Correlation-based Frequency Warping for Voice Conversion

Xiaohai Tian[1,2]*, Zhizheng Wu*[3]*, S. W. Lee*[4]*, and Eng Siong Chng*[1,2]

[1]School of Computer Engineering, Nanyang Technological University, Singapore
[2]Joint NTU-UBC Research Centre of Excellence in Active Living for the Elderly, Nanyang Technological University, Singapore
[3]Centre for Speech Technology Research, University of Edinburgh, United Kingdom,
[4]Human Language Technology, Institute for Infocomm Research, Singapore
xhtian@ntu.edu.sg, zhizheng.wu@ed.ac.uk, swylee@i2r.a-star.edu.sg, aseschng@ntu.edu.sg

Abstract

Frequency warping (FW) based voice conversion aims to modify the frequency axis of source spectra towards that of the target. In previous works, the optimal warping function was calculated by minimizing the spectral distance of converted and target spectra without considering the spectral shape. Nevertheless, speaker timbre and identity greatly depend on vocal tract peaks and valleys of spectrum. In this paper, we propose a method to define the warping function by maximizing the correlation between the converted and target spectra. Different from the conventional warping methods, the correlation-based optimization is not determined by the magnitude of the spectra. Instead, both spectral peaks and valleys are considered in the optimization process, which also improves the performance of amplitude scaling. Experiments were conducted on VOICES database, and the results show that after amplitude scaling our proposed method reduced the mel-spectral distortion from 5.85 dB to 5.60 dB. The subjective listening tests also confirmed the effectiveness of the proposed method.

Index Terms: Speech synthesis, Voice conversion, Frequency warping, Correlation

1. Introduction

Voice conversion (VC) seeks to convert one speaker's speech (source) to sound like a target speaker's speech without changing the language content. A voice conversion system contains *training phase* and *conversion phase*. During *training phase*, a conversion function is estimated from parallel source and target feature vector sequences. In *conversion phase*, the conversion function is applied on features extracted from new input speech of source speaker, then the modified features are used to reconstruct the converted speech. As the spectral envelope plays a key role in both capturing speaker's timbre and determining the identity of speaker, the conversion function is also learned from this feature.

A number of statistical and frequency warping methods have been proposed to implement a robust conversion function. In the statistical parametric framework, the joint density Gaussian mixture model (JD-GMM) [1, 2], partial least squares regression [3], artificial neural network [4], and kernel partial least squares regression [5] method aims to implement linear and nonlinear conversion functions to map the source spectral features into the target space. The statistical parametric methods try to minimize the mean square error or maximize the joint probability between the source and target spectral features. This objective will cause the parametric model to capture the average property of the speech signals and lead to over-smoothing of converted speech [3, 6].

Alternative to statistical approaches, frequency warping (FW) approach aims to shift the source spectra frequency axes to match spectral shapes of the target. Since the frequency warping method does not remove the spectral details, it generally yields more natural converted speech as compared to the statistical parametric approaches. Generally, the frequency warping methods can be grouped into two categories. The first type of method, such as vocal tract length normalization (VTLN) [7, 8, 9] and bilinear frequency warping (BLFW) [10, 11], uses limited number of parameters to define the warping function. As these methods use a single warping factor to control the spectral shape, voice conversion systems benefit from the flexibility of those approaches. However, the dynamics in spectral shapes cannot be well modelled by those methods and a degraded conversion performance is generally observed [9].

The second type of FW method defines the warping function by a sequence of aligned frequency axis pairs. Dynamic frequency warping (DFW) technique was proposed in [12, 13, 14] to minimize the spectral distance between the source and target spectra. This method operates on the high-dimensional spectral feature directly and is able to achieve low spectral distortion. However, the conversion quality is moderate because the slopes of spectra are not considered. In [15, 16, 14], low-dimensional spectral features representing the formant positions, were used to train the FW functions. A combination of statistical method and FW method was proposed in [17]. Because the functions are trained by the corresponding formants, its performance depends on the accuracy of formant estimation.

In general, the methods of previous studies ignore the physical property constraint of the spectra, and only consider the spectral peak information. Inspired by the works of correlation optimized warping in [18, 19], we propose a correlation-based frequency warping method. Other than minimizing the spectral distance between the converted and target spectra, our proposed method maximizes the correlation of the spectra pairs. There are several advantages of the correlation-based frequency warping. First, with the criterion of maximizing correlation, the frequency warping function allocates not only the formants, but also the overall spectral shape, including spectral valleys and tilts. Second, although formant information is still needed in our method, by cutting the spectral envelope into several seg-

[1]This research is supported in part by Interactive and Digital Media Programme Office (IDMPO), National Research Foundation (NRF) hosted at Media Development Authority (MDA) of Singapore (Grant No.: MDA/IDM/2012/8/8-2 VOL 01).

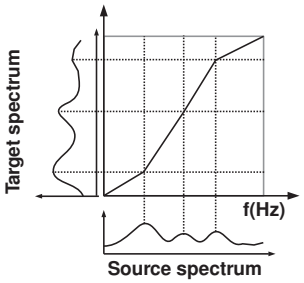

Figure 1: An example of piecewise linear FW function. The source and target spectral frequency axis were split into several segments. A linear warping function was determined by frequency pairs of source and target for each segment.

ments and introducing a varying window to relax the boundaries of each segment, the FW function will be more robust. The last but not least, our frequency warping function could obtain a higher correlation to the target, and the amplitude scaling (AS) technique [16] could works more efficiently.

2. Conventional frequency warping methods

The goal of frequency warping based voice conversion is using FW function, $W(f)$, to transform the frequency axis of source spectrum towards that of target. Here f denotes frequency.

In training phase, the $W(f)$ is learned from parallel frames of source and target. In conversion phase, this warping function will be employed on the spectrum of the input frame, $S(f)$, and the converted results can be expressed as: $S^*(f) = S(W^{-1}(f))$.

The piecewise linear based warping is widely used in previous works. Figure 1 shows the frequency warping function of two spectra. Assuming N frequency pairs $\{(f_1^s, f_1^t), \ldots, (f_N^s, f_N^t)\}$ denoting the boundaries for source and target frequency segments are given, the warping functions can be expressed as, $\{W_1(f), \ldots, W_N(f)\}$.

The sequence of frequency pairs is rephrasing the significant frequency points of spectra, which is aligned by a certain method. The accuracy of warping function greatly relies on the precision of finding these frequency pairs. In the following subsections, we will introduce two frequency warping methods as our baselines.

2.1. Dynamic frequency warping

Dynamic frequency warping (DFW), which is known as the most important method in frequency warping based voice conversion, was first introduced in [12].

The algorithm of DFW could be summarized by following steps. First, given a pair of spectra, the log-magnitude spectra of source and target can be treated as two frequency sequences and the distance matrix will be calculated. Second, by applying the dynamic time warping [20] on frequency sequences, the frequency pairs, $\{(f_1^s, f_1^t), \ldots, (f_N^s, f_N^t)\}$, will be aligned by minimizing the spectral distance. The dynamic spectral feature [21], which depicts the spectral variation, was also used for the distance matrix calculation. This could reduce the impact of spectral tilt efficiently. Then, the optimal warping function could be determined based on the aligned frequency pairs, with a constraint of $|(f_n^s - f_n^t)| < 1500$Hz to prevent very large warping factor.

Operating directly on high-dimensional spectral envelope, DFW leads to a large number of frequency segments, which could improve the flexibility of FW function and efficiently decrease the distortion between converted and target spectra. However, the frequency pairs are chosen by minimizing the spectral distance ignoring spectral formants and valleys. This will decrease the accuracy of FW function and the quality of converted speech.

2.2. Automatic mapping of formants

In this subsection, we briefly summarize the automatic mapping of formants (AMF) FW method, proposed in [15]. Similar to DFW, AMF also aims to obtain an optimal piecewise linear FW function by minimizing the spectral distance. The difference is AMF operates on formant frequency pairs. Since formant locations could be easily calculated by LPC coefficients [22, 23], which in turn could be represented by LSF, we use LSF features in our implementation to calculate the corresponding formant locations.

First, given two LSF vectors representing the source and target spectral envelope, the poles of LPC (formants) are computed from the LSF features and stored in ascending order. Second, these formant frequencies of source and target spectra are treated as two frequency sequences. Then, similar to DFW, the spectral distances of different warping paths are calculated based on the formant frequency sequences. The optimal FW function is defined by the warping path with the lowest spectral distance. Specifically, the derivative of the spectrum is included in the spectral distance calculation to decrease the influence of spectral tilt in warping path optimization. For details of the algorithm, please find in [24].

Comparing to DFW, AMF is able to obtain a more precise FW function by using formants, due to the high correlation between source and target spectra [15]. However, the AMF algorithm highly depends on the accuracy of formant estimation. The method operates on low-dimensional feature (LSF), which dismisses the spectral details during the FW function training, is another drawback of AMF.

3. Correlation-based frequency warping

To overcome the drawbacks caused by DFW and AMF, we propose to use correlation-based frequency warping (CFW). Different to DFW and AMF algorithms, the optimal warping path of CFW is achieved by maximizing the correlation. During the warping process, the spectra are split into segments. Then, the optimization problem could be solved by modifying the segment length to maximize the segment correlation.

Given two spectral envelopes, they are divided into equal number of segments. The boundaries of which is written as $\{(p_{s,1}^b, p_{s,1}^e), \ldots, (p_{s,N}^b, p_{s,N}^e)\}$ for source and $\{(p_{t,1}^b, p_{t,1}^e), \ldots, (p_{t,N}^b, p_{t,N}^e)\}$ for target. We note the ending point of current segment is the starting point of the following segment. The length of the n^{th} source and target segments were defined as $l_{s,n}$ and $l_{t,n}$, respectively.

During the warping process, the target segment boundaries are fixed, while the segments boundaries of source spectra will be varied. This will compensate the inaccuracy of boundary estimation efficiently. For each source segment, we always fix one segment point and the shift range of the other boundary point is defined by h. h is fixed for all the segments and should be smaller than segment length [18].

With segment boundaries of the source and target paired spectral envelopes, the correlation-based frequency warping can be performed through a backward step and a tracking step.

- *Backward step*: The backward step will be illustrated by an example. Figure 2 shows a pair of target and source spectral envelopes, with the segment number and the boundary shift set to 4 and 1 respectively. The process starts from the last (4^{th}) segment. Fixing the ending point of last segment of the source spectral envelope, the starting point varies from

$p_{s,4}-1$ to $p_{s,4}+1$. Then $p_{s,4}-1$, $p_{s,4}$ and $p_{s,4}+1$ are the possible ending points for the 3^{rd} segment. As the ending points increase, the starting point of 3^{rd} segment extends to a wider range, from $p_{s,3}-2$ to $p_{s,3}+2$, as shown in the figure. Since the position of the starting point is fixed, and the segment length is constraint from $l-1$ to $l+1$, where l is the predetermined length of the first segment, the ending point for the first segment can vary only from $p_{s,2}-1$ to $p_{s,2}+1$ instead a wider range. In other words, the possible starting and ending points of the $(n-1)^{th}$ segment will vary according to the segment boundaries of the $(n)^{th}$, $(n+1)^{th}$, ... segment. The constraint on segment length is always applied.

Each starting and ending point pair defines a possible segment. We perform linear interpolation to make the source-target segments have the same length, and s_n and t_n denote the $(n)^{th}$ segment of source and target. Correlation coefficient between s_n and t_n is calculated by $\gamma(s_n, t_n) = \mathrm{cov}(s_n, t_n)/\sqrt{\mathrm{var}(s_n)\cdot\mathrm{var}(t_n)}$, where $\mathrm{cov}(s_n, t_n)$ denotes the covariance between s_n and t_n; var denotes the variance.

In segment warping step, the cumulative correlation for each path will be saved at the starting point. Since the starting points of different paths may overlap, we only store the maximum correlation. The procedure continues until is finished all the correlation calculation of possible start points of segments.

- *Tracking step*: When all the correlation of segments are calculated, the segment wise optimal solution are defined. The path with maximum cumulative correlation is selected as the optimal solution. Simultaneously, the segment boundary points are also determined by the path. Note that, the first point and the last point of each source spectrum are fixed. Thus the invalid paths will be removed.

Obviously, the accuracy of CFW is affected by the segment boundaries and shift. In this work, the LSF features are used to define segments boundaries of source and target spectra. It is useful to control the number of formants and valleys in each segment. The shift is defined by the minimum length of the segments, which will compensate the imprecision problem of LSFs and lead to a FW function with higher accuracy.

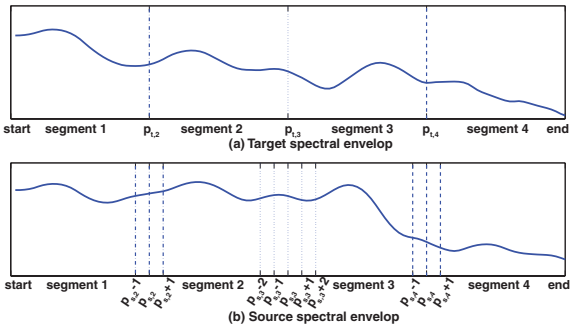

Figure 2: An example of the procedure of correlation-base frequency warping. Figure (a) is the target spectral envelope and predetermined boundaries. Figure (b) depicts the source spectral envelope and the range of segment boundaries.

4. Evaluation

4.1. Acoustic database

The VOICES database [25] was used in our experiments. Four speakers were selected from this database: two male speakers, *jal* and *jcs*, and two female speakers, *leb* and *sll*. The evaluations were conducted on intra-gender, *jal* to *jcs*, and *leb* to *sll*, and inter-gender, *jal* to *sll* and *leb* to *jcs*. For each pair, 20 parallel sentences were selected as training set, and another 20 sentences which were not included in the training set were used for test set.

The speech signals were downsampled to 16 kHz. STRAIGHT [26] was used as the vocoder of our experiments. An 513-dimensional spectral envelope was extracted, then it was represented by LSFs and mel-generalized coefficient (MGC) with the dimension of 14 and 24 respectively. The MGCs were used for source and target frame alignment and objective evaluation; LSFs were used for JD-GMM modeling and classification; and 513-dimensional spectral envelopes were used for conversion. For training of FW functions, LSFs and high-dimensional spectral envelope were used in different FW methods, and will be described in the next subsection.

4.2. Baseline methods and settings

Two baseline methods were used in our evaluation experiments. As JD-GMM has been compared with frequency warping based method in [15], [16], frequency warping methods achieved better performance than JD-GMM, and the focus of this paper is the warping function, thus the classic JD-GMM is not used as a baseline system.

- DFW: The frequency warping functions were calculated by Dynamic Frequency Warping (DFW) [12]. Dynamic spectral feature was used for distance matrix calculation. According to the results of our preliminary experiments, we set the weight as 0.2 and 0.8 for distance matrix of spectral feature and dynamic spectral feature respectively. DFW was applied on the feature of 513-dimensional spectral envelope.
- AMF: For the automatic mapping of formants (AMF) method, we used the same setting as [15].
- CFW (proposed): The correlation-based frequency warping (CFW) was used for the calculation of FW functions on 513-dimensional spectral envelope. Since the warping function was calculated on the mean spectral envelope of each acoustic classes, the segment length and shift were automatically determined by the correspond LSF feature pair, as illustrated in Section 3.

FW methods generally modify the frequency axis of source spectra, while the relative amplitude still remains. In order to minimize the error between the converted and target signals, amplitude scaling (AS) function was used. For evaluating the performance of different frequency alignment, all the experiments with AS shared the same AS function, mentioned in [16]. The efficiency of AS function is related to FW function, it will be improved by high accuracy FW function.

The idea of weighted frequency warping [15] method was used in this work. According to the results of our preliminary experiments, in our database, the optimal Gaussian components number for AMF method was eight; while for DFW and CFW, the optimal number of Gaussian components was 32.

In objective experiment, two conditions were evaluated: with and without AS. Similar to [15], in this work, only voiced frames were transformed and assessed. F0 is converted by normalising the mean and variance of the source speaker to match that of the target speaker in log-scale.

4.3. Objective evaluation

The mel-cepstral distortion (MCD) measured between converted and target spectrum, was used as the objective evaluation. The MCD for k^{th} frame was calculated as: $\mathrm{MCD[dB]} = 10/\ln 10\sqrt{2\sum_{i=1}^{24}(c_{i,k}-c_{i,k}^{conv})^2}$, where $c_{i,k}$ and $c_{i,k}^{conv}$ indicated the i^{th} target and converted MGC in frame k, respectively.

The average MCD over all evaluation pairs was reported. A lower MCD indicates smaller distortion.

Table 1 shows the results of different frequency warping methods with and without AS. Obviously, among the results without AS applied, DFW, which minimizes the spectral distance directly, gives the lowest spectral distortion (6.46 dB). However, as the physical meaning of the spectral shape is ignored, the resulted speech quality might not as good as other methods. AMF obtains the MCD at 6.79 dB, which is higher than that of CFW (6.71 dB). The results indicate that AMF is less accurate than CFW. This maybe due to the fact that AMF uses formants for warping function calculation, without considering the spectral shape.

Table 1: *Comparison of spectral distortion of different methods.*

method	MCD (dB)	method	MCD (dB)
DFW	6.46	DFW+AS	5.85
AMF	6.79	AMF+AS	5.95
CFW	6.71	**CFW+AS**	5.60

After AS, CFW+AS achieves the lowest MCD (5.60 dB), which is 0.25 dB and 0.35 dB lower than that of DFW+AS and AMF+AS methods, respectively. The result of DFW+AS has very limited compensation by AS method. The MCD decreases 0.61 dB. While the MCD of AMF+AS decreases a lot, achieving at 5.95 dB. The AS improves CFW result significantly, the gap of MCD is 1.11 dB. This is because the spectral shape is considered in CFW. After warping, a high correlation achieves between warped and target spectral envelope, which improve the effectiveness of AS function and the performance of VC.

4.4. Subjective evaluation

Listening tests were conducted to assess speech quality and speaker similarity of the proposed methods. Here AS was always used with one of the three methods, since AS has been found to be useful for enhancing speech quality [16]. 20 utterances of each speaker pair were converted for evaluation. Thus, the evaluation dataset contains 80 converted utterances for each method. Ten subjects participated in all the listening tests.

We first performed AB preference test to assess speech quality. 20 pairs were randomly selected from the 80 paired samples. In each pair, A and B were the samples from the proposed CFW+AS and DFW+AS/AMF+AS, respectively, in a random order. Each listener was asked to listen to both samples and then decide which sample is better in term of naturalness.

We then conducted an XAB test to assess the speaker similarity. In the test, similar to the AB preference test above, 20 pairs were randomly selected from the 80 paired samples. In each pair, X was the reference target sample, A and B were the converted samples of the proposed CFW+AS and DFW+AS/AMF+AS, respectively, in a random order. Note that X, A and B have the same language content. The listeners were asked to listen to the sample X first and then A and B, after that, they were asked to decide which sample is more closed to the reference target sample.

Table 2 shows the average quality and similarity results with 95% confidence intervals. In quality tests, our proposed method achieves 78.5% preference against DFW+AS and obtain the similar performance with AMF+AS. In similarity test, our proposed method gets higher preference score than the baseline methods. Since DFW method ignores the physical properties of spectral, both the quality and similarity preference scores of DFW+AS is much lower than that of our proposed method. AMF+AS and CFW+AS, both consider the physical properties of spectrum. It is expected that they have similar preference in the quality test. However, AMF+AS is affected by the accuracy of formant estimation and AMF method does not consider that spectral shape information in estimating the warping function. This explains why AMF+AS has a lower preference score in the similarity test.

Table 2: *Results of average quality and similarity preference tests with 95% confidence intervals for different methods.*

FW method	Preference score(%) (95% confidence interval)	
	Quality test	Similarity test
DFW+AS	21.5 ($\pm$ 5.85)	34 ($\pm$ 9.99)
CFW+AS	78.5 ($\pm$ 5.85)	66 ($\pm$ 9.99)
AMF+AS	49 ($\pm$ 6.82)	45.5 ($\pm$ 5.92)
CFW+AS	51 ($\pm$ 6.82)	54.5 ($\pm$ 5.92)

4.5. Discussion

The upper part of figure 3 presents the examples of spectral envelopes converted by CFW and the two baseline methods without AS function. It is obvious that the DFW result is closest to the target spectral envelope. However, DFW tries to minimize the spectral distance between the converted and the target without considering the spectral envelope. For example, in the frequency band between 2000Hz to 3000Hz, the converted spectral of DFW is almost a straight line; the spectral peak vanishes. On the other hand, the spectral envelope converted by AMF method could keep more spectral details. However, due to the inaccurate formant estimation, there is still some mismatch between the peak locations of converted and target spectrum. The converted result of CFW method is most correlated to the target.

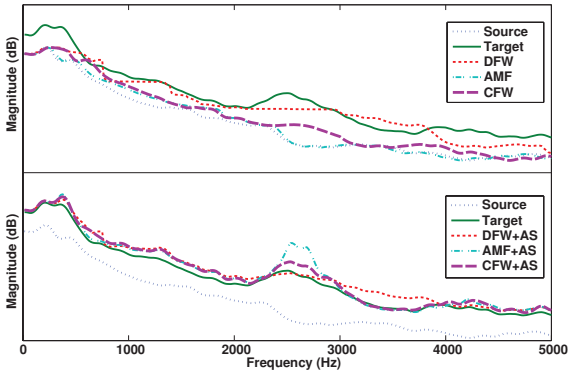

Figure 3: Illustration of spectral envelope converted by different frequency warping methods before and after AS function.

The bottom of figure 3 shows the results after AS function. Due to the accurate FW function, the converted result of CFW method is improved by AS in both distance and details. AS function also works well on the result of AMF method. However, in some parts where the FW function of AMF is not precise, such as the peak between 2500 Hz to 2800 Hz, the AS function is less effective. Because the FW function of DFW method loses lots of details, compare to other two methods, the effectiveness of AS function is significantly decreased.

5. Conclusions

In this paper, we propose a correlation-based FW method for voice conversion. Comparing to conventional FW methods, by using correlation to optimize the warping path, CFW+AS obtains more accurate FW functions. Moreover, CFW also improves the effectiveness of AS function. By allowing a soft decision on frequency axis segmentation, CFW is found to be robust to inaccurate formant estimation. The experimental results indicate that, comparing to the baseline methods, our proposed method improves both quality and speaker identity of converted speech effectively.

6. References

[1] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.

[2] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[3] E. Helander, T. Virtanen, J. Nurminen, and M. Gabbouj, "Voice conversion using partial least squares regression," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 912–921, 2010.

[4] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad, "Voice conversion using artificial neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2009, pp. 3893–3896.

[5] E. Helander, H. Silén, T. Virtanen, and M. Gabbouj, "Voice conversion using dynamic kernel partial least squares regression," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 806–817, 2012.

[6] Z. Wu, T. Virtanen, E. S. Chng, and H. Li, "Exemplar-based sparse representation with residual compensation for voice conversion," *IEEE Transactions on Audio, Speech and Language Processing*, 2014.

[7] D. Sundermann and H. Ney, "VTLN-based voice conversion," in *IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*. IEEE, 2003, pp. 556–559.

[8] M. Pitz and H. Ney, "Vocal tract normalization equals linear transformation in cepstral space," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 930–944, 2005.

[9] D. Sundermann, H. Ney, and H. Hoge, "VTLN-based cross-language voice conversion," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2003, pp. 676–681.

[10] V. Popa, J. Nurminen, and M. Gabbouj, "A novel technique for voice conversion based on style and content decomposition with bilinear models." in *INTERSPEECH*, 2009, pp. 2655–2658.

[11] D. Erro, E. Navas, and I. Hernaez, "Parametric voice conversion based on bilinear frequency warping plus amplitude scaling," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 3, pp. 556–566, 2013.

[12] H. Valbret, E. Moulines, and J.-P. Tubach, "Voice transformation using PSOLA technique," *Speech Communication*, vol. 11, no. 2, pp. 175–187, 1992.

[13] T. Toda, H. Saruwatari, and K. Shikano, "Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of straight spectrum," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2. IEEE, 2001, pp. 841–844.

[14] S. H. Mohammadi and A. Kain, "Transmutative voice conversion," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2013, pp. 6920–6924.

[15] D. Erro, A. Moreno, and A. Bonafonte, "Voice conversion based on weighted frequency warping," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 922–931, 2010.

[16] E. Godoy, O. Rosec, and T. Chonavel, "Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1313–1323, 2012.

[17] D. Erro, T. Polyakova, and A. Moreno, "On combining statistical methods and frequency warping for high-quality voice conversion," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2008, pp. 4665–4668.

[18] N.-P. V. Nielsen, J. M. Carstensen, and J. Smedsgaard, "Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping," *Journal of Chromatography A*, vol. 805, no. 1, pp. 17–35, 1998.

[19] G. Tomasi, F. van den Berg, and C. Andersson, "Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data," *Journal of Chemometrics*, vol. 18, no. 5, pp. 231–241, 2004.

[20] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 26, no. 1, pp. 43–49, 1978.

[21] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, no. 1, pp. 52–59, 1986.

[22] S. McCandless, "An algorithm for automatic formant extraction using linear prediction spectra," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 22, no. 2, pp. 135–141, 1974.

[23] R. C. Snell and F. Milinazzo, "Formant location from LPC analysis data," *IEEE Transactions on Speech and Audio Processing*, vol. 1, no. 2, pp. 129–134, 1993.

[24] D. Erro, "Intra-lingual and cross-lingual voice conversion using harmonic plus stochastic models," Ph.D. dissertation, Universitat Politècnica de Catalunya, 2008.

[25] A. B. Kain, "High resolution voice transformation," Ph.D. dissertation, Rockford College, 2001.

[26] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech communication*, vol. 27, no. 3, pp. 187–207, 1999.