Available online at www.sciencedirect.com
ScienceDirect
Computer Speech and Language xxx (2014) xxx–xxx
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

Interpretable parametric voice conversion functions based on Gaussian mixture models and constrained transformations

Daniel Erro [a,b,*], Agustin Alonso [a], Luis Serrano [a], Eva Navas [a], Inma Hernaez [a]

[a] *Aholab, University of the Basque Country, Bilbao, Spain*
[b] *Ikerbasque, Basque Foundation for Science, Bilbao, Spain*

Received 14 October 2013; received in revised form 7 February 2014; accepted 7 March 2014

Abstract

Voice conversion functions based on Gaussian mixture models and parametric speech signal representations are opaque in the sense that it is not straightforward to interpret the physical meaning of the conversion parameters. Following the line of recent works based on the frequency warping plus amplitude scaling paradigm, in this article we show that voice conversion functions can be designed according to physically meaningful constraints in such manner that they become highly informative. The resulting voice conversion method can be used to visualize the differences between source and target voices or styles in terms of formant location in frequency, spectral tilt and amplitude in a number of spectral bands.
© 2014 Elsevier Ltd. All rights reserved.

Keywords: Voice conversion; Gaussian mixture models; Frequency warping; Amplitude scaling; Spectral tilt

1. Introduction

Voice conversion (VC) (Moulines and Sagisaka, 1995) is the technology that allows transforming the voice characteristics of a speaker (the source speaker) into those of another speaker (the target speaker) without altering the linguistic message. The applications of VC include the personalization of artificial speaking devices, the transformation of voices in the movie, music and computer game industries, and the real-time repair of pathological voices.

Among all possible voice characteristics, the timbre, which is closely related to the short-time spectral envelope, has attracted most of the attention of researchers. During the training phase, given a number of speech recordings from the two involved speakers, VC systems extract their corresponding acoustic information and then learn a mapping function to transform the source speaker's acoustic space into that of the target speaker. During the conversion phase, this function is applied to transform new input utterances from the source speaker. Various types of VC techniques have been studied in the literature: vector quantization and mapping codebooks (Abe et al., 1988), more sophisticated solutions based on fuzzy vector quantization (Arslan, 1999), frequency warping transformations (Rentzos et al., 2004; Shuang et al., 2006; Suendermann and Ney, 2003; Valbret et al., 1992), artificial neural networks (Desai et al., 2010; Narendranath et al., 1995), hidden Markov models (Duxans et al., 2004; Lee et al., 2010; Zen et al., 2011), and Gaussian

* Corresponding author at: Aholab, University of the Basque Country, Bilbao, Spain. Tel.: +34 946017245.
E-mail address: derro@aholab.ehu.es (D. Erro).

http://dx.doi.org/10.1016/j.csl.2014.03.001
0885-2308/© 2014 Elsevier Ltd. All rights reserved.

Please cite this article in press as: Erro, D., et al., Interpretable parametric voice conversion functions based on Gaussian mixture models and constrained transformations. Comput. Speech Lang. (2014), http://dx.doi.org/10.1016/j.csl.2014.03.001

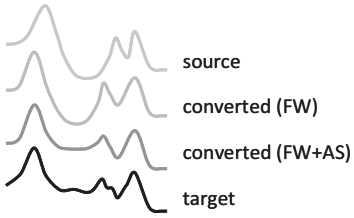

Fig. 1. Graphical explanation of FW + AS transformations applied to spectral envelopes.

mixture model (GMM) based VC (Benisty and Malah, 2011; Helander et al., 2010; Kain, 2001; Stylianou et al., 1998; Toda et al., 2007), which currently is the dominant technique.

Recently, the set of linear transforms characterizing the traditional GMM-based VC systems were replaced by a set of frequency warping (FW) plus amplitude scaling (AS) transforms (Erro et al., 2010; Godoy et al., 2012; Tamura et al., 2011; Toda et al., 2001) to improve the quality and naturalness of the converted speech. Unlike the former, FW + AS transformations have a clear physical interpretation (see Fig. 1). FW is a nonlinear operation that maps the frequency axis of the source speaker's spectrum into that of the target speaker. Since it does not remove any detail of the source spectrum but just moves it to a different location in frequency, FW preserves the quality of the converted speech well. However, the conversion accuracy achieved via FW is moderate because it does not modify the relative amplitude of meaningful parts of the spectrum. For this reason, FW is complemented with AS to compensate for the differences in the amplitude axis, typically by means of smooth corrective filters.

In the works referenced above, particular signal representations were required for the specific FW + AS methods to be applicable, whereas current trends in speech synthesis technologies are pushing research toward methods that can be applied to well known parametric representations of speech. That is why it was shown in Zorila et al. (2012) that GMM-based FW + AS methods can be applied to a simple cepstral representation of speech, overcoming the need of specifically designed vocoders. In Erro et al. (2012), the FW functions were constrained to be bilinear (BLFW), which led to a more elegant formulation of FW + AS in the cepstral domain with very few conversion parameters. The performance of BLFW + AS was found to be as good as that of the best existing GMM-based parametric VC methods (Erro et al., 2013b).

Following the line of BLFW + AS and in continuation of our preliminary work (Erro et al., 2013a), this paper goes one step beyond in making VC functions more understandable and controllable by users while reducing even more the number of involved conversion parameters. We suggest imposing constraints to the AS part of the VC function as it was done previously with the FW part. More specifically, we propose a new way of expressing the AS function as a combination of a spectral tilt related term and a set of smooth bandpass filters. We will show that the resulting VC functions are very informative in the sense that all their parameters can be interpreted from a physical point of view. Therefore, the method can be applied not only to synthesize high-quality converted voices but also to analyze the differences between the involved voices.

The remainder of the paper is structured as follows. Section 2 contains a brief description of the BLFW + AS VC method. In Section 3 we present the modified method and describe the corresponding automatic training procedures. The performance of this method is experimentally evaluated and discussed in Section 4. Section 5 shows how the proposed method can be used as an analysis and visualization tool. Finally, the conclusions of this work are summarized in Section 6.

2. Overview of BLFW + AS

In the cepstral domain, FW transformations are equivalent to multiplicative matrices (Pitz and Ney, 2005) and AS can be implemented by means of additive cepstral terms (Godoy et al., 2012). Given an input p-dimensional cepstral vector $\mathbf{x}$ and a GMM θ, the BLFW + AS operation proposed by Erro et al. (2013b) can be formulated mathematically as follows:

$$F(\mathbf{x}) = \mathbf{W}_{\alpha(\mathbf{x},\theta)}\mathbf{x} + \mathbf{s}(\mathbf{x}, \theta) \quad (1)$$

Please cite this article in press as: Erro, D., et al., Interpretable parametric voice conversion functions based on Gaussian mixture models and constrained transformations. Comput. Speech Lang. (2014), http://dx.doi.org/10.1016/j.csl.2014.03.001

Download English Version:

https://daneshyari.com/en/article/6951566

Download Persian Version:

https://daneshyari.com/article/6951566

Daneshyari.com