Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 58 (2014) 124–138

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Voice conversion based on Gaussian processes by coherent and asymmetric training with limited training data [☆]

Ning Xu [a,b,c,*], Yibing Tang [a], Jingyi Bao [d], Aiming Jiang [a,b], Xiaofeng Liu [a,b], Zhen Yang [e]

[a] *College of IoT Engineering, Hohai University, Changzhou, China*
[b] *Changzhou Key Laboratory of Robotics and Intelligent Technology, Hohai University, Changzhou, China*
[c] *Ministry of Education Key Lab of Broadband Wireless Communication and Sensor Network Technology, Nanjing University of Posts and Telecommunications, China*
[d] *School of Electronic Information and Electric Engineering, Changzhou Institute of Technology, Changzhou, China*
[e] *College of Telecommunication and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing, China*

Received 8 May 2013; received in revised form 28 October 2013; accepted 14 November 2013
Available online 26 November 2013

Abstract

Voice conversion (VC) is a technique aiming to mapping the individuality of a source speaker to that of a target speaker, wherein Gaussian mixture model (GMM) based methods are evidently prevalent. Despite their wide use, two major problems remains to be resolved, i.e., over-smoothing and over-fitting. The latter one arises naturally when the structure of model is too complicated given limited amount of training data.

Recently, a new voice conversion method based on Gaussian processes (GPs) was proposed, whose nonparametric nature ensures that the over-fitting problem can be alleviated significantly. Meanwhile, it is flexible to perform non-linear mapping under the framework of GPs by introducing sophisticated kernel functions. Thus this kind of method deserves to be explored thoroughly in this paper. To further improve the performance of the GP-based method, a strategy for mapping prosodic and spectral features coherently is adopted, making the best use of the intercorrelations embedded among both excitation and vocal tract features. Moreover, the accuracy in computing the kernel functions of GP can be improved by resorting to an asymmetric training strategy that allows the dimensionality of input vectors being reasonably higher than that of the output vectors without additional computational costs. Experiments have been conducted to confirm the effectiveness of the proposed method both objectively and subjectively, which have demonstrated that improvements can be obtained by GP-based method compared to the traditional GMM-based approach.

Keywords: Asymmetric training; Coherent training; Gaussian processes; Gaussian mixture model; Voice conversion

1. Introduction

Voice conversion, in a word, is a technique that aims to modify the speaker-dependent information of a source speech so as to match that of a target speech with message unaltered. There are many applications of VC, such as customizing voices for text-to-speech (TTS) systems (Stylianou, 2009), transforming somebody's voice so that it sounds like that of a well-known celebrity (Ye and Young, 2006), and improving the intelligibility of deficient voices uttered by a person with speech difficulty (Erro, 2008).

[☆] The work is supported in part by the Grant from the National Natural Science Foundation of China (11274092, 61271335), the Grant from the Fundamental Research Funds for the Central Universities (2011B11114, 2011B11314, 2012B07314, 2012B04014), the Grant from the National Natural Science Foundation for Young Scholars of China (61101158, 61201301, 31101643), the Grant from the Jiangsu Province Natural Science Foundation for Young Scholars of China (BK20130238), and the open research fund of Key Lab of Broadband Wireless Communication and Sensor Network Technology (Nanjing University of Posts and Telecommunications), Ministry of Education (NYKL201305).

[*] Corresponding author at: College of IoT Engineering, Hohai University, Changzhou, China.

E-mail addresses: xuningdlts@gmail.com (N. Xu), tangyb@hhuc.edu.cn (Y. Tang), baojy@czu.cn (J. Bao), jiangam@hhuc.edu.cn (A. Jiang), liuxf@hhuc.edu.cn (X. Liu), yangz@njupt.edu.cn (Z. Yang).

0167-6393/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2013.11.005

In general, the process of VC mainly consists of two stages, namely training and transformation, wherein the objective is to learn a mapping function from training observations (training stage) that can then be used to mapping arbitrary test features (including both prosodic and spectral features) of a source speech onto the acoustic space of a target speech (transformation stage). Various approaches have been proposed in the literature, e.g. the standard codebook mapping (Abe et al., 1988), the weighted codebook mapping (Arslan, 1999), the artificial neural network (ANN) mapping (Narendranath et al., 1995; Desai et al., 2010) as well as the Gaussian mixture model (GMM)-based mapping (Stylianou et al., 1998; Kain, 2001; Lee, 2007; Toda et al., 2008, 2007; Erro et al., 2010). The standard codebook mapping method, so-called vector quantization (VQ), is used extensively during the early days of VC, where a one-to-one correspondence between source and target spectral codebooks is derived in the training stage. In the transformation stage, the obtained relationship is used to transform the short time spectral envelope of a source speech into an estimated envelope that is close to the desired one. Since this transformation is achieved as a linear combination of the target codebook centroids of a limited set of vectors, it inevitably leads to discontinuities in the transformed speech and suffers from the problem of degraded speech quality. Then, an enhanced algorithm named weighted codebook mapping is proposed. In this case, the converted vector is calculated by weighting all of the target codebooks, wherein the weighting factor is obtained according to the contribution of line spectral frequencies (LSFs). This approach solves the discontinuity problem to some extent. Although ANN is capable of handling nonlinear mapping rules for spectral envelope, giving fairly good results, the performance of the VC system using ANN has been confirmed to be relatively inferior to that obtained by Gaussian mixture model (GMM). It is worth noting that GMM-based statistical mapping methods have made great contributions to VC by significantly improving the quality and similarity of the converted speech, compared to other alternatives. Thus, a massive of variants of GMM have been proposed and used extensively in most of the state-of-the-art VC systems (Lee, 2007; Toda et al., 2008; Erro et al., 2010).

On the other hand, it should be noted that GMM-based methods mainly suffer from the *over-smoothing* and *over-fitting* problems: (a) the *over-smoothing* phenomenon arises as the fact that the converted spectra are excessively smoothed compared to the natural ones, which may be attributed to the averaging nature of GMM. The nature of statistical modeling inevitably leads to the reduction of details of spectrum, thus causing the degradation of the synthesized speech. A lot of attentions have been paid to addressing this problem. For example, Chen et al. (2003) have found that most of the values in the cross covariance matrices of GMM are extremely small, which may lead to *over-smoothing* that makes converted speech sound muffled. So they proposed to design a mapping function based on the concept of maximum-a posteriori adaptation in order to alleviate this problem. Later on, the idea of perceptual post-filtering (Ye and Young, 2006) has been proposed to avoid the excessive broadening of the formants caused by the *over-smoothing* effect. Recently, Toda et al. (2007) have demonstrated that the variances of the converted spectra are less versatile than those of natural ones so that they introduced an enhanced version of GMM, which takes the global variances into consideration. Obviously, the *over-smoothing* problem of GMM is so well-established that a myriad of methods have been proposed intensively in the literature. (b) The problem of *over-fitting* is referred to the fact that a trained model gives very good results for the training data while being poor to predict new test data that are unseen. This problem arises from the fact that the structure of model is probably too complicated given that the amount of training data is limited. For example, suppose the dimensionality of feature parameter is 20, thus, a joint GMM model (Kain, 2001) with 256 mixtures trained by using 10 utterances may inevitably result in *over-fitting*. Compared to the vast proposals to the *over-smoothing* problem, relatively few attentions have been focused on the *over-fitting* problem of GMM until now. Among the limited amount of literatures, Helander et al. (2008) proposed to take the inter-relationship of LSFs into consideration, making GMM more reliable under the condition when small training data set is available. Later on, they proceeded by resorting to the combination of partial least squares (PLS) regression and GMM modeling, restricting the degrees of freedom in mapping functions by selecting a suitable number of components adaptively (Helander et al., 2012). In addition, variational Bayesian techniques were also used to obtain the estimates of the parameters of GMM in a full Bayes way, alleviating the *over-fitting* problem to a certain extent (Marume et al., 2007; Xu and Yang, 2010).

Recently, Pilkington et al. proposed to use Gaussian processes (GPs) experts to perform the task of spectral conversion, which is found to be insensitive to the *over-fitting* problem and can predict the target spectra accurately (Pilkington et al., 2011). In essence, GP provides a unified, principled and probabilistic approach for machine learning problems by resorting to Bayesian formalism (Rasmussen and Williams, 2006), which is intrinsically a nonparametric model giving advantage of making full use of the information involved in training set without strong artificial assumptions. It should be emphasized that the *over-fitting* problem can be largely suppressed due to the nonparametric nature of GP, which in fact allows relatively few degrees of freedom of model. Moreover, it is reported that GP can be interpreted as a Bayes-