Available online at www.sciencedirect.com

SciVerse ScienceDirect

Speech Communication 55 (2013) 347–357

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

An intuitive style control technique in HMM-based expressive speech synthesis using subjective style intensity and multiple-regression global variance model

Takashi Nose *, Takao Kobayashi

Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Yokohama 226-8502, Japan

Received 3 March 2012; received in revised form 10 August 2012; accepted 10 September 2012
Available online 24 September 2012

Abstract

To control intuitively the intensities of emotional expressions and speaking styles for synthetic speech, we introduce subjective style intensities and multiple-regression global variance (MRGV) models into hidden Markov model (HMM)-based expressive speech synthesis. A problem in the conventional parametric style modeling and style control techniques is that the intensities of styles appearing in synthetic speech strongly depend on the training data. To alleviate this problem, the proposed technique explicitly takes into account subjective style intensities perceived for respective training utterances using multiple-regression hidden semi-Markov models (MRHSMMs). As a result, synthetic speech becomes less sensitive to the variation of style expressivity existing in the training data. Another problem is that the synthetic speech generally suffers from the over-smoothing effect of model parameters in the model training, so the variance of the generated speech parameter trajectory becomes smaller than that of the natural speech. To alleviate this problem for the case of style control, we extend the conventional variance compensation method based on a GV model for a single-style speech to the case of multiple styles with variable style intensities by deriving the MRGV modeling. The objective and subjective experimental results show that these two techniques significantly enhance the intuitive style control of synthetic speech, which is essential for the speech synthesis system to communicate para-linguistic information correctly to the listeners.

Keywords: HMM-based expressive speech synthesis; Multiple-regression HSMM; Style control; Style intensity; Multiple-regression global variance model

1. Introduction

Expressive speech synthesis that is able to express various kinds of para-linguistic information, e.g., emotions, speaking styles, intentions, emphasis, and attitudes, is one of the key technologies to achieve more advanced and natural human-computer interaction (HCI). To synthesize natural-sounding expressive speech, a variety of techniques have been studied (Schröder, 2001; Erickson, 2005). A concatenative approach using unit selection with large-scale training corpora was successful in neutral-style speech synthesis and was also applied to the expressive speech (Iida et al., 2003; Pitrelli et al., 2006). However, in some expressions such as angry and joyful, the prosodic variation of speech is much larger than that in the neutral one, and it is not always possible to keep both the naturalness and expressivity of synthetic speech.

Recently, statistical parametric speech synthesis based on hidden Markov models (HMMs) has been widely studied (Zen et al., 2009) because of its flexibility and compactness. In the HMM-based approach, para-linguistic expressions globally appearing in speech such as emotional expressions and speaking styles, which we call simply *styles*, can be well modeled using the same training and generation algorithms as in the neutral-style speech without any modification (Yamagishi et al., 2003a). In addition, since the spectral and prosodic features are directly and simultaneously modeled and parametrized using HMMs

* Corresponding author. Tel.: +81 45 924 5030; fax: +81 45 924 5055.
E-mail addresses: takashi.nose@ip.titech.ac.jp (T. Nose), takao.kobayashi@ip.titech.ac.jp (T. Kobayashi).

0167-6393/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2012.09.003

(Yoshimura et al., 1999), it is easy to modify the parameters and add the new characteristics to the synthetic speech.

In the previous studies of expressive speech synthesis, most of the approaches have focused only on synthesizing speech of a certain categorized style where the expressivity of the synthetic speech cannot be changed in the synthesis stage. However, in our real-life communication, the intensity of style appearing in speech is not always constant and varies depending on the emotional state, the situation of the speaker, and so on (Cowie et al., 2003). One of the approaches to overcoming this problem is a style control technique (Nose et al., 2007), which we focus on in this paper. Style control is an idea for flexibly controlling style expressivity, i.e., the types and intensities of emotional expressions and/or speaking styles, appearing in synthetic speech. In the realization of the style control, the multiple-regression HMMs (Fujinaga et al., 2001; Miyanaga et al., 2004) or the multiple-regression hidden semi-Markov models (MRHSMMs) (Niwase et al., 2005) are used to simultaneously model and control multiple styles and their intensities. In the MRHSMM-based style control, the mean parameter in each state of a synthesis unit is represented by a multiple regression of a low-dimensional vector named a style vector. It has been shown that we can control subjective intensities of styles by changing the style vector in the speech parameter generation process.

Although the various styles can be well modeled and controlled using the MRHSMM-based style control framework, there are still two problems in the modeling and synthesis processes. The first is that the intensity of style expressivity perceived by listeners completely depends on that in the original speech used for the model training. This problem exists in both the HMM-based approach and most of the conventional expressive speech synthesis techniques. One of the most important factors in advanced HCI is how correctly the system can communicate paralinguistic information to the users as well as linguistic information. The conventional techniques ignored this essential issue and did not take the perceived style intensity into account. This is also true in the MRHSMM-based style control though style intensities of synthetic speech can be controlled. For instance, if the average intensity of a certain style in the training data is much lower than that users expect, intensity of the style in synthetic speech also becomes consistently lower than the users' demand. Consequently, it becomes difficult to communicate the intended para-linguistic information in the dialogue. In order to solve this problem, the perceived style intensities should be explicitly taken into account in the model training. The second is an over-smoothing problem that is well known in HMM-based speech synthesis. This problem is caused by the statistical averaging procedure through the model training. For the neutral-style speech, the variance compensation for spectral features global variance (GV) (Toda et al., 2007) has been shown to be effective for improving the subjective quality of the synthetic speech. Though the GV-based variance compensation can also be applied to a fundamental frequency (F0) trajectory, Toda et al. (2007) reported that there was no significant improvement in the neutral-style speech. However, the prosodic variation plays an important role in expressing some emotions, e.g., anger, fear, and happiness (Scherer and Oshinsky, 1977), and hence the variance compensation for excitation features should also be effective when synthesizing such stylized speech.

In this paper, we propose two novel techniques to alleviate the above problems and achieve more intuitive control of style intensities for MRHSMM-based expressive speech synthesis. First, we introduce subjective style intensities in the training of MRHSMMs. Specifically, the style intensities perceived for respective training utterances are quantified by listening tests and used as the style vectors in the model training. Though the approach is different, Tsuzuki et al. (2004) also introduced para-linguistic information related to emotional expressions in the HMM-based expressive speech synthesis. They quantified emotional intensities of training data on the basis of listening tests and used the voting results as the contexts of HMMs in the model training. Our approach is also similar to adaptive training techniques, e.g., speaker adaptive training (Anastasakos et al., 1996; Gales, 1998), cluster adaptive training (Gales, 2000), and context adaptive training (Yu et al., 2011). The difference from these approaches is that the adaptation in the proposed technique is based on the actual style intensities obtained by subjective listening tests whereas the conventional adaptive training techniques do not use such perceptual information and the matrices or weighting factors for adaptation are determined automatically in the model training. Introducing subjective style intensities into the model training brings us an additional advantage in style control: a style and its intensity can be modeled only using the speech data of the target style whereas the conventional MRHSMM-based style control technique (Nose et al., 2007) needs two or more styles for the model training. The second is multiple-regression GV (MRGV) modeling for speech parameter generation. When we attempt to model multiple styles' data using MRHSMMs, the GV can also vary depending on the kinds and intensities of those styles. To reflect the variation of the GV in the parameter generation, we formulate the GV model parameter control using a multiple regression of a style vector.

2. Modeling of the variation of subjective style intensities

In this section, we describe how to model the variation of the subjective intensities of styles appearing in expressive speech. To quantify the intensities of styles perceived for respective utterances, a subjective experiment is conducted and style intensity scores are determined from the rating results. The scores are then explicitly taken into account in the subsequent model training. MRHSMMs are used for modeling the variation of subjective style intensities, which is described in Section 2.3.