Towards a Robust Dynamic Frequency Warping Text-Independent Voice Conversion System

Tudor-Cătălin Zorilă[1], Daniel Erro[2], Yannis Stylianou[3,2], and Inma Hernáez[2]

[1]Telecommunication Department, Politehnica University of Bucharest, Romania
[2]Aholab Signal Processing Laboratory, University of the Basque Country, Bilbao, Spain
[3]Institute of Computer Science, FORTH, and Multimedia Informatics Lab, CSD, UoC, Greece
ztudorc@gmail.com,{derro,inma}@aholab.ehu.es,yannis@csd.uoc.gr

Abstract. In this work we investigate several issues related to the use of Dynamic Frequency Warping in the context of text-independent voice conversion. For this type of systems, given an average spectral representation of each acoustic/phonetic class, dynamic programming is applied to compute the best alignment path between the frequency axis of the source and target speakers. In order to increase the robustness of the system, we suggest estimating such average spectral information using the Multi-Frame Analysis framework, while we compare several different local slope constraints for the dynamic programming procedure. Objective measurements show that the suggested approach provides better results than a state-of-the-art histogram-based solution in transforming the source spectral envelope towards the target one for all the dynamic programming constraints we considered.

Keywords: Voice conversion, voice quality, dynamic frequency warping

1 Introduction

In a society more and more interested in the latest technological breakthroughs, speech technologies are fast evolving to keep up with the trends. Telephony, music, film and computer games industries are insatiable when it comes to new speech technologies applications. In this context, voice conversion (VC) has emerged as a powerful technology that allows to modify the voice produced by one speaker (the source speaker) in such manner that it is perceived as that of another speaker (the target speaker), thus providing artificially generated speech with a high degree of flexibility. To achieve this goal, the speech individuality of the target speaker has to be modeled and transferred to the source speaker. The speech individuality involves both segmental and supra-segmental speech features, but there is a general agreement that the spectral envelope plays the determinant role among them [1-7]. In this paper the spectral envelope conversion will be referred to as 'voice conversion'.

In a typical VC system, mapping functions between source and target spectral envelopes are trained using speech utterances recorded from the source and target speakers. In the case of parallel training corpora, the speakers record the same collection of utterances, thus facilitating the search for a correspondence between source

and target features. In practice, it is not always possible to obtain parallel corpora from the involved speakers (in cross-lingual VC, for instance). As result, nonparallel training methods have been proposed to make VC systems text-independent [3].

Although probabilistic mapping methods based on Gaussian mixture models (GMMs) have prevailed over other VC methods [1-4], they have some important limitations, the most remarkable one being known as over-smoothing [4]. This phenomenon is produced by the restricted capability of a statistical VC function to capture the source-target correspondence accurately. Instead, the converted features yielded by the system tend to be too close to the mean values of the corresponding acoustic class. Consequently, the generated speech lacks naturalness and variability. Another source of quality loss which is inherent to GMM-based VC systems is the use of low dimensional spectral envelope parameterizations provided by vocoders.

Dynamic Frequency Warping-based VC method (DFW) [5, 8] is a non-parametric approach that maps significant points of the frequency axis of the source speaker into that of the target speaker, thus offering the premises of a higher perceptual quality due to the fact that spectral details are preserved during the transformation stage. In exchange, the individuality of the target speaker is not well transferred by pure DFW-based conversion approaches because the relative amplitude of the different spectral bands is not modified. In recently suggested DFW-based VC systems a suitable amplitude correction step is introduced after DFW to cope with this problem [6-7]. Furthermore, the solution proposed in [7] allows text-independent VC by training frequency warping functions from single representatives of the content of the phonetic classes. The training procedure of such a DFW-based system consists of three steps: (i) the training data are segmented into a number of acoustic/phonetic classes; (ii) for each class, using only the data therein, representative spectral information (RSI) is calculated for both the source and the target speaker; (iii) the frequency warping path that produces minimum-distortion between the two RSI is determined. The main disadvantage of this system is that, even when phonetic labels are available (as it is assumed in this work), the accuracy of the DFW-based VC system is strongly conditioned by the way steps (ii) and (iii) are implemented.

A high quality VC system implies estimating robust class-dependent RSI, which is related to robust spectral envelope estimation. In this work, we assume that a spectral envelope represents the magnitude of the frequency response of the filter which models speaker's vocal tract. In this context, for the voiced speech, a good/robust class-dependent RSI should be fundamental frequency independent. This leads to the main problem to be solved, namely how to combine the contribution of multiple training frames into a single RSI. One possible solution is to average data sets [5], but this approach will lead to flat spectrum, so spectral details will be lost during transformation. Godoy et al. [7] suggested a better way to solve the problem by calculating class-dependent frequency histograms from meaningful (formant related) peaks of the magnitude spectra. These histograms are used as class-dependent RSI, where the k-th element is the probability that a formant peak is found in the k-th frequency bin. However, the main disadvantage of this approach is its sensitivity to the analysis method used to construct the histograms. The solution proposed by Godoy et al. will be denoted as Hist-VC from now on. In this paper we suggest a more

robust way of estimating the RSI of each class, i.e. Multi-Frame Analysis (MFA) [9]. Throughout the paper this method is denoted MFA-VC.

Given the source and target RSI that corresponds to a specific class, frequency alignment is carried out by means of an automatic dynamic programming technique. Due to vocal tract length and shape differences, aligning the speakers' RSI is not a trivial task (even manually!). Therefore, in this work several local slope constraints in the implementation of the DFW are investigated.

The paper is organized as follows. Section 2 provides a brief description for the methods used throughout the paper. Section 3 describes the experimental setup and presents the results. Finally, section 4 summarizes the conclusions and future work.

2 Methods Towards DFW-based VC

Figure 1 presents the training stage block diagram for the suggested MFA-VC system (solid line). A baseline approach for a DFW-based VC, Hist-VC (similar to the one described in [7]), is also shown (dashed line) for comparison purposes.

Segments of speech signals from the same phonetic/acoustic class (for both speakers) are divided into frames of length twice the pitch period and then the magnitude spectrum of these frames is computed using DFT. Next, a first Peak Picking is applied on the DFT magnitude spectra. The spectral peaks are pairs of magnitude and frequency bins, which correspond to the harmonics in voiced frames. These peaks are used as input for both MFA-VC and Hist-VC systems. The MFA technique is briefly presented in the following sub-section.

The block called "Spline + Peak Picking + Hist" is the module that generates the RSI for the Hist-VC system. It uses the peaks provided by the Peak Picking block and it computes spectral envelope by spline interpolating between these peaks (in voiced frames, this eliminates the pitch dependencies). Then, a second peak-picking is used to extract the local maxima from the estimated spectral envelope, thus the frequencies of these new peaks are an estimation of the formants' location. After analyzing all the training frames, for each phonetic class, histograms are built to model the distribution of the formants in frequency.

The DFW block aligns the RSI for both systems and generates an optimal frequency warping path per phonetic/acoustic class. The Amplitude Correction (AC) compensates for the remaining differences between the frequency-warped source and target spectral envelopes [7].

Using histograms as RSI is advantageous with respect to an average magnitude spectra solution because it removes the effect of spectral tilt and avoids spectral flattening. However, it is sensitive to the analysis parameters, such as: the analysis window type and length, the masking functions for selecting the representative peaks from the spectrum or the smoothing approach used to obtain the histograms.

As alternative, the suggested MFA-based VC uses the powerful Multi-frame Analysis core to extract robust RSI and offers in the same time the ability to generate robust AC functions (see Fig. 1). The dual gain from using MFA (in both estimation

robustness and system complexity reduction) can be boosted by the usage of a proper (physically motivated) DFW technique.

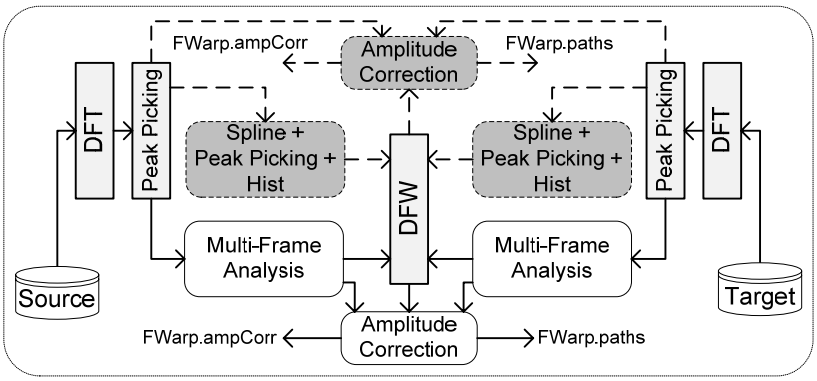

Fig. 1. Training stage block diagram for the suggested MFA-VC (solid line) and a Hist-VC systems (dashed line)

The following sections present the MFA spectral envelope estimation, the DFW and the AC internals.

2.1 Multi-frame Analysis Spectral Envelope Estimation

Due to quasi-periodic nature of voiced speech and due to changes of the fundamental frequency during speaking, the simple averaging of the magnitude spectrum from each acoustical class does not offer a robust estimator for the RSI of an acoustic/phonetic class. Assuming that such information exist (or it is valid) for a class, we may consider that a specific frame provides simply a sample of that information. However, the usage of a single frame will not provide a good estimation of the average spectral representation of an acoustic class. This is actually similar to the problem of computing the mean of a (multidimensional, in this case) stochastic variable using only one realization. Multi Frame Analysis offers a mechanism to estimate in a robust way this average spectral information (avoiding however any smoothing and whitening that a simple average operator will produce) [9].

MFA has been presented before in the context of speech synthesis but not in the context of voice conversion.

MFA uses cepstrum and Least Squares Estimation for all the frames from a phonetic category. Denoting $c[n]$ as the cepstrum for a discrete sequence $x[n]$, then the log magnitude spectral envelope can be approximated as

$$\ln|X(f)| = \sum_{n=-p}^{p} c[n]\cos(2\pi fn) \tag{1}$$

where p is the cepstrum's order.

Supposing M frames from a specific acoustic/phonetic class, these will be prior analyzed in terms of their log-amplitudes a_k^l and frequencies f_k^l (in this paper we consider the spectral peaks extracted by the Peak Picking block – see Fig.1), where the upper index denotes the l-th spectral peak and the lower index denotes the k-th speech frame. The cepstrum coefficients are estimated using least squares, considering the error ε to be minimized as:

$$\varepsilon = \sum_{k=1}^{M} \sum_{l=1}^{N_k} \frac{w(f_k^l)}{N_k} \left[a_k^l - d_k - \sum_{n=-p}^{p} c[n] \cos(2\pi n f_k^l / F_s) \right] \quad (2)$$

where $w(f)$ represents a weighting function used to put more emphasis on the lower part of spectrum, N_k is the number of harmonics in each frame, d_k is an offset factor used to correct the total power of each frame such that ε gets minimized and F_s is the sampling frequency.

In (2), the minimization of ε is a non-linear procedure since the error depends on $c[n]$ and d_k. Assuming an initial value for d_k, an iterative approach is adopted [9]. The iterations stop when the error is lower than a predefined threshold, providing the MFA-based p order cepstrum (1).

2.2 Dynamic Frequency Warping

To automate the process of finding the correspondence between RSI, dynamic programming technique is used [5, 7-8, 10-11]. The correspondence between RSI is a core component of a DFW-based VC system, as result we are considering different ways of implementing the dynamic programming in order to achieve the best physically meaningful warping paths. As result, in this paper three variations for the dynamic programming are considered, with different complexity and flexibility.

The first DFW technique accounted (denoted as DFW_1) has the maximum degree of freedom in searching for the optimal warping paths. Its local slope constraint is given by:

$$C_1[i,j] = \min \begin{Bmatrix} C_1[i-1,j-1] \\ C_1[i-1,j] \\ C_1[i,j-1] \end{Bmatrix} + S[i,j] \quad (3)$$

where C_1 is the cost matrix associated with the dynamic programming and S is the similarity matrix between the two RSI. For the similarity matrix the following distance is used (for all the methods):

$$S[i,j] = \left| RSI_1[i] - RSI_2[j] \right|,\ i = \overline{1, N_1},\ j = \overline{1, N_2} \quad (4)$$

where N_1 and N_2 are the lengths for RSI_1 and RSI_2, respectively. In this context, the goal of DFW is to find the minimum cost warping path $P = \{(i_1, j_1),(i_2, j_2),\ldots,(i_K, j_K)\}$ [8, 10-11].

The DFW_1 freedom for the warping path search is not always an advantage because it could lead to physically unsustainable warping functions. Therefore, quite complex global slope constraint policies are considered very often [10], with impact on the overall performance of the system. An intermediate step towards such a complex approach is by using a "built-in" global slope constraint (by the way the local slope is defined), with limited impact on the system's performance. Our extended experiments using different well known speech recognition local slope constraints scenarios [10-11] recommend the following form (referred to as DFW_2):

$$C_2[i, j] = \min\begin{Bmatrix} C_2[i-1, j-1] \\ C_2[i-2, j-1] \\ C_2[i, j-1] \end{Bmatrix} + S[i, j] \tag{5}$$

The third tested technique (denoted as DFW_3) is based on DFW_1, but introduces a higher degree of control to select the minimum cost frequency warping paths, following the work of Matsumoto and Wakita for non-uniform talker normalization [8]. By imposing a complex set of restrictions (boundary, continuity and nonlinearity conditions) the frequency warping functions are searched within a physically significant area. As result, with a higher computational price, physical meaningful warping paths can be obtained between RSI. The boundary conditions used in our implementation are 0.6 and respectively 1.4 for the slopes defining the search area. The values are chosen to cover a large dynamic for the ratios between speakers' min and max vocal tract lengths [8]. The advantage of this method is that speaker adaptive estimates for the vocal tract lengths could be easily integrated into the framework, therefore enhancing the DFW and promising higher VC quality.

2.3 Amplitude Correction

Amplitude correction is implemented as a corrective filter (CF) [12]. For the MFA-VC system the AC is implemented as a 12 order CF, computed from the difference between the frequency-warped source and target average spectral envelopes. The average spectral envelopes are generated from the MFA estimates. In the case of Hist-VC, an approach closer to the implementation mentioned in [7] is employed. The same 12 order CF is used, but in this case the average spectral envelopes (frequency warped for the source) are computed by means of discrete cepstrum [7, 13].

3 Evaluation and Results

In this experiment we are using CMU_ARCTIC speech database [14], consisting of phonetically balanced US English single speaker recordings at 16 kHz sampling frequency. For the purpose of spectral envelope conversion we have selected four speakers: two female speakers (*slt* and *clb*) and two male speakers (*bdl* and *rms*). All the voice conversion directions are considered: male to female (bdl to slt, denoted *bdl2slt*), female to male (clb to rms, denoted *clb2rms*), male to male (rms to bdl, denoted *rms2bdl*) and female to female (slt to clb, denoted *slt2clb*). A number of 100 parallel training sentences per speaker are selected, while a different set of 20 randomly selected sentences are considered for testing purposes.

During the training stage, the speech frames are extracted from segments defined by the labels provided by the ARCTIC's phonetic segmentation. The segmentation consists of 44 American English phonetic classes. The voiced speech segments are analyzed using a Hanning window with length twice the average pitch period, while for unvoiced segments a fixed 100 Hz pitch frequency is imposed. The training is done using only frames from the middle section of the segments to avoid boundary artifacts.

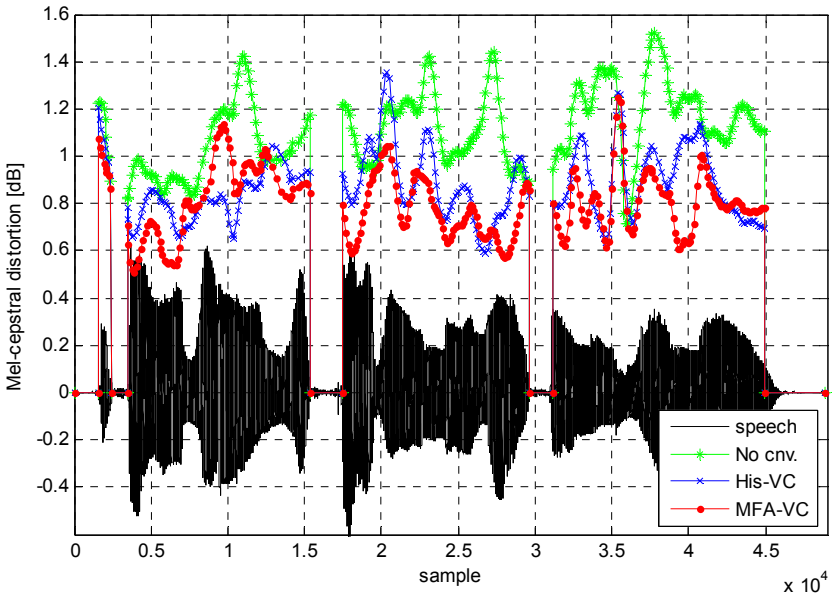

Fig. 2. The MCD scores per voiced frames for the "If I ever needed a fighter in my life I need one now" sentence, using DFW2 approach and the following methods for the *bdl2slt* VC direction: no conversion applied (the green with stars curve), Hist-VC (the blue with x-s curve) and MFA-VC (the red with dots curve)

The MFA-VC system is objectively compared against the Hist-VC for different speaker pairs and for different dynamic programming methods. The objective measure used is the average Mel-cepstral distortion (MCD) between the converted and the target spectral envelopes [4]:

$$MCD[dB] = \frac{10}{\ln 10}\sqrt{2\sum_{k=1}^{K}\left(mc_k^{tgt} - mc_k^{src\to tgt}\right)^2} \quad (6)$$

The average MCD is computed only for voiced frames.

In Fig.2 is presented an example of spectral envelope conversion between a male and a female speaker in terms of MCD over time, using MFA-VC and Hist-VC systems. For comparison purposes the "no conversion" MCD is also depicted. It is obvious that the voice conversion is effective because the MCD drops. The general trend shows that MFA-VC offers lower distortions than Hist-VC. Another observation is that the distortion contours present an oscillatory behavior, with minimum corresponding in general to the center of the phonetic segments. This is normal because in this implementation the transformation applied per acoustic class is the same for all the frames inside a segment, thus the error is higher during segment transitions. Inter-segment interpolation between warping paths and CFs will be considered in our future work.

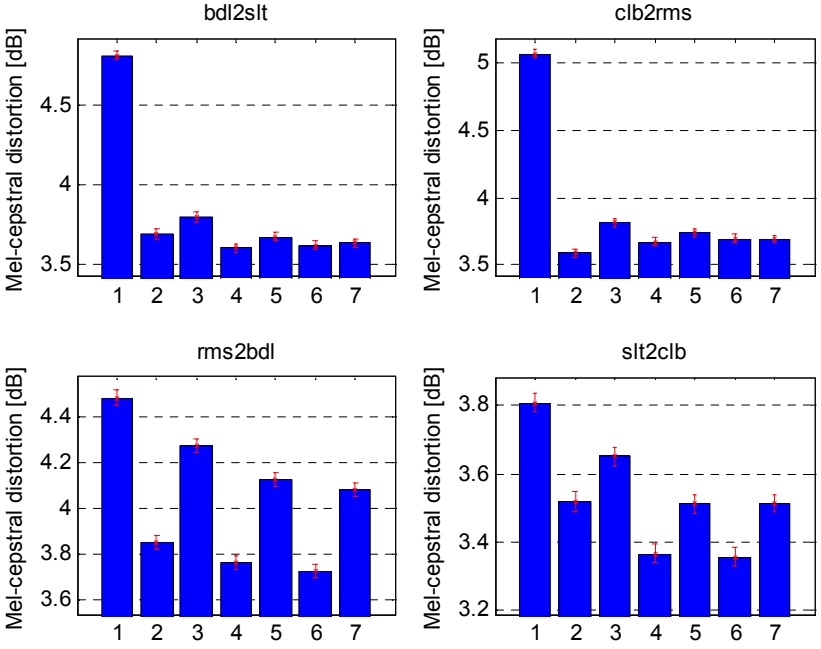

Fig. 3. Average MCD scores with 95% confidence intervals for all the VC directions and for different DFW approaches, in the case of MFA and histogram based VC systems, as follows: 1– no conversion, 2 - MFA-DFW1, 3 – Hist-DFW1, 4 – MFA-DFW2, 5 – Hist-DFW2, 6 – MFA-DFW3, 7 – Hist-DFW3

Table 1. Objective comparison between the MFA-VC and Hist-VC systems for the considered DFW techniques. The MCD values are averaged for all the VC directions

	No cnv.	MFA-DFW1	Hist-DFW1	MFA-DFW2	Hist-DFW2	MFA-DFW3	Hist-DFW3
Average MCD (dB)	4.54± 0.03	3.66 ± 0.03	3.88 ± 0.03	3.60 ± 0.03	3.76 ± 0.03	3.60 ± 0.03	3.73 ± 0.03

In Fig.3 the average MCD scores for all the VC directions are shown, with 95% confidence intervals values around ±0.03 dB. The lowest MCD for all the directions is obtained for the MFA-VC system. Furthermore, the superiority of MFA is maintained for all the considered DFW methods. As it is easy to see from Table 1, DFW_2 and DFW_3 offer the best performances. Taking into account the higher arithmetical complexity given by DFW_3, the optimal solution remains DFW_2. Even so, we believe that the power (in terms of robust and meaningful warping paths) given by DFW_3 is not completely unleashed. In our future work we will consider speaker adaptive parameters for the warping functions search area to increase the VC robustness and quality.

4 Conclusions and future work

Objective measures indicate that the suggested MFA-VC framework outperforms the Hist-VC opponent, both in terms of adopted warping strategy and conversion direction. By integrating prosody modifications as well, extended subjective tests between our framework and other state of the art VC systems are scheduled as a future work.

Acknowledgements. This work has been partially supported by the Romanian Ministry of Labor, Family and Social Protection (financial agreement POSDRU/88/1.5/S/61178), the Spanish Ministry of Science and Innovation (Buceador, TEC2009-14094-C04-02) and the Basque Government (Berbatek, IE09-262; ZURE_TTS, SPE11UN081). Y.S. holds a Visiting Professorship from UPV/EHU.

References

1. Y. Stylianou, O. Cappé, E. Moulines, “Continuous probabilistic transform for voice conversion”, IEEE Trans. Speech and Audio Process., vol. 6, pp. 131-142, 1998
2. A. Kain, “High resolution voice transformation”, Ph.D. thesis, Oregon Health & Science University, 2001
3. A. Mouchtaris, , J. Van der Spiegel, P. Mueller, "Nonparallel training for voice conversion based on a parameter adaptation approach," IEEE Transactions on Audio, Speech and Language Processing, vol. 14, pp. 952-963, 2006

4. T. Toda, A.W. Black, K. Tokuda, “Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory”, IEEE Trans. Audio, Speech, Lang. Process., vol. 15(8), pp. 2222-2235, 2007
5. H. Valbret, E. Moulines, J.P. Tubach, “Voice transformation using PSOLA technique”, Proc. ICASSP, pp. 145-148, 1992
6. D. Erro, A. Moreno, A. Bonafonte, “Voice conversion based on weighted frequency warping”, IEEE Trans. Audio, Speech, Lang. Process., vol. 18(5), pp. 922-931, 2010
7. E. Godoy, O. Rosec, T. Chonovel, "Voice Conversion Using Dynamic Frequency Warping With Amplitude Scaling, for Parallel or Nonparallel Corpora," IEEE Trans. Audio, Speech, Lang. Process., vol. 20, pp. 1313-1323, 2011
8. H. Matsumoto, H. Wakita, “Frequency warping for nonuniform talker normalization”, IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 1979
9. Y. Shiga, S. King, "Estimating the Spectral Envelope of Voiced Speech Using Multi-Frame Analysis," in Proc. EUROSPEECH 2003, 2003, pp. 1737-1740
10. H. Sakoe, S. Chiba, “Dynamic programming algorithm optimization for spoken word recognition”, IEEE Transactions on Acoustics, Speech, and Signal Processing, 1978, 26(1): p. 43-49
11. L. Rabiner, B.H. Juang, Fundamentals of speech recognition: Prentice Hall, 1993
12. Y. Stylianou, "Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification," PhD Thesis, Ecole Nationale Supérieure des Télécommunications, 1996
13. O. Cappé, J. Laroche, E. Moulines, “Regularized estimation of cepstrum envelope from discrete frequency points”, IEEE Workshop on Apps. Signal Process. to Audio & Acoustics, pp.213-216, 1995
14. CMU ARCTIC speech synthesis databases. Online: http://festvox.org/cmu_arctic/