論文の内容の要旨

論文題目： Korean Corpus-based Text-to-Speech Synthesis System
（大容量のデータベース基盤の韓国語の音声合成器）

氏名： 金 相勳 (KIM, Sang-hun)

Speech synthesis is an emerging technology with many potential applications; especially fax reading, directory and reverse directory listing, information retrieval, proofreader, a talking character, dictation and navigational systems. To satisfy the user's needs, the speech synthesis has to create human-like voice as closely as possible. However, the conventional concatenative TTS systems based on prosody control still produces machine-like synthetic speech. It comes from the excessive signal processing for prosodic modifications. In general, the conventional methods have the limited available synthesis units (typically 1,000~2,000 demisyllables or diphones) in terms of acoustic and prosodic point of view. It is necessary to modify the prosody to represent the various prosodic phenomena with the defected speech database. The conventional synthesis methods seem to reach limits in the point of the synthetic speech quality. Therefore, it is time to shift the paradigm of speech synthesis.

This thesis describes a new Korean Text-to-Speech (TTS) system. To cope with the problems mentioned before, we have implemented a new Korean corpus-based TTS system using a large speech database without prosodic modification. The new TTS system has adopted the context sensitive units (i.e., triphone) as a synthesis unit. We have designed a new sentence set maximizing phonetic or prosodic coverage of Korean triphones. All the utterances were segmented through semi-automatic ways and constructed to synthesis database reflecting a synthesis unit cost of zero if two synthesis units were located consecutively in an utterance. This operation reduces the number of concatenating points that may occur due to concatenating mismatches. By doing so, we could create human voice-like synthetic speech without prosodic modification. From the informal listening test, we found that the proposed TTS system showed greater naturalness than did the baseline TTS system based on TD-PSOLA technique.

We have tried to detect the phrase break strength from the utterances. The various prosodic features were extracted. As a detection algorithm, the CART classification method was adopted. The detection performance of CART was 81.7% for four levels of phrases break strength except sentence final label '6'. To predict phrase break strength on texts, we have adopted an HMM-like part-of-speech sequence model. To reflect major prosodic variations, the phrase break strength was divided into four kinds of phrase types based on pause length. The performance of the prediction

model has shown 73.5% accuracy for four level phrase break strength prediction. We have also investigated phrasing style of several speakers. There is 88% agreement between speakers in phrasing style. After reflecting different phrasing types of speakers, the prediction accuracy was 80.8% for two level (Break or Non-Break) phrase break strength.

In the thesis, a new intonation stylization (i.e., LH stylization) was proposed. As a classification algorithm, we have adopted multi-layer perceptron one of neural network classifier. Complex intonation contours were stylized using step functions. We have extracted four kinds of prosodic features based on the stylized intonation contours. In the experiment of discriminating five major different boundary tones, the performance has shown 82.4% of boundary tone classification.

Finally, we have proposed a new pruning method called weighted vector quantization (WVQ) to eliminate useless instances from the synthesis database. As usual, a large-scale synthesis database for a unit selection based synthesis method retains redundant synthesis unit instances, which are useless to the synthetic speech quality. The WVQ reflects relative importance of each synthesis unit instance when clustering the similar instances using vector quantization (VQ) technique. The proposed method was compared with two conventional pruning methods through the objective and subjective evaluations of the synthetic speech quality: one to simply limit maximum number of instance, and the other based on normal VQ-based clustering. The proposed method showed the best performance under 50% reduction rates. Over 50% of reduction rates, the synthetic speech quality is not seriously but perceptibly degraded. The synthesis database can be efficiently reduced without serious degradation of the synthetic speech quality using the proposed method

Thanks to the corpus-based synthesis method, the TTS systems for information retrieval purposes seem to be successfully applied in real applications in spite of lack of user requirements. However, it is unilateral communication. In the near future, the TTS systems should be responded to the user's requirements interactively. In that case, the TTS systems should be able to output dialogue style speech. Furthermore, the TTS systems will express the emotion for the natural man-machine interface. Thus, this challenging topic will be hot issue in speech synthesis area.