CCIT Report #843 November 2013

Sequential Voice Conversion Using Grid-Based Approximation

Hadas Benisty, David Malah, and Koby Crammer

Abstract

The goal of voice conversion is to modify a source speaker's speech to sound as if spoken by a target speaker. Common conversion methods are based on Gaussian Mixture Modeling (GMM), which require exhaustive training (typically lasting hours), often leading to ill-conditioning, if the dataset used is too small. Additionally, the training process is based on a one-to-one match between the source and target vectors, requiring time alignment. We propose a new conversion method that is trained in seconds, using either small or large scale datasets (50-200 sentences). It requires a parallel dataset but without time alignment. The proposed Grid-Based (GB) method is based on sequential Bayesian tracking, by which the conversion process is expressed as a sequential estimation problem of tracking the target spectrum based on the observed source spectrum. The converted MFCC vectors are sequentially evaluated using a weighted sum of the target training set used as grid-points.

To improve the perceived quality of the synthesized signals, we use a post-processing block for enhancing the global variance. Objective and subjective evaluations show that the enhanced-GB method is comparable to classic GMM-based methods in terms of quality and comparable to their enhanced versions in terms of individuality.

1 Introduction

Voice conversion systems aim to modify the perceived identity of a source speaker saying a sentence, to that of a given target speaker. This kind of transformation is useful for personalization of Text-To-Speech (TTS) systems, voice restoration in case of vocal pathology, obtaining a false identity when answering the phone (for safety reasons, for example), and also for entertainment purposes such as online role-playing games.

The identity of a speaker is associated with the spectral envelope of the speech signal, and with its prosody attributes: pitch, duration, and energy. Most voice conversion methods aim to transform the spectral envelope of the source speaker to the spectral envelope of the target speaker. The pitch contour is commonly converted by a linear transformation based on the global mean and standard deviation values of the pitch frequency.