Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 48 (2006) 1262–1272
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Dynamic programming approach to voice transformation

Özgül Salor *,[1], Mübeccel Demirekler

Department of Electrical and Electronics Engineering, Middle East Technical University, 06531 Ankara, Turkey

Received 7 June 2005; received in revised form 14 April 2006; accepted 19 June 2006

Abstract

This paper presents a voice transformation algorithm which modifies the speech of a source speaker such that it is perceived as if spoken by a target speaker. A novel method which is based on dynamic programming approach is proposed. The designed system obtains speaker-specific codebooks of line spectral frequencies (LSFs) for both source and target speakers. Those codebooks are used to train a mapping histogram matrix, which is used for LSF transformation from one speaker to the other. The baseline system uses the maxima of the histogram matrix for LSF transformation. The shortcomings of this system, which are the limitations of the target LSF space and the spectral discontinuities due to independent mapping of subsequent frames, have been overcome by applying the dynamic programming approach. Dynamic programming approach tries to model the long-term behaviour of LSFs of the target speaker, while it is trying to preserve the relationship between the subsequent frames of the source LSFs, during transformation. Both objective and subjective evaluations have been conducted and it has been shown that dynamic programming approach improves the performance of the system in terms of both the speech quality and speaker similarity.

Keywords: Voice transformation; Speaker transformation; Codebook; Line spectral frequencies; Dynamic programming

1. Introduction

The aim of voice transformation (VT) is to modify the speech of a source speaker such that it is perceived as if spoken by a target speaker. A considerable amount of effort has been dedicated to the problem of voice transformation in the last two decades (Abe et al., 1988; Valbret et al., 1992; Childers, 1995; Mizuno and Abe, 1995; Lee et al., 1995; Stylianou et al., 1998; Arslan, 1999; Kain, 2001). There are various applications of a VT system. Using VT technology, new synthesis voices can be created by transforming the voice of the existing inventory to a new speaker's voice in a text-to-speech system. VT system would require a much smaller inventory than the original text-to-speech inventory, which saves time and disk space. Another application can be developing the voice of a speaking-impaired person, who can provide limited amount of speech data. A VT system could also be used as a preliminary step to speech recognition to reduce speaker variability.

In general, all VT systems have two modes: training and transformation. In the training mode, the

* Corresponding author. Tel.: +90 312 2101310; fax: +90 312 2101315.
E-mail address: ozgul.salor@bilten.metu.edu.tr (Ö. Salor).
[1] Present address: Institute of Space Technologies Research, TÜBİTAK, METU Campus, Ankara, Turkey.

0167-6393/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2006.06.003

system uses source and target speech inventory to estimate a transformation function that maps the acoustic space of the source speaker to that of the target speaker. Once the training is achieved, the system is ready to transform the source speaker's speech to the target speaker's speech. The acoustic space of the speakers can be represented by various acoustic features. Formant frequencies (Abe et al., 1988; Mizuno and Abe, 1995), LPC cepstrum coefficients (Lee et al., 1995; Stylianou et al., 1998), and line spectral frequencies (Arslan, 1999; Kain, 2001; Salor et al., 2003) have been used. The transformation function can be a continuous function applied to the features (Stylianou, 1999; Kain, 2001; Toda, 2003; Salor, 2005), or it can be a discrete mapping from the feature space of the source speaker to that of the target speaker (Abe et al., 1988; Arslan, 1999; Salor and Demirekler, 2004). The discrete mapping is in general a codebook mapping, in which a one-to-one correspondence between the spectral codebooks of the source speaker and the target speaker is developed. These methods usually face several problems such as degradation of the speech quality because the parameter space of the converted envelope is limited to a discrete set of envelopes. These methods may also result in high distortions between LPC spectrums of the neighboring frames due to independent transformation of the successive frames, which cause audible buzzy sounds or clicks.

In this work, we have aimed to obtain a voice transformation system inside the decoder part of a MELP speech coding algorithm. The idea is that the coded parameters could be used to produce the voice of another person at the end point of the coder. Therefore, we have focused on improving the quality of a codebook based voice transformation system. Here, we propose a dynamic programming approach to codebook based VT methods to overcome the problems of discontinuities and high distortions in speech. Dynamic programming approach considers the spectral distance between successive frames of the source speaker during transformation, while it is giving the chance to one of several target codewords to be selected at every frame instead of using a one-to-one mapping between the source and target speaker codewords. It has been observed that dynamic programming increases speech quality.

2. Algorithm description

This section provides a general description of the voice transformation algorithm. An overview block diagram of the VT system is given in Fig. 1. The speech model is based on the traditional *Linear Prediction Coding* (LPC) parametric model. The spectral characteristics are represented by line spectral frequencies (LSFs) and the spectral transformation from source to target is applied to the source speaker's LSFs. The reason for selecting LSFs is that these parameters are closely related to formant frequencies which carry speaker individualities (Arslan, 1999). LSFs have been used to represent the vocal tract parameters for VT throughout this work.

LPC residual is used as an approximation to the excitation signal during synthesis and average pitch of the excitation signal is modified such that the average pitch of the transformed sentence is the same with that of the target speaker. The algorithm will be described under two main sections: training and transformation.

2.1. Training

The output of the training part will be the histogram matrix, denoted by *Hist* and the target speaker's transition probability matrix denoted by T. Both of these matrices are square matrices of size $L \times L$, where L is the codebook size of both speakers. Details of obtaining these matrices are explained in detail in Sections 2.1.2 and 2.1.3 after

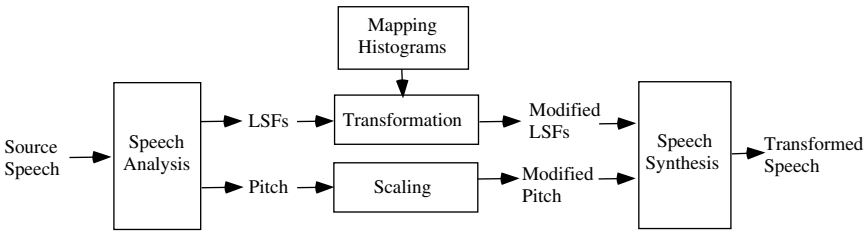

Fig. 1. Overview block diagram for the voice transformation system.