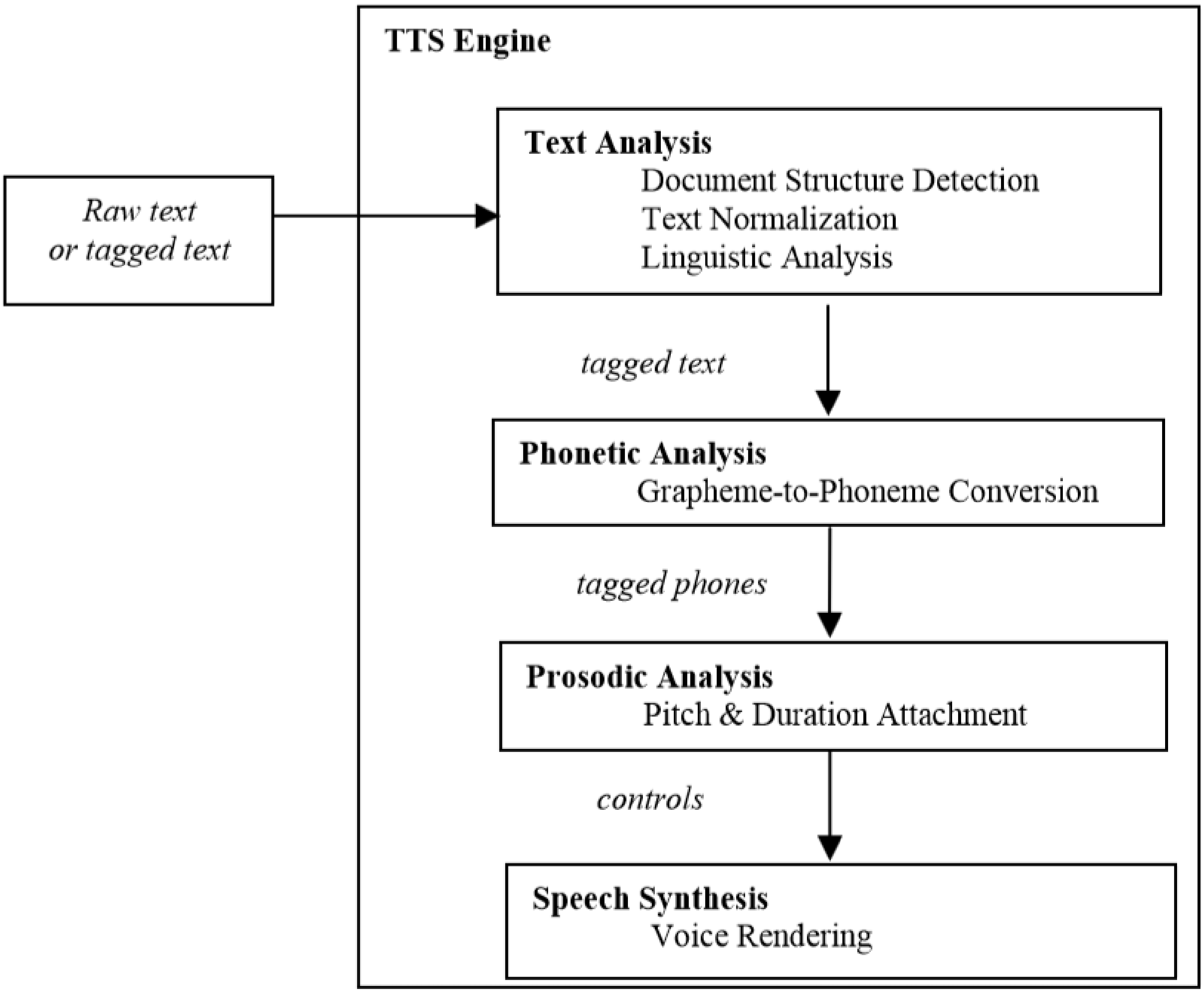

Figure 1 Text-to-Speech system architecture

In the TTS, the text analysis component is typically responsible for determining document structure, conversion of non-orthographic words, and parsing of language structure and meaning. The phonetic analysis component converts the orthographic symbols into phonemes. Then, the prosodic analysis component add the prosodic features to the phonemes according to text information. At last, the speech synthesis component generates the sound from a trained source-filter model or just concatenating the sound of units.

The TTS systems today haven't approached optimal quality in the Turing test. However, a large number of experimental and commercial systems have provided a lot of insight.

Phoneme recognition system with HMM

The phoneme recognition system is an important part of automatic speech recognition system. It is responsible for the front-end process of speech signal before the ASR system determines the text output. The phoneme recognition system with HMM utilizes the Hidden Markov model to represent the speech features extracted from the speech utterances. The model parameters are produced from the training process of the model and represents the vocal information of the phonemes. During the recognition process, the likelihood ratio of the test utterance would be computed and used as an important score to determine the type of the phoneme. Also , in some systems, the length of each phoneme are also produced from the recognition.

Concatenative synthesis

Concatenative synthesis is one of the most popular and important techniques in various speech synthesis methods including formant synthesis, articulatory synthesis and HMM-based synthesis. The concatenative method is based on the concatenation of segments of recorded speech. Generally, concatenative synthesis produces the most natural-sounding synthesized speech. However, differences between natural variations in speech and the nature of the automated techniques for segmenting the waveforms sometimes result in audible glithes in the output.

There are also different types of concatenative synthesis methods like unit selection synthesis, Diphone synthesis and domain-specific synthesis.

Ⅲ.System Design and Implementation

My cross-lingual synthesis system consists of three parts including a English phoneme recognizer trained by English data, a Chinese words to English phonemes dictionary and a concatenating component. A standard synthesis process of my system has been shown as follows:

Step 1, an English training corpus become the input of the phoneme recognizer, then the recognized phoneme sequences are produced.

Step 2, the speech data of each phoneme are sliced from the English corpus according to the recognition result.

Step 3, the speech data of Chinese words are recorded

Step 4, each Chinese word speech utterance is transformed to the phoneme recognizer to be converted into speakable phonemes. The Chinese words to English phonemes dictionary described before are produced here.

Step 5, the concatenating component connects the phonemes together and does some normalization and smoothing process.

The English phoneme recognizer is a mono-phone context-independent model which was trained using HMM Toolkit(HTK).It consists of 40 phone units. Each phone model has 3 states, and each state has 16 Gaussian mixture components. The acoustic model uses the MFCC feature.

The training corpus 40 English utterances separately by 20 female and 20 male. Each utterance is very short at the length of just 3 or 4 seconds.

The Chinese word speech utterance are recorded directly on the Lenovo Yoga laptop by the integrated microphone. The speech is stored in Microsoft .wav format with 8kHz sampling rate, 16 bits per sample and mono-channel.

The code is written mainly in Matlab 2013b.

IV. Results

The designed method didn't perform well in the producing of Chinese sentences "你好，早上好" and "东北菜，好吃". In fact, there is no any information could be transformed in the final result.

However, after some revising in the step 4 on the dictionary. The result becomes acceptable at least.

Here are the comparing result of the waveforms of the recording speech and synthesized speech in figure 2 and 3.

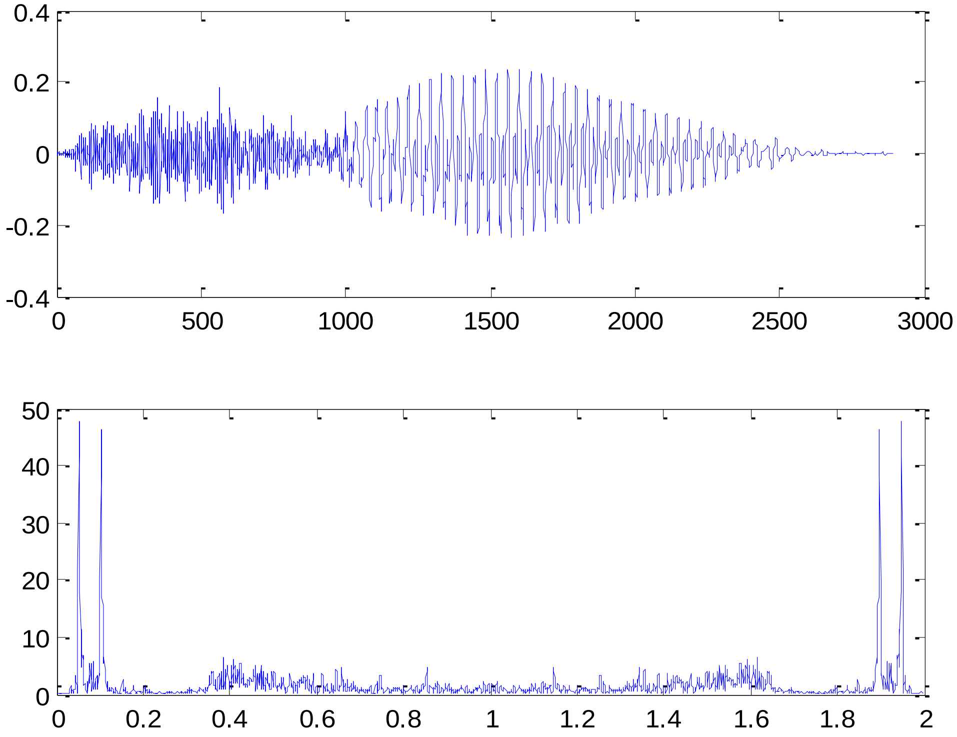

Figure 2 time waveform and spectral analysis of the recorded ‘吃’

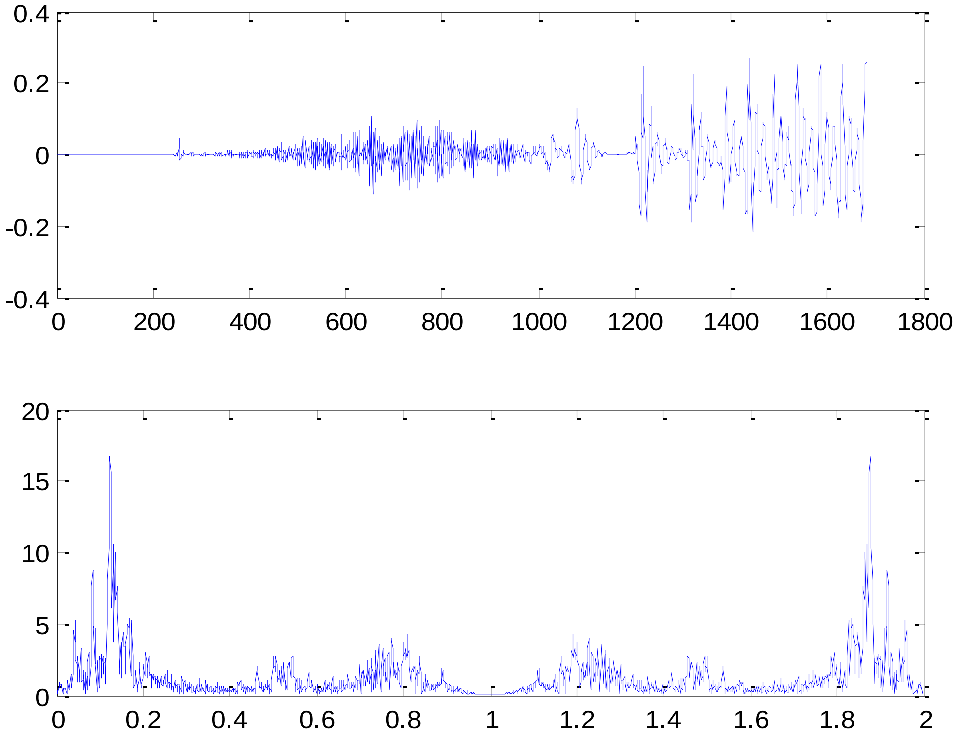

Figure 3 time waveform and spectral analysis of the synthesized ‘吃’

V. Discussion

The original system design didn’t perform well because of 4 reasons which are also the challenging part during the cross-lingual speech synthesis.

First, the English phoneme speech slides cut out from the training corpus isn’t one hundred percent accurate. There are some miss classification which would cause the failure of other parts.
The phoneme speech slides also varies a lot on the amplitude, length, environment noise and pitch. It is almost impossible to define a standard pronunciation since they are both right even though different considering the context they are from.
Second, the free alignment from Chinese word to English phonemes performs poorly. The consonants haven’ t been recognized at all. And the resulted phoneme sequence to each word also doesn’ t make sense from the view of artificially generating the phoneme sequence. The reason might be the tiny difference on the pronunciation of phonemes and even missing of some phonemes.

Third, the tone and pitch cannot be captured in this system. Especially, in Chinese, the tone are different even for the same word in different sentences. Forth, the length and amplitude of each phoneme is hard to choose automatically since we don' t have a standard version of phoneme pronunciation but a set of phoneme speeches.

However, after artificially adjustment in the building of the dictionary, the result becomes much better immediately. At least, this shows the possibility of using English data and model to synthesize Chinese speech. The four weaknesses could be overcome by using more complex word representing model including tone and pitch and other techniques. The benefits and flexibility of this method have showed us the hope to build different language synthesizer from one language synthesis system. Only small change on the system and different dictionary are needed in a new language system.

VI. Conclusion

This report shows the design and implementation of a cross lingual speech synthesizer by phone concatenating. The result of the system are also reported. The weaknesses and benefits of the system are discussed. To be concluded, this method has a long way to go to catch the performance of the state-of-art speech synthesis system. But the experiment on two Chinese sentences have shown the possibility of this method. The possible solutions to improve are also proposed in the discussion.

References

1. Huang X, Acero A, Hon H W. Spoken language processing[M]. New Jersey: Prentice Hall PTR, 2001.
2. Tokuda K, Nankaku Y, Toda T, et al. Speech synthesis based on hidden markov models[J]. 2013.
3. Wikipedia Article 'Speech synthesis' (http://en.wikipedia.org/wiki/Speech_synthesis#Formant_synthesis)
4. Baghdasaryan A G, Beex A A. Automatic Phoneme Recognition with Segmental Hidden Markov Models[C]//Signals, Systems and Computers (ASILOMAR), 2011 Conference Record of the Forty Fifth Asilomar Conference on. IEEE, 2011: 569-574.