Quality improvement of HMM-based synthesized speech based on decomposition of naturalness and intelligibility using non-negative matrix factorization

Anh-Tuan Dinh*, Masato Akagi*
*School of Information Science, Japan Advanced Institute of Science and Technology
E-mail: {tuan.dinh, akagi}@jaist.ac.jp

HMM-based synthesized voices are intelligible but not natural especially in limited data condition because of over-smoothing speech spectra. Improving naturalness is a critical problem of HMM-based speech synthesis. One solution for the problem is using voice conversion techniques to convert over-smoothed spectra to natural spectra [Jiao 2014]. Although conventional conversion techniques transform speech spectra to natural ones to improve naturalness, they cause unexpected distortions on acceptable intelligibility of synthesized speech. To improve naturalness without violating acceptable intelligibility, a novel asymmetric bilinear model (ABM) [Tenenbaum 2000] was employed to separate the naturalness and intelligibility of synthesized speech. Two problems in applying ABM are: finding efficient acoustic feature as input and factorizing two factors naturalness and intelligibility. Overcoming these problems, in the paper, an ABM was implemented on modulation spectrum domain of Mel-cepstral coefficient sequence to enhance fine structures of spectral parameter trajectories generated from HMMs. In addition, to avoid unrealistic subtraction among intelligibility factors and naturalness factors, subtractive combinations are avoided by applying non-negative constrain using non-negative matrix factorization (NMF).

Two CMU datasets (RMS and SLT) were used to train two HMM-based voices (HMM). For both speakers, 10 different samples are synthesized. The samples were improved in large data condition using proposed method with singular value decomposition (as SVD), proposed method with NMF (as NMF), global variance (GV) method [Toda 2007] and modulation spectrum filter (MS) [Takamachi 2015]. In limited data condition (with 5 natural sentences), GV cannot be used because there is not enough data to train GV model for each phoneme in specific context. These result in 400 pairs in large data condition test and 240 pairs in limited data condition test. Seven listeners (non-native but all fluent in English) participated in the test. For each pair, the subjects were asked which of the samples is more natural. Here, we define natural speech is actual human speech.

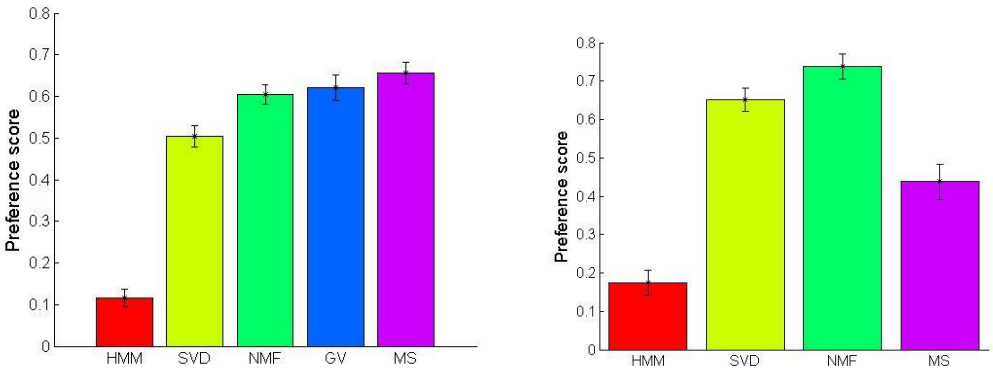

Figure 1: Preference scores with 95% confidence interval in large data and limited data, respectively

The proposed methods (SVD and NMF) can significantly improve the naturalness of HMM-based synthesized speech (HMM). It is competitive with GV and MS in large data condition. In limited data condition, proposed methods outperform MS in naturalness. The non-negative constraint also proved its efficiency in separating naturalness and intelligibility when NMF outperforms SVD.