Intra-Gender Statistical Singing Voice Conversion With Direct Waveform Modification Using Log-Spectral Differential

Kazuhiro Kobayashi, Tomoki Toda, Satoshi Nakamura

PII: S0167-6393(17)30371-0
DOI: 10.1016/j.specom.2018.03.011
Reference: SPECOM 2551

To appear in: *Speech Communication*

Received date: 6 October 2017
Revised date: 12 February 2018
Accepted date: 31 March 2018

Please cite this article as: Kazuhiro Kobayashi, Tomoki Toda, Satoshi Nakamura, Intra-Gender Statistical Singing Voice Conversion With Direct Waveform Modification Using Log-Spectral Differential, *Speech Communication* (2018), doi: 10.1016/j.specom.2018.03.011

Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 00 (2018) 1–16

Speech Communi-cation

Intra-Gender Statistical Singing Voice Conversion With Direct Waveform Modification Using Log-Spectral Differential

Kazuhiro Kobayashi[a], Tomoki Toda[a], Satoshi Nakamura[b]

[a]*Information Technology Center, Nagoya University, Japan.*
[b]*Graduate School of Information Science, Nara Institute of Science and Technology (NAIST), Japan.*

Abstract

This paper presents a novel intra-gender statistical singing voice conversion (SVC) technique with direct waveform modification based on the log-spectrum differential (DIFFSVC) that can convert the voice timbre of a source singer into that of a target singer without vocoder-based waveform generation of the converted singing voice. SVC makes it possible to convert the singing voice characteristics of an arbitrary source singer into those of an arbitrary target singer by converting some of its acoustic features, such as F_0, aperiodicity, and spectral features based on a statistical conversion function. However, the sound quality of the converted singing voice is typically degraded compared with that of a natural singing voice, owing to various factors, such as analysis and modeling errors in the vocoding process and over-smoothing of the converted feature trajectory. To alleviate sound quality degradation, we propose a statistical conversion process that directly modifies the signal in the waveform domain by estimating the difference in the spectra of the source and target singers' singing voices. Additionally, we propose the following several techniques for the DIFFSVC method: 1) derivation of a differential Gaussian mixture model (DIFFGMM) from a conventional Gaussian mixture model (GMM) and 2) a parameter generation algorithm considering the global variance (GV). The experimental results demonstrate that the proposed DIFFSVC methods enable significant improvements in the sound quality of the converted singing voice, while preserving the conversion accuracy of the singer's identity compared with conventional SVC.

Keywords: statistical singing voice conversion, direct waveform modification, log-spectral differential, global variance, Gaussian mixture model

1. Introduction

A singing voice is one of the most expressive components in music. In addition to pitch, dynamics, and rhythm, the linguistic information of the lyrics can be used by singers to express a greater variety of expression than with other music instruments. Although singers can also expressively control their voice timbre to some degree, they usually have difficulty in changing it widely (e.g., changing their own voice timbre into that of another specific singer) owing to physical constraints in speech production. If singers could freely control their voice timbre beyond their physical constraints, it would open up entirely new ways for singers to express a greater variety of expression.

Singing synthesis systems [1, 2, 3, 4, 5] for generating an arbitrary singing voice have attracted growing interest in computer-based music technology. By entering notes and lyrics into a singing synthesis engine, users (e.g., composers) can easily produce a synthesized singing voice with a specific singer's voice characteristics different from those of the user. Techniques have been proposed for flexibly control the synthesized singing voice in accordance with the user's preferences by manually [6] or automatically [7, 8] adjusting the parameters of the singing synthesis system to create a more expressive synthesized singing voice. Although these technologies are effective for producing singing voices designed by users, it is essentially difficult to produce synthesized singing voices by controlling all of the singing voice components including the lyrics in real time.

Singing voice conversion (SVC), on the other hand, con-