Spectral Envelope Transformation using DFW and Amplitude Scaling for Voice Conversion with Parallel or Nonparallel Corpora

Elizabeth Godoy¹, Olivier Rosec¹, Thierry Chonavel²

¹Orange Labs, Lannion, France ²Telecom Bretagne, Brest, France
godoyec@gmail.com, olivier.rosec@orange-ftgroup.com, thierry.chonavel@telecom-bretagne.eu

Abstract

Dynamic Frequency Warping (DFW) offers an appealing alternative to GMM-based voice conversion, which suffers from “over-smoothing” that hinders speech quality. However, to adjust spectral power after DFW, previous work returns to GMM-transformation. This paper proposes a more effective DFW with amplitude scaling (DFWA) that functions on the acoustic class level and is independent of GMM-transformation. The amplitude scaling compares average target and warped source log amplitude spectra for each class. DFWA outperforms the GMM in terms of both speech quality and timbre conversion, as confirmed in objective and subjective testing. Moreover, DFWA performance is equivalent using parallel or nonparallel corpora.

Index Terms: voice conversion, DFW, spectral transformation

1. Introduction

From personalization of messaging systems to film dubbing to medical devices helping people with speech disabilities, Voice Conversion (VC) finds applications in numerous industries. The goal of VC is to modify the speech of a source speaker so that it sounds like that of a particular target speaker. To this end, the spectral speech envelope plays a crucial role in VC as it is critical both in capturing a speaker’s identity and in generating high quality synthesized speech.

A VC system has three stages. First, in the analysis stage, pertinent acoustic features (e.g. spectral envelopes) are extracted from the speech signals. Second, in the training or learning stage, a mapping between the source and target features is estimated. Third, in the transformation stage, the source parameters are transformed according to the learned mapping so that the speech synthesized with these modified parameters resembles that of the target speaker. In order to estimate the mapping between source and target parameters, most VC approaches rely on a time-alignment of source and target speech frames. It is this time-alignment that often imposes a significant constraint on VC systems that speech corpora be parallel (i.e. the source and target speakers utter the same phrases).

Currently, standard approaches to VC use Gaussian Mixture Models (GMMs) that explicitly seek to exploit the statistical correlation between time-aligned source and target frame parameters [1]. Unfortunately, this correlation is weak in the GMM [2]. As a result, transformed spectral envelopes do not deviate significantly from average representations of target spectra that are “overly-smooth” and lack spectral details [2], [3], [4]. Consequently, while more-or-less capturing the target timbre, the converted speech is not of a high quality.

Instead of explicitly depending on the source-target parameter correlation, Dynamic Frequency Warping (DFW) estimates a frequency warping function that is applied to a source spectral envelope so that it resembles that of the target [5]. In this case, spectral details are not attenuated as in the GMM, so the synthesized speech is of a higher quality. However, since spectral power is not adjusted explicitly in DFW, the speaker timbre is less successfully converted.

In order to generate high quality speech that also sounds like the target speech, an adjustment of spectral power after DFW is needed. To this end, previous work proposes a hybrid approach, combining a DFW- and GMM-transformed spectral envelope for each frame [4], [6]. Unfortunately, these combinations rely on arbitrary weighting [4] or smoothing [6] that diminish spectral details.

This work proposes a more effective way to adjust spectral power after DFW that is independent from GMM-transformation and a consequent reliance on time-alignment of speech frames; rather, the proposed Dynamic Frequency Warping with Amplitude Scaling (DFWA) compares source and target parameter statistics on the acoustic class level. First, acoustic classes are defined in a way that ensures a one-to-one correspondence between the source and target spaces. Then, the DFW is estimated by comparing source and target distributions of spectral peak occurrences in frequency. Next, for a given acoustic class, the amplitude scaling function is defined as the difference between the average target and frequency-warped source log amplitude spectra. In functioning on an acoustic class level, DFWA is able to both capture average trends of the target parameters and treat spectral details in a coherent way, without requiring arbitrary weighting or smoothing. As a consequence, the speech converted using DFWA is more natural sounding than the speech converted by standard GMM-based transformation while, at the same time, equally successful in capturing the target speaker identity. In addition to objective indications, these observations are confirmed with subjective evaluations for all gender conversion directions using data from multiple speakers.

Furthermore, an important consequence of eliminating a reliance on time-alignment of source and target speech is that DFWA does not require parallel corpora for VC. In fact, given contextual (e.g. phonetic) information to classify frames, DFWA performs equally as well in VC without parallel corpora, as shown in this work. This property removes a significant constraint for most VC applications, thus opening up spectral transformation with DFWA to a larger range of voice transformation applications.

The structure of this paper is as follows. Section 2 first describes details of the proposed method, DFW with amplitude scaling (DFWA). Section 3 then presents results from both objective and subjective evaluations, demonstrating that DFWA outperforms the GMM. Section 4 extends the application of DFWA to VC with nonparallel corpora. Finally, section 5 con-
cludes and discusses future implications in voice transformation technologies.

2. Dynamic Frequency Warping with Amplitude Scaling (DFWA)

2.1. Association of Acoustic Classes

Generally speaking, acoustic classes for the source and target speech can be generated and associated by clustering data. The choice of clustering approach depends on the available information in the speech data. Consequently, many different approaches are examined in VC works. For example, if only acoustic information is available, Vector Quantization (VQ) or a GMM can be used. In the case of source and target frames being aligned in time, this clustering can be carried out on the joint frames. Otherwise, a distance metric can be used to align separately generated source and target classes that have the closest means. Alternatively, when contextual information is available in the speech data, generation of source and target classes, as well as their association, can be carried out using symbolic information, such as phoneme labels.

Given a clustering approach that generates \( q \in \{1, 2, \ldots, Q\} \) acoustic classes aligned between the source and target spaces and that allocates each frame to a class, the DFW and amplitude scaling estimation are carried out strictly on the acoustic class level.

2.2. DFW Estimation

The frequency warping estimation in DFWA is inspired by a method discussed briefly in [7]. Specifically, the DFW estimation is based on aligning maxima in smoothed histograms of spectral peak occurrences in frequency, this alignment is carried out between a single source and target peak-occurrence histogram for each acoustic class \( q \). In DFWA, the spectral peaks used to generate these histograms are taken from peak-picking on the Discrete Fourier Transform (DFT) magnitude of a speech frame (as described in [8]), rather than the approach presented in [7] that is based on poles in a fixed-order Auto-Regressive (AR) spectral analysis. Considering the frames belonging to class \( q \), each spectral peak occurrence is accumulated into the histogram for the class. For each class histogram, the most informative features are the prominent maxima, which indicate the most probable spectral peak locations. Accordingly, these histogram maxima are the information used to define the intervals of the frequency warping function described in (1)-(3). In order to assure only the most relevant maxima are selected in this analysis, the histograms for each class are smoothed to reduce rapid fluctuations. Then, for each class \( q \), estimation of the DFW function aims to determine the “best” association between the source and target histogram maxima, i.e., the set of \( M_q \) pairings of these probable peak locations \( \left(f_{q,m}^{x}, f_{q,m}^{y}\right)\) that minimizes the sum of the absolute difference between the smoothed target and warped source peak occurrence histograms. For this purpose, a Dijkstra algorithm is used and constraints are introduced so that the resulting source and target histogram maxima are in a one-to-one correspondence [9]. From this correspondence, the DFW function for class \( q \) is given for frequency intervals \( f \in [f_{q,m}, f_{q,m+1}) \) by

\[
W_q(f) = B_{q,m}f + C_{q,m} \quad (1)
\]

where \( f_{q,m}^{x} = f_{q,m}^{y} = 0 \), \( f_{q,M_q}^{x} = f_{q,M_q+1}^{y} = f_{q} \), and

\[
B_{q,m} = \frac{f_{q,m+1}^{y} - f_{q,m}^{y}}{f_{q,m+1}^{x} - f_{q,m}^{x}} \quad (2)
\]

\[
C_{q,m} = f_{q,m}^{y} - B_{m}f_{q,m}^{x} \quad (3)
\]

There are several advantages to this type of DFW estimation. First, the DFW estimation does not rely on individual frame-by-frame comparisons of source and target spectral envelopes. Second, considering comparisons on the acoustic class level, spectral peak occurrence histograms are more informative than mean spectral envelopes, which are considered in [6]. In particular, one of these histograms can isolate distinct peaks that are near in frequency though appear often in different instances, while an averaged spectral envelope might combine them to form one larger or “blurred” peak. Third, by using histograms of spectral peak occurrences, the frequency warping estimation in DFWA avoids a spectral distortion criterion and is therefore not biased by spectral tilt. The spectral tilt is then accounted for solely in the amplitude scaling.

2.3. Amplitude Scaling

The goal of amplitude scaling after DFW is to adapt the warped source spectral envelopes so that they better resemble those of the target speaker. Consequently, the amplitude scaling function in DFWA is estimated using statistics observed from the target and warped source data.

Within a class \( q \), the amplitude scaling function adjusts the mean warped source log magnitude spectrum towards the mean target log magnitude spectrum. Thus, the amplitude scaling function for class \( q \), \( A_q(f) \), is such that:

\[
\log(A_q(f)) = \log(S_q^T(f)) - \log(S_q^W(f)) \quad (4)
\]

where \( \log(S_q^T(f)) \) and \( \log(S_q^W(f)) \) denote the average of the target and warped source log spectral envelopes, respectively, in class \( q \).

Given this amplitude scaling function, the transformation for frame \( n \) of source speech belonging to class \( q \) is given by

\[
S_q^{W,n}(f) = A_q(f)S_q^{T,n}(W_q^{-1}(f)) \quad (5)
\]

There are two important features distinguishing the amplitude scaling function in DFWA from other approaches adjusting spectral power of a warped spectral envelope. First, the separation of mean target and warped source log spectral envelopes in (4) shows that the comparison between source and target frames is carried out strictly on the acoustic class level, thus avoiding the need for individual frames to be aligned in time. Second, within each class \( q \), the average trends in the spectral envelopes for each speaker are exploited for each frequency, without blurring from arbitrary weighting or smoothing, thus maintaining integrity in the data.

3. Evaluations

3.1. Speech Data

The speech data used in all evaluations is from the CMU ARCTIC database for US English and is sampled at \( f_s = 16 \) kHz. The speakers include two males (bdl, rms) and two females (clb, slt), allowing for both intra and cross-gender conversion. The corpora are parallel, labeled and segmented. Given the availability of phonetic information, the acoustic classes in DFWA
are generated by phoneme; frame classification in DFWA is also carried out by phoneme. All speech data is analyzed pitch synchronously using the harmonic-plus-noise model [10] with cutoff frequency set to 8kHz. For voiced frames, the spectral envelope is estimated from the harmonic amplitudes. The envelopes considered in this work are parameterized by discrete cepstral coefficients of order 40 and objective metrics are calculated using these parameters.

In addition to DFWA, spectral transformation using a baseline GMM with the same number of acoustic classes is examined. The discrete cepstral coefficients are used as feature vectors (without the zeroth order coefficient) in the GMM learning, which is carried out on the joint vectors [11]. The marginal and cross-covariance matrices are diagonal and the transformation function is given in [11].

In all evaluations, the training sets contain 200 sentences and testing for objective results is carried out on 100 sentences separate from training. Voiced frames are used in learning and testing for objective evaluations. For the data sets, the source and target frames are aligned in time as follows: the three center (“stable”) frames of each source and target phone are automatically aligned; the remaining frames are aligned proportionally in time until the phone boundary. While DFWA does not rely on time-alignment of source and target speech frames, this alignment is necessary for learning the baseline GMM and for objectively evaluating transformation methods.

### 3.2. Objective Results

For the objective evaluations in this work, the different criteria described in [8] are considered. First, a standard MSE (normalized by the target energy) is used to indicate the average frame-to-frame accuracy in transformation. In addition to this error metric, the average ratio of the transformed data variance to the target data variance is a more global indicator of the ability of the transformation method to mimic realistic variations in the converted speech. This Variance Ratio (VR) is given by:

$$ VR = \frac{1}{Q} \sum_{q=1}^{Q} \frac{1}{P} \sum_{p=1}^{P} \left( \frac{\sigma_q(p)}{\sigma_q^*(p)} \right)^2 $$

where $\sigma_q(p)$ and $\sigma_q^*(p)$ are the standard deviations of the $p^{th}$ cepstral coefficient, calculated as the sample covariance of the transformed and target data, respectively.

Table 1 presents the average MSE (in dB) and VR for all gender-pair conversions between the 4 CMU ARCTIC speakers. As seen in Table 1, the GMM yields the lowest MSE, though there is little-to-no variance exhibited in the transformed data. These observations confirm the over-smoothing effect that hinders high-quality VC in GMM transformations. On the other hand, the DFWA-transformed envelopes exhibit variance comparable to that of natural speech, as is evident in comparing the VR for DFWA with that for the source. It is important to note that this variance in the DFWA transformed data results directly from variations in the warped source envelopes (5) and not from supplementary scaling factors or imposed adaptations in transformation, as in [2] and [3].

### 3.3. Subjective Test Results

For the converted speech in subjective tests, the harmonic amplitudes for voiced frames are re-sampled from the transformed spectral envelopes and the harmonic phases are taken from nearest-neighbor sampling of the source phases. The fundamental frequencies are modified as in [12], with the means and standard deviations estimated from voiced frames in the source and target phoneme. A separate conversion model is learned and applied to unvoiced frames, which are individually synthesized by passing white noise through an AR filter representing the spectral envelope. Finally, all speech synthesis is carried out with TD-PSOLA.

In the subjective evaluations, 17 English-speaking listeners were asked to judge the speech quality and perceived identity of speakers in 16 different trials following a MUSHRA methodology [13]. All possible source-target speaker pairs were included among the trials, covering both inter and intra-gender VC. Each trial contained the following stimuli: i) one sentence spoken by the source speaker (source), ii) the original source sentence pitch-modified using TD-PSOLA (TD-PSOLA), iii) the pitch-modified sentence converted according to the indicated transformation method (e.g. GMM or DFWA). The sentences used in the subjective evaluations were not part of the learning data set. It is important to note here that natural speech was included in the listening tests. While this inclusion might result in quality scores for converted speech that seem low, evaluations of converted speech quality in VC should not be limited (as in [6]) by a reference TTS synthesized voice that suffers from artifacts such as unnatural prosody, buzziness and concatenation discontinuities. Finally, considering the classification of frames by phoneme, there were no particular artifacts consistently heard or noted in transitioning between phones.

In the first test evaluating the speech quality, listeners were asked to rate the quality of the speech according to the following scale: 0-20 (bad), 20-40 (poor), 40-60 (fair), 60-80 (good), 80-100 (excellent). Responses were then normalized to a 5-point scale, as is shown in typical Mean Opinion Score (MOS) results. The results of the quality evaluation are shown in Fig. 1 for all trials (ALL) as well as for the particular gender-conversion directions: Male-to-Female (MF), Male-to-Male (MM), etc. For all subjective evaluations, results are indicated with their 95% confidence interval.

Fig. 1 shows that the converted speech using DFWA is consistently judged to be of a higher quality than that of the GMM. Furthermore, the subjective evaluations of speech quality in Fig. 1 more closely follow indications given by the VR in the objective results. This observation suggests that the MSE alone is not an efficient enough indicator of converted speech quality, as it does not reasonably penalize transformation that is fixed near the target class means. These results ultimately confirm that maintaining spectral details in transformation has a larger influence on converted speech quality than respecting only average trends.

In the second subjective test evaluating the similarity of the speaker to a reference speaker, listeners were first asked to listen to the same sentence used in the trial, spoken by the target speaker. For each sentence in the trial, the listener then considered the question: “How similar is the speaker to the reference speaker?” and then rated their response according to the following scale: 0-20 (completely different), 20-40 (fairly different), 40-60 (somewhat similar), 60-80 (fairly similar), 80-100 (very similar). Table 1: Objective Results

<table>
<thead>
<tr>
<th></th>
<th>GMM</th>
<th>DFWA</th>
<th>source</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSE (dB)</td>
<td>0.04</td>
<td>0.91</td>
<td>0.05</td>
</tr>
<tr>
<td>VR</td>
<td>0.90</td>
<td>0.93</td>
<td>1.09</td>
</tr>
</tbody>
</table>
 Listeners were asked to try to ignore differences in quality and prosody among the sentences. Again, responses were then normalized to a 5-point scale. Fig. 2 shows the results of the speaker identity evaluation. Examining Fig. 2, the transformation methods achieve similar success in converting speaker identity. In all cases comparing the source to the pitch-modified and then converted speech, there is an evident conversion in speaker timbre that is happening, going beyond simple pitch modifications. Considering the scores more closely, DFWA is equally (if not more) successful in converting speaker timbre when compared to the GMM.

4. Implications in VC with Nonparallel Corpora

The previous section shows that spectral envelope transformation with DFWA outperforms that of current baseline approaches in VC using the GMM. In addition to these results, there are significant, practical advantages to using DFWA. By functioning on the acoustic class level and eliminating GMM-based transformation using joint statistics, DFWA offers a more flexible framework for spectral envelope transformation that does not require individual source and target frames to be aligned in time. Consequently, parallel corpora are not required. In particular, DFWA performance is equivalent using parallel and nonparallel corpora in learning. These results are shown in Table 2, where sets of 200 identical (parallel) or disjoint (non-parallel) source and target phrases are used in DFWA learning. The same parallel sets of 100 phrases (separated from all the learning sets) are used in testing for both cases, where the DFWA performance is virtually identical. The two cases also proved to be equivalent in informal subjective evaluations conducted by the authors. Note that the subjective tests presented in the previous section also effectively represent the case of VC using nonparallel corpora, as DFWA transformation never depends on the source and target corpora being parallel.

5. Conclusion

This paper has presented a new alternative method to GMM-based spectral transformation in VC that uses DFW with amplitude scaling. Compared with the GMM, DFWA both i) achieves higher converted speech quality and ii) is equally (if not more) successful in transforming speaker identity. Moreover, DFWA offers a more flexible framework for spectral envelope transformation that eliminates the need for parallel corpora. In fact, DFWA performance is shown to be equivalent for VC using parallel or nonparallel corpora. The higher achievable converted speech quality combined with a less constraining transformation framework ultimately makes DFWA appealing to a wider range of voice transformation technologies, such as cross-language conversion or speaker adaptation.

6. References


