Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 34 (2015) 308–322

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

HMM-based expressive singing voice synthesis with singing style control and robust pitch modeling[☆]

Takashi Nose[a,*], Misa Kanemoto[b], Tomoki Koriyama[b], Takao Kobayashi[b]

[a] *Graduate School of Engineering, Tohoku University, 6-6-05 Aramaki aza Aoba, Aoba-ku, Sendai 980-0011, Japan*
[b] *Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Yokohama 226-8502, Japan*

Received 23 May 2014; received in revised form 28 March 2015; accepted 7 April 2015
Available online 16 April 2015

Abstract

This paper proposes a singing style control technique based on multiple regression hidden semi-Markov models (MRHSMMs) for changing singing styles and their intensities appearing in synthetic singing voices. In the proposed technique, singing styles and their intensities are represented by low-dimensional vectors called style vectors and are modeled in accordance with the assumption that mean parameters of acoustic models are given as multiple regressions of the style vectors. In the synthesis process, we can weaken or emphasize the intensities of singing styles by setting a desired style vector. In addition, the idea of pitch adaptive training is extended to the case of the MRHSMM to improve the modeling accuracy of pitch associated with musical notes. A novel vibrato modeling technique is also presented to extract vibrato parameters from singing voices that sometimes have unclear vibrato expressions. Subjective evaluations show that we can intuitively control singing styles and their intensities while maintaining the naturalness of synthetic singing voices comparable to the conventional HSMM-based singing voice synthesis.

Keywords: HMM-based singing voice synthesis; Singing style control; Multiple-regression HSMM; Pitch adaptive training; Vibrato modeling

1. Introduction

Speech synthesis is the key technologies for human computer interaction (HCI) systems, and the interactive robot is one of the most typical and important applications to be realized in HCI systems. Recently, a humanoid robot named HRP-4C (Kaneko et al., 2009) was developed whose appearance is quite close to that of a human (Nakaoka et al., 2009). For such a state-of-the-art interactive robot, more advanced speech synthesis with rich para-linguistic and non-linguistic information, e.g., affections, emotions, speaking styles, and speaker characteristics, is indispensable. In addition, a function of synthesizing not only speech but also singing voice is desirable to achieve HCI systems that is capable of making the speech communication more diverse and rich like a human. This is because in our daily life there are a variety of music pieces including singing voices that are capable of relaxing or exciting us and some people communicate their feeling to others by singing. If an interactive robot has a function of singing voice synthesis with

[☆] This paper has been recommended for acceptance by Roger K. Moore.
[*] Corresponding author. Tel.: +81 22 795 7112.
E-mail addresses: tnose@m.tohoku.ac.jp (T. Nose), koriyama@ip.titech.ac.jp (T. Koriyama), takao.kobayashi@ip.titech.ac.jp (T. Kobayashi).

http://dx.doi.org/10.1016/j.csl.2015.04.001
0885-2308/© 2015 Elsevier Ltd. All rights reserved.

various singing styles, the application of the robot will expand not only to home entertainment but also to business showcase, exhibition, musical concert, and so on. Also in the education area, such an advanced and sophisticated interactive robot will give a good impact at the music class. These applications have a possibility of providing a new communication/interaction style between a human and a robot to our future life.

Singing voice synthesis is becoming an attractive application for speech synthesis in these days, and several products such as VOCALOID (Kenmochi and Ohshita, 2007) have become popular in the entertainment industry, especially in Japan. In the singing voice synthesis, users can easily create singing voices of certain singers or characters by inputting arbitrary musical notes (or MIDI codes) and lyrics, and this provides composers with an assistance method for adding original singing voices to their compositions. Recently, singing voice synthesizers have been utilized not only for hobby use but also for professional music production, a singing robot, karaoke, and live music, which shows the potential capability for entertainment and amusement applications (Tachibana et al., 2010; Kenmochi, 2012).

To develop a singing voice synthesis system, various approaches have been proposed (Cook, 1996; Rodet, 2002). The techniques based on speech production models, e.g., (Cook, 1993), and formant synthesis, e.g., (Sundberg, 2006), have an advantage that their model parameters have physical meanings and hence we are able to modify the synthetic singing voice by carefully controlling the parameters. However, the synthesis performance highly depends on the target voice, and the quality of output voice is not always satisfactory, which is the main drawback in singing voice synthesis. In terms of spectral reproducibility, concatenative synthesis, e.g, (Macon et al., 1997; Bonada et al., 2003; Kenmochi and Ohshita, 2007), outperforms the above techniques because recorded singing samples, such as diphones and sustained vowels, are directly used without physical modeling. A limitation of the concatenative synthesis is that singing samples, which are used as synthesis units, are recorded separately whereas we sing a song in a continuous manner beyond phonemes, syllables, and words. As a result, it is difficult to model the prosodic characteristics of singers and singing styles, and rule-based prosody generation is generally used. However, this heuristic approach is not always sufficient to synthesize singing voices with a wide variety of singing styles of various singers.

Singing voice synthesis based on hidden Markov models (HMMs) (Sako et al., 2004; Saino et al., 2006) is an alternative approach that enables simultaneous modeling of spectral and prosodic characteristics of singers and singing styles from a continuous singing voice corpus. The basic framework of the HMM-based singing voice synthesis is the same as that of HMM-based speech synthesis (Yoshimura et al., 1999). Although the baseline quality of the HMM-based singing voice synthesis has been steadily improved by introducing several techniques such as time-lag modeling (Saino et al., 2006), frame-based vibrato modeling (Oura et al., 2010), and pitch adaptive training (Oura et al., 2012), studies for diversifying singing voices are rather limited (Saino et al., 2010). By contrast, a variety of techniques, e.g., speaker adaptation (Tamura et al., 2001), style interpolation (Tachibana et al., 2005), and style control (Nose et al., 2007), have been proposed in HMM-based speech synthesis research area for adding or controlling various speaker and style characteristics (Yamagishi et al., 2009; Nose and Kobayashi, 2011). However, few studies have so far applied these techniques to singing voice synthesis (Sung et al., 2011).

In this study, we apply the style control technique of synthetic speech (Nose et al., 2007) to the HMM-based singing voice synthesis, which enables users to change the singing style expressivity in an intuitive and continuous manner.[1] In the proposed technique, multiple singing styles and their expressivity are represented by a low dimensional vector named a style vector and are simultaneously modeled using multiple-regression hidden semi-Markov models (MRHSMMs) (Niwase et al., 2005). The style vector is used as a explanatory variable of the MRHSMM where the mean parameter of each probability density function (pdf) is assumed to be given by a multiple regression of the style vector. In the model training, the parameters of MRHSMMs are estimated with training songs and the corresponding style vectors using maximum likelihood estimation with the EM algorithm. In the singing voice synthesis, we can control the singing style expressivity by changing the style vector.

To improve the modeling accuracy of pitch in the case of a limited amount of training data, we extend the model-space pitch adaptive training (Oura et al., 2012) into the feature-space one for the MRHSMM. In the proposed modeling, the observation features of static log fundamental frequency (F0) values are normalized using the pitch of the corresponding note in the parameter re-estimation process by the EM algorithm. By using pitch adaptive training, we can generate better F0 trajectories that closely follow the original pitch of given notes. In the singing voice synthesis, the vibrato modeling is also important for natural sounding synthetic singing voices (Maher and Beauchamp, 1990). However, the

[1] Part of this work was presented at INTERSPEECH 2013 (Nose et al., 2013).