論文の内容の要旨

論文題目　Advanced Control of Prosody in HMM-based Mandarin Speech Synthesis
（HMM 中国語音声合成における韻律制御の高度化）

氏　名　汪 淼淼

Speech synthesis is the artificial production of human speech. A computer system used for this purpose is called a speech synthesizer, and can be implemented in software or hardware. A text-to-speech (TTS) system converts normal language text into speech; other systems render symbolic linguistic representations like phonetic transcriptions into speech. Synthesized speech can be created by concatenating pieces of recorded speech that are stored in a database. Systems differ in the size of the stored speech units; a system that stores phones or diphones provides the largest output range, but may lack clarity.

Recently in speech synthesis community, attention has been attracted by HMM-based speech synthesis, in which short term spectra, fundamental frequency (F0) and duration are simultaneously modeled by the corresponding HMMs. It has compact and flexible representation of voice characteristics and has been successfully applied to TTS system in many different languages, e.g., Japanese, English and Mandarin. Compared with the large corpus, example the unit selection based speech synthesis, HMM-based synthesis is statistically oriented and model based. The speech generated by the HMMs is fairly smooth and exhibits no concatenation glitches occur in unit-selection synthesis. To change the segmental or supra-segmental quality of the generated speech, we can modify HMM parameters flexibly.

We are working hard to turn our ideas into reality and improve spoken-language technologies, enable human-computer voice interaction, and enrich human-to-human voice communications. We are current focus speech synthesis to enable computers to speak with a human-sounding voice, to respond and provide information, and to read; and spoken-document retrieval and processing to enrich communication between people. Hence our goals in building a computer system capable of speaking are to build a system that first of all clearly gets across the message and secondly does this using a human-like voice. Within the research community, these goals are referred to as intelligibility and naturalness.

Prosody is the rhythm, stress, and intonation of speech. It’s highly related with the intelligibility and the naturalness of synthetic speech. Prosodic features are suprasegmental. They are not confined to any one segment, but occur in some higher level of an utterance. However, the prosodic features, like F0 and duration trajectories, generated by HMM-based speech synthesis are often excessively smoothed and lack prosodic variance. Prosodic features of speech are known to play an important role

in the transmission of linguistic information in human conversation for any languages. But it's critical to Mandarin, or Standard Chinese.

Mandarin is a typical tonal language and each tone presents different meanings. F0 and duration are the two of the most important prosodic features contributing to the perceived naturalness of synthetic speech. The current state-of-art HMM-based TTS can produce highly intelligible output speech and deliver a decent segmental quality. However, its prosody, especially at the phrase or sentence level, tends to be bland.

In my study, in order to model the prosodic features within the standard HMM framework, firstly, we propose a new minimum v/u error approach to F0 trajectory synthesis for HMM-based TTS. The new approach is for producing more consistent and better v/u prediction in synthesis than the conventional baseline system. A prior knowledge of v/u label for each Mandarin phone is incorporated into v/u prediction and accumulated v/u probabilities are used to search for the optimal v/u switching point. Comparing with the baseline system, the new approach can significantly reduce v/u prediction errors in F0 generation and produce more pleasant synthesized voice. Then secondly, I developed a corpus-based method of synthesizing F0 contours in the framework of the generation process model (F0 model), which represents continues sentence F0 contours as a superposition of tone components on phrase components. By handing F0 contours in the F0 model framework, a clear relationship is obtainable between generated F0 contours and their background linguistic (and para-/non-linguistic) information, enabling "flexible" control of prosodic features. The F0 generation process model is used to re-estimate F0 values in the regions of pitch tracking errors, as well as in unvoiced regions. A prior knowledge of VU is imposed in each Mandarin phoneme and they are used for VU decision. Also it's necessary to predict segmental durations (including pauses) according to syntax information from the text. Firstly, this will help generation of prosody automatically at the backend in a system. Secondly, from the view of human speech production, underlying and surface syntax representation of the utterance is the step before phonetic representation in human speech production process. Syntax information might provide important cues for segmental duration prediction. We design a set of syntax features to improve Mandarin phoneme duration prediction. Instead of using manually extracted syntax information as previous researches do, we acquire these syntax features from an automatic Chinese syntax parser. Results show that even though the automatically extracted syntax information has limited precision; it could still improve Mandarin segmental duration prediction.