Analysis of emotional speech based on ARX-LF model for emotional speech conversion

Yongwei Li (1210059)

School of Information Science,
Japan Advanced Institute of Science and Technology

February 12, 2014

Keywords: Speech analysis, ARX-LF model, emotional speech.

As one of the most convenient and important ways for communication among humans, speech expresses both linguistic information (i.e, semantic meaning) and non-linguistic information (emotion, age, gender, etc.), in which emotion is a very important element. Although recent speech synthesis systems can provide relatively good neutral speech to deliver the content, they are not able to synthesize emotion to present non-linguistic information. So some emotional speech conversion is widely studied based on neutral speech in recent years.

Many emotional speech conversion methods have been proposed. Most of them used statistics-based method, and Gausian Mixture Model (GMM) method. The first step of GMM method is to find out the training conversion functions by machine-learning, and the second step is to convert emotional speech from neutral speech by using training conversion function. A large database is needed in this method. It is well known that emotional state has different degree in our real-life and it can be described in activation and valence dimension. The database including all of emotional degrees is difficult to construct. The human speech system (speech perception and production) was not considered in GMM method. Thus, converting a lot of emotional degrees as human is diffcult.

Cahn first considered human perceptual ability and proposed two-layer perceptual model. However, human ability to perceive emotions from the

speech was not directly based on acustics. Thus, Huang and Akagi proposed a three-layer perceptual model to simulate human ability to perceive emotions from speech.

In order to convert a lot of emotional degrees like human as much as possible, we need come back to consider the analysis of origin of human speech production and connect with perceptual model. The main work of this research is to focus on speech production mechanism. The ARX-LF model is used to simulate speech production process. Since the period of joy and anger speech is very short, even small mistake may cause big error of analyzed results. However, the parameters of GCI and GOI are as input parameters of LF model and undervalued in current ARX-LF model-based methods.

In order to accurately estimate glottal source wave of emotional speech, the mean-based signal method is selected for estimating GCI parameter, and the GOI is estimated from the Hilbert envelop of LP residual. By searching the smallest values of mean square equation error (MSEE) between original speech waveform and re-synthesized speech waveform, an optimal set of estimation can be ontained. The results of our approach were compared with previous approach, and the compared results show that our approach is more suitable for estimating emotional speech.

The glottal source wave of emotional speech was analyzed by our approach, and the different parameters of LF model were obtained among different emotional speech. The results of this research show that the glottal source parameters have strong relationship with activation dimension and the analyzed results can be used for converting different degree of emotional speech in activation dimension.