Available online at www.sciencedirect.com

ScienceDirect

October 2017, 24(5): 60–67
www.sciencedirect.com/science/journal/10058885

The Journal of China Universities of Posts and Telecommunications

http://jcupt.bupt.edu.cn

Speaker conversion using kernel non-negative matrix factorization

Xu Qinyu[1,2], Lu Guanming [1,2](✉), Yan Jingjie [1,2], Li Haibo[1,2], Cheng Xiao [1,2]

1. College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
2. Jiangsu Province Key Laboratory on Image Processing and Image Communication, Nanjing 210003, China

Abstract

Voice conversion (VC) based on Gaussian mixture model (GMM) is the most classic and common method which converts the source spectrum to target spectrum. However this method is prone to over-fitting because of its frame-by-frame conversion. The VC with non-negative matrix factorization (NMF) is presented in this paper, which can keep spectrum from over-fitting by adjusting the size of basis vector (dictionary). In order to realize the non-linear mapping better, kernel NMF (KNMF) is adopted to achieve spectrum mapping. In addition, to increase the accuracy of conversion, KNMF combined with GMM (GKNMF) is also introduced into VC. In the end, KNMF, GKNMF, GMM, principal component regression (PCR), PCR combined with GMM (GPCR), partial least square regression (PLSR), NMF correlation-based frequency warping (NMF-CFW) and deep neural network (DNN) methods are compared with each other. The proposed GKNMF gets better performance in both objective evaluation and subjective evaluation.

Keywords VC, kernel, NMF, spectrum mapping

1 Introduction

With the maturity of speech signal processing techniques, VC has attracted more and more attention, which converts the voice of source speaker into the voice of target speaker while retaining linguistic information. VC has been widely used in emotion conversion, speech enhancement, speaking assistance and speaker conversion. For example, VC can be used to disguise the speaker's identity in secure communication systems. In the field of speech synthesis, VC can also be used to synthesize personalized speech. In most cases, a speech analysis and synthesis model based on source-filter is needed. In this paper, the spectrum mapping in speaker conversion is studied.

In the last decades, a variety of statistical approaches to VC have been proposed. Abe et al. [1] proposed a VC technique through vector quantization (VQ) and spectrum mapping. The basic idea of the technique was to make mapping codebooks to represent the correspondence between different speaker's codebooks. However, this method would not produce accurate results in scenarios when the speech spectrum has more nonlinear components. Valbret et al. [2] achieved better result than VQ through the linear multivariate regression (LMR) method, but the result was still not satisfactory. Stylianou et al. [3] proposed a GMM-based VC approach by representing the relation between two spectral envelopes corresponding to the source and target speaker. Kain et al. [4] proposed a GMM-based VC model which worked by analyzing the spectral parameters and mapping the spectrums of the source and target speaker. Toda et al. [5] introduced dynamic features and the global variance (GV) of the converted spectra over a time sequence. Toda et al. [6] also proposed eigen VC based on GMM (EV-GMM). Helander et al. [7] implemented VC using PLSR.

NMF is considered to be a popular method for speech enhancement [8] and monaural singing voice separation [9]. Takashima et al. [10] employed NMF method to realize an exemplar-based VC for noisy source signals. They also made a lot of improvements on NMF algorithm

Received date: 23-01-2017
Corresponding author: Lu Guanming, E-mail: lugm@njupt.edu.cn
DOI: 10.1016/S1005-8885(17)60234-6

[11–13]. Wu et al. [14] proposed an exemplar-based VC approach using joint nonnegative matrix factorization and improved the conversion quality.

In the field of VC, the existing methods still have some problems to be solved. In GMM-based VC, a segregate mapping function often leads to over-fitting, and GMMs may be over-fitted to the training data, especially when the training data set is small. In regression methods such as PCR and PLSR, the single linear transformation is insufficient to meet the requirement for VC.

In order to solve those problems, the KNMF is introduced to VC. KNMF has a larger range of applications and better results than NMF, which can overcome the limitations of NMF. Firstly, KNMF can uncover hidden features better from data with nonlinear mapping. Secondly, by using some specific kernel functions, KNMF can handle negative data. Thirdly, KNMF can be used to deal with the data which only knows the similarity. However, a single nonlinear transform is not effective for the data. To settle this problem, GMM is introduced to combine with KNMF, which is called GKNMF. GKNMF has the advantages of using kernel trick and GMM. GKNMF can overcome the over-fitting problem and realize the nonlinear mapping very well. In the experiments, GKNMF achieves better results for VC.

The paper is organized as follows: Sect. 2 describes the conventional GMM-based mapping with dynamic feature. Sect. 3 describes the NMF and KNMF based VC methods and also combined KNMF with GMM. Sect. 4 presents the objective and subjective evaluations. Finally, this paper is summarized in Sect. 5.

2 Conventional GMM-based mapping with dynamic feature

In this part, Trajectory-based conversion process [3] is used to map the speaker spectral envelopes. Source and target feature vectors are represented as $\boldsymbol{x}=[x_1,x_2,...,x_n]$ and $\boldsymbol{y}=[y_1,y_2,...,y_n]$ respectively, which are aligned by the dynamic time warping (DTW) algorithm. Considering the feature correlation between frames, $\boldsymbol{X}_t=[x_t,\Delta x_t,\Delta^2 x_t]$ and $\boldsymbol{Y}_t=[y_t,\Delta y_t,\Delta^2 y_t]$ are replaced for $\boldsymbol{x}=[x_1,x_2,...,x_n]$ and $\boldsymbol{y}=[y_1,y_2,...,y_n]$, which consists of static and dynamic features at frame *t*. The coefficient of delta and delta-delta window is set to [−0.5, 0, 0.5] and [1, −2, 1] respectively. The transformation vector $\boldsymbol{W}$ is specifically introduced in Toda's paper [5,6]

$$\left.\begin{aligned}\boldsymbol{X}&=\boldsymbol{W}\boldsymbol{x}\\ \boldsymbol{Y}&=\boldsymbol{W}\boldsymbol{y}\end{aligned}\right\} \tag{1}$$

Establishing a model for joint feature $\boldsymbol{Z}_i=[\boldsymbol{X}_i^{\mathrm{T}},\boldsymbol{Y}_i^{\mathrm{T}}]^{\mathrm{T}}$ by GMM [4]

$$p(\boldsymbol{Z})=\sum_{i=1}^{M}\alpha_i N(\boldsymbol{Z},\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i) \tag{2}$$

$$\left.\begin{aligned}\boldsymbol{\mu}_i&=\begin{bmatrix}\boldsymbol{\mu}_i^{x}\\ \boldsymbol{\mu}_i^{y}\end{bmatrix}\\ \boldsymbol{\Sigma}_i&=\begin{bmatrix}\boldsymbol{\Sigma}_i^{xx} & \boldsymbol{\Sigma}_i^{xy}\\ \boldsymbol{\Sigma}_i^{yx} & \boldsymbol{\Sigma}_i^{yy}\end{bmatrix}\end{aligned}\right\} \tag{3}$$

where α_i, $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ can be obtained by expectation maximization (EM) algorithm. Then by using the minimum mean-squared error (MMSE) algorithm method, the conversion function is obtained

$$\boldsymbol{Y}_t=\sum_{i=1}^{M}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W})^{-1}\boldsymbol{W}^{\mathrm{T}}p(i\mid\boldsymbol{X}_t)(\boldsymbol{A}_i\boldsymbol{X}_t+\boldsymbol{D}_i) \tag{4}$$

$$\left.\begin{aligned}\boldsymbol{A}_i&=\boldsymbol{\Sigma}_i^{yx}(\boldsymbol{\Sigma}_i^{xx})^{-1}\\ \boldsymbol{D}_i&=\boldsymbol{\mu}_i^{y}-\boldsymbol{A}_i\boldsymbol{\mu}_i^{x}\end{aligned}\right\} \tag{5}$$

$$P(i\mid\boldsymbol{X}_t)=\frac{\alpha_i N(\boldsymbol{X}_t,\boldsymbol{\mu}_i^{x},\boldsymbol{\Sigma}_i^{xx})}{\sum_{i=1}^{M}\alpha_i N(\boldsymbol{X}_t,\boldsymbol{\mu}_i^{x},\boldsymbol{\Sigma}_i^{xx})} \tag{6}$$

where *M* is the number of GMM centers. $P(i\mid\boldsymbol{X}_t)$ is the posterior probability of the *t*th frame $\boldsymbol{X}_t$. Experimental results show that the introduction of dynamic features is an effective way to solve the discontinuity problem.

3 Spectrum mapping using KNMF

3.1 Spectral conversion with NMF

In order to prevent the over-fitting problems of GMM, NMF is used to implement VC. NMF is not a statistical approach, so there is no over-fitting problem and it also can get a more natural voice.

The NMF decomposes a non-negative matrix into two non-negative factors

$$\boldsymbol{V}_{+}^{s}\approx\boldsymbol{B}_{+}^{s}\boldsymbol{H}_{+}^{s} \tag{7}$$

where superscript *s* and subscript + mean source speaker and non-negative respectively. $\boldsymbol{V}=[v_1,v_2,...,v_n]\in\mathbf{R}^{m\times n}$ is the input feature of the training data such as linear prediction coding (LPC), linear spectral pairs (LSP) or