Abstract

Recently with the advancements in the corpus-base speech synthesis technology, it becomes to be possible to obtain synthetic speech sounds with a high quality. Generally different context environment and F0 changes possibly induce some changes in the frequency characteristics. For this reason, it is necessary to collect huge speech data to cover all phoneme environment changes. This is not feasible in fact. Moreover, to synthesize the emotional speech and/or personalized voice, it requires database to have variety of speech with each style. To build a friendly user interface between human and familiar equipments it requires us to develop a compact speech synthesis method that can synthesize the desired voice quality and speech styles. For this purpose, a human mechanism based speech synthesis method is one of the solutions. This method has some definite advantages. One of them is that the database can be compressed greatly because the parameters used in the method vary slowly with time as human articulation. The other advantage is that this model is able to produce various synthetic speech sounds by manipulating the parameters instead of collecting the speech data. So far, a number of human mechanism based models have been proposed, unfortunately, few models can be used practically. One of the problems is that the physiological mechanism of speech production has not fully understood yet because of the difficulty in measuring speech organs during speech. The other problem is that the sound source and the vocal tract shape have been modeled independently, the interaction between them was not considered sufficiently. In addition, there are not enough the vocal tract data with high quality for constructing the model. Due to the development of the MRI technology in recent years, the measurement of the vocal tract shape under various conditions has become to be possible. Therefore, we attempt to develop a practical speech synthesis method based on human speech production mechanism by means of the advanced MRI technology.

In this paper, we proposed a novel speech synthesis method based on human speech production mechanism, where the human speech production is modeled as a combination of a sound source and a filter, the resonance property of the vocal tract. Based on the new observations using MRI technology, we refine the sound source model and the filter (vocal tract) part respectively, and develop a speech synthesis method by taking the interaction between the source and filter using a 2D physiological speech organs model. Furthermore, we challenge the bottleneck, the sound quality, towards practical use of such a method. To break the bottleneck and develop a practical system, we proposed a vocal-tract area function model by applying an accurate 3D measurement method on dynamic vocal tract shapes. The study was carried out in the following procedures. At first we investigated F0 control mechanism by analyzing the laryngeal complex and proposed a control method. Then, we proposed a physiological model of the speech organs and used it to confirm our observation of F0 control mechanism. Finally, we proposed a speech synthesis method which is suitable for practical application of text-to-speech synthesis system by using vocal-tract area function model.

To investigate F0 control mechanism, the MR images were measured during phonations with different F0 levels. It is found that a rotation of the cricoid cartilage was always associated with laryngeal descent during lowering F0. The function of the rotation is to shorten the vocal folds, and the mechanism was realized by vertical sliding motion of the posterior plate of the cricoid cartilage along the physiological curvature of the cervical vertebrae. This mechanism showed that the laryngeal descent and strap muscle activity are responsible for F0 lowering. Based on this mechanism, changes in F0 may cause a change of the tongue shape, and vice versa. To investigate this relation, a physiological model of speech production was designed to represent the interaction between F0 change and tongue shape change. The position of the speech organs was computed by driving their static equilibrium using muscle forces. Speech was synthesized based on the calculated vocal-tract shape and the length of vocal-fold. Simulation results of this model demonstrated that the proposed model can represent the observed F0 control mechanism and realize the interaction between F0 control and articulatory activity.

However, the proposed model cannot reproduce an acceptable individual speech. Poor sound quality is the bottleneck for this method to be a practical one. To solve the problem we extend the optimized parameters of vocal-tract from 2D to 3D. Accurate 3D vocal tract shapes were estimated from 3D MRI Movie data with reference to the recorded speech data. The vocal-tract shape was represented by a vocal-tract area function model. The results of a comparison between synthesized and recorded speech sounds showed that the proposed method can provide high quality speech sounds by using 3D MRI data. In the future, a compact text-to-speech system is planed to be developed on the proposed approach by taking the F0 adjustment mechanism into account.