On Using Warping Function for LSFs Transformation in a Voice Conversion System

Zdeněk Hanzlíček and Jindřich Matoušek
University of West Bohemia, Faculty of Applied Sciences, Department of Cybernetics
Univerzitní 8, 306 14 Plzeň, Czech Republic
{zhanzlic, jmatouse}@kky.zcu.cz

Abstract

In this paper, a new approach to line spectral frequencies transformation is introduced and employed in the voice conversion framework. This approach stems from the fact that LSFs are some specific points on the frequency axis and their positions determine the shape of the spectral envelope. Thus, they could be transformed directly by frequency axis warping. Two warping functions were designed specially for LSFs and compared with the traditional GMM-based conversion function. Listening tests and mathematical evaluation revealed that speech transformed by using proposed warping functions is of higher quality and does not suffer from oversmoothing which is common for GMM-based transformation. On the other hand, the speaker identity is slightly better transformed by GMM-based conversion. However, it is possible to combine these two approaches to obtain a compromise between quality and speaker identity.

1. Introduction

In this paper, a new function for transformation of line spectral frequencies – LSFs [1] is introduced. It is also employed and tested in our voice conversion system [2]. Usually, LSFs are transformed by employing vector functions. However, we proposed a new approach flowing from the fact that LSFs are some specific points on the (normalised) frequency axis; they are mostly located near formant frequencies. The transformation of LSFs can be interpreted as a shift of particular frequencies. Thus, it should be possible to transform LSFs directly by frequency axis warping. The main idea is depicted in Fig. 1.

Two different warping functions proposed specially for LSF transformation were compared with traditional GMM-based conversion function [3].

This paper is organized as follows. In Section 2, our baseline conversion system is briefly described. Two warping functions for LSF transformation are proposed in Section 3. Experiments and results are presented in Section 4. Finally, Section 5 concludes this paper.

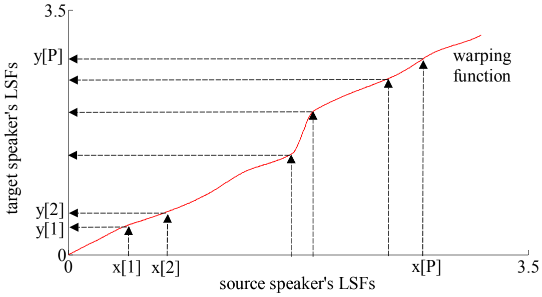

Fig 1. LSF transformation by using a warping function.

2. Baseline conversion system

Our voice conversion system employs parallel training data. Voiced speech is analysed pitch synchronously; each segment is three pitch periods long and the shift of analysis window is one pitch period. Unvoiced segments are 10 msec long with 5 msec overlap. The spectral envelope of each frame is obtained by using true envelope estimator [4]. Then, the envelope is approximated with spectrum of an all-pole model which is represented by its line spectral frequencies. Experimentally, 35 was selected as an optimal parameter order for sufficient spectral envelope approximation. Moreover, the residual (excitation) signal for each speech segment can be obtained by inverse filtration with the corresponding all-pole model. Frames of particular pairs of utterances are time-aligned by using DTW algorithm.

In our baseline system, LSF features are converted by employing traditional GMM-based transformation function [3]; fundamental frequency is converted by using Gaussian (mean/variance) normalization. And the suitable target residual signal is estimated by so-called residual prediction [5]. For more detailed information on our baseline system, see [2].

3. LSF warping functions

Transformation by using warping functions is a popular approach within the voice conversion framework. Recently, Erro at al. [6] proposed a combination of statistical methods and warping function. Usually, warping functions are used for transformation of amplitude spectrum (or spectral envelope). In our approach, warping function is used for shifting the position of particular line spectral frequencies. Thus, such a warping function is a scalar function *f* which transforms all the components *x*[*i*] (for *i* = 1,... *P*) of source LSF vector *x*

$$\tilde{y}[i] = f(x[i]) \qquad (1)$$

Similarly as in [6], the joint LSF feature space is divided into *K* classes and each class has its respective warping function. Aligned training data $z = [x^T, y^T]^T$ can be described with a GMM with *K* components where each component corresponds to one class. However for larger number of classes or less amount of training data, a non-probabilistic clustering (e.g. by using k-means algorithm) is better to use instead. Then, each class is described by its centroid

$$\mu_z^k = \left[\mu_x^k[1], \ldots \mu_x^k[P], \mu_y^k[1], \ldots \mu_y^k[P]\right]^T$$

and diagonal covariance matrix

$$\Sigma_z^k = diag\left[\sigma_x^k[1], \ldots \sigma_x^k[P], \sigma_y^k[1], \ldots \sigma_y^k[P]\right].$$

Both μ_z and Σ_z can be decomposed into parts which correspond to source and target speaker: μ_x, μ_y, Σ_x, Σ_y.

For each class *k*, an individual warping function f_k is defined and the final position of *i*-th LSF is given as a weighted average over all classes

$$\tilde{y}[i] = \sum_{k=1}^{K} w_k(x) f_k(x[i]) . \qquad (2)$$

In the case of GMM-based LSF space description, the weight $w_k(x)$ of the *k*-th class is given as the conditional probability $p(k|x)$. Or in the case of non-probabilistic clustering, the weight is defined as

$$w_k(x) = \frac{1/d_k(x)}{\sum_{j=1}^{1} 1/d_j(x)} , \qquad (3)$$

where $d_k(x)$ is Mahalanobis distance

$$d_k(x) = \left[(x-\mu_x^k)^T (\Sigma_x^k)^{-1} (x-\mu_x^k)\right]^\gamma \qquad (4)$$

Parameter γ controls the smoothing among results of particular classes. The higher is the value of γ, the higher are the weights of closer classes in comparison with weights of remoter classes. Experimentally, we set $\gamma = 4$, though the value probably depends on the number of classes too.

3.1. Piecewise linear warping function (WL)

First, the mean vectors of all classes are extended

$$\begin{gathered} \mu_x^k[0] = \mu_y^k[0] = 0 \\ \mu_x^k[P+1] = \mu_y^k[P+1] = \pi \end{gathered} \qquad (5)$$

The warping function is divided into *P* + 1 intervals. For the *j*-th interval

$$x[i] \in \left\langle \mu_x^k[j-1], \mu_x^k[j] \right\rangle \qquad (6)$$

the warping function is defined as a linear function

$$\tilde{y}[i] = f_k^j(x[i]) = a_k^j x[i] + b_k^j . \qquad (7)$$

All the unknown parameters are determined from the requirements for the boundary points of particular intervals; e.g. for the *j*-th interval, we require

$$\begin{gathered} f_k^j(\mu_x^k[j-1]) = \mu_y^k[j-1] \\ f_k^j(\mu_x^k[j]) = \mu_y^k[j] \end{gathered} \qquad (8)$$

3.2 Piecewise cubic warping function (WC)

Gaussian (or mean-variance) normalization for two scalar Gaussian variables

$$X_1 \sim \boldsymbol{N}\{\mu_1, \sigma_1\} \qquad X_2 \sim \boldsymbol{N}\{\mu_2, \sigma_2\} \qquad (9)$$

is given as

$$X_2 = \mu_2 + \frac{\sigma_2}{\sigma_1}(X_1 - \mu_1) . \qquad (10)$$

The tangent of that transformation function is given as σ_2/σ_1. A similar feature will be demanded for our warping function. First, we introduce

$$\begin{gathered} \delta_k[i] = \sigma_y^k[i]/\sigma_x^k[i] \qquad i = 1, 2, \ldots P \\ \delta_k[0] = \delta_k[P+1] = 1 \end{gathered} \qquad (11)$$

Warping function in the *j*-th interval

$$x[i] \in \left\langle \mu_x^k[j-1], \mu_x^k[j] \right\rangle \qquad (12)$$

is defined as a cubic function

$$f_k^j(x[i]) = a_k^j x^3[i] + b_k^j x^2[i] + c_k^j x[i] + d_k^j \quad (13)$$

and its derivation is given as

$$g_k^j(x[i]) = 3\,a_k^j x^2[i] + 2\,b_k^j x[i] + c_k^j \quad (14)$$

Again, all unknown parameters are determined from the requirements for the boundary points of particular intervals; for the warping function, we require

$$\begin{aligned} f_k^j(\mu_x^k[j-1]) &= \mu_y^k[j-1] \\ f_k^j(\mu_x^k[j]) &= \mu_y^k[j] \end{aligned} \quad (15)$$

and for its derivation (in accordance with (10))

$$\begin{aligned} g_k^j(\mu_x^k[j-1]) &= \delta_k[j-1] \\ g_k^j(\mu_x^k[j]) &= \delta_k[j] \end{aligned} \quad (16)$$

Comparison of piecewise linear and piecewise cubic warping function is presented in Fig. 2.

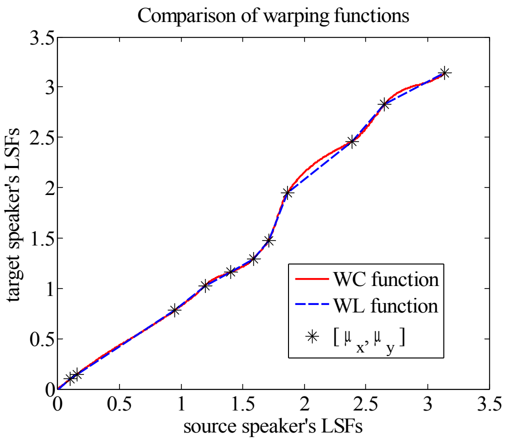

Fig. 2. Comparison between WC and WL function.

4. Experiments and results

In all experiments, 40 utterances were used for training and another 15 utterances for assessment. All utterances were Czech sentences, about 6-7 words long. First, one female (reference) speaker recorded all the utterances. Then 4 other speakers (2 males and 2 females, denoted M1, M2, F1 and F2) listened to these reference recordings and repeated them. In our experiments, conversion from reference speaker to all other speakers was performed.

4.1 Objective evaluation

For performance evaluation of our VC system, so-called performance index P_{LSF} was employed

$$P_{LSF} = 1 - \frac{\sum_{n=1}^{N} d(\tilde{y}_n, y_n)}{\sum_{n=1}^{N} d(x_n, y_n)} . \quad (17)$$

The higher is the P_{LSF} value, the higher is the similarity between transformed and target utterances in comparison with the original similarity between source and target utterances.

Comparison between WL and WC transformation is presented in Fig. 3. Obviously, for a lower parameter order, WC slightly outperforms WL. However, for a higher parameter order, warping functions for WL and WC are defined by more points and their shapes differ insignificantly. Thus WL and WC perform practically the same.

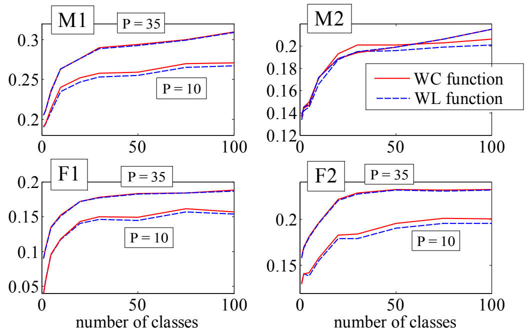

Fig. 3. Dependency between number of classes and P_{LSF} for particular speakers. Parameter order was set to 10 and 35.

The direct comparison between GMM-based and warping transformation is quite difficult. GMM-based function performs best for a low number of mixtures (about 10). For a higher number of mixtures, the value of P_{LSF} decreases; probably more training data are necessary for a superior estimation of GMM parameters. On the other hand WC or WL graduate their performance for a high number of classes (above 100). The comparison is presented in Tab. 1.

Tab. 1. Comparison of P_{LSF}.

Function	Number of classes	Target speaker			
		M1	M2	F1	F2
GMM	5	0.424	0.326	0.303	0.346
	10	0.425	0.328	0.301	0.345
	20	0.421	0.317	0.296	0.336
WC	5	0.235	0.149	0.135	0.184
	10	0.264	0.172	0.153	0.203
	20	0.279	0.188	0.172	0.221
	150	0.314	0.219	0.189	0.240

The degree of parameter oversmoothing can be scored by average global variance ratio

$$R_{GV} = \frac{1}{P} \sum_{p=1}^{P} \frac{GV(\tilde{y}[p])}{GV(y[p])} \tag{18}$$

where $GV(y[p])$ is the (global) variance of the p-th component of parameter vector y over the whole utterance. The desired value of R_{GV} is about 1; lower values signify oversmoothed parameters.

The results for GMM-based and WC transformation are compared in Tab. 2. Obviously, R_{GV} for WC is significantly better. However, for a higher number of classes, R_{GV} value decreases. Probably, the smoothing parameter γ in (4) should be selected with regard to number of classes.

Tab. 2. Comparison of R_{GV}.

Function	Number of classes	Target speaker			
		M1	M2	F1	F2
GMM	5	0.642	0.676	0.670	0.686
	10	0.655	0.691	0.700	0.683
	20	0.667	0.705	0.703	0.706
WC	5	1.081	0.968	1.005	0.960
	10	1.038	0.953	1.004	0.958
	20	0.997	0.939	0.984	0.946
	150	0.935	0.912	0.955	0.912

4.2 Listening tests

The proposed transformation methods were also evaluated in listening tests. 10 participants took part in those tests; each of them listened to 20 quintuples of utterances: from source and target speakers and 3 converted utterances (in random order): transformed by using GMM-based transformation, WC function and a compromise between them given by

$$\tilde{y}_{GW} = \beta \, \tilde{y}_{GMM} + (1-\beta) \, \tilde{y}_{WC} \tag{19}$$

where β was set to 0.5.

Listeners should order the converted utterances descending according to their quality and also (independently) according to the voice similarity with the target voice. Then, average quality and similarity rankings were calculated for each method. The lower value means the better quality or similarity. Results are presented in Tab. 3.

Results of listening tests were analysed by using the paired t-test. The system utilizing the WC function produces speech of higher quality (but of less similarity) than the baseline system with P-value less than 0.0001, which is extremely statistically significant.

Tab. 3. Average rankings of particular methods.

	GMM	GW	WC
Similarity	1.83 ± 0.56	1.94 ± 0.31	2.23 ± 0.59
Quality	2.34 ± 0.55	2.02 ± 0.45	1.65 ± 0.62

5. Conclusion

In this paper, a new approach to line spectral frequencies transformation, based on frequency axes warping was introduced. Two special functions were proposed – piecewise linear and piecewise cubic warping function.

Both objective and subjective comparison with traditional GMM-based transformation revealed that speech transformed by using warping function is of higher quality and does not suffer from oversmoothing. On the other hand, the speaker identity is better transformed by GMM-based conversion. However, it is possible to combine these two approaches and obtain a compromise between quality and speaker identity.

6. Acknowledgement

Support for this work was provided by the Ministry of Education of the Czech Rep., project No. 2C06020.

7. References

[1] T. Bäckström, "Linear Predictive Modelling of Speech Constraints and Line Spectrum Pair Decomposition", Ph.D. dissertation, Helsinki University of Technology, 2004.

[2] Z. Hanzlíček and J. Matoušek: "F0 Transformation within the Voice Conversion Framework", *Interspeech*, Antwerp, Belgium, 2007, pp. 1961-1964.

[3] 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.

[4] F. Villavicencio, A. Röbel and X. Rodet, "Improving LPC Spectral Envelope Extraction of Voiced Speech by True Envelope Estimation", *ICASSP*, Toulouse, France, 2006, pp. 869-872.

[5] D. Sündermann, A. Bonafonte, H. Ney and H. Höge, "A Study on Residual Prediction Techniques for Voice Conversion", *ICASSP*, Philadelphia, USA, 2005, pp. 13-16.

[6] D. Erro, T. Polyakova and A. Moreno, "On Combining Statistical Methods and Frequency Warping for High-Quality Voice Conversion", *ICASSP*, Las Vegas, USA, 2008, pp. 4665-4668.