Journal of Southeast University (English Edition) Vol. 28, No. 2, pp. 140 - 144 June 2012 ISSN 1003—7985

Efficient fundamental frequency transformation for voice conversion

Song Peng[1] Jin Yun[1,2] Bao Yongqiang[3] Zhao Li[1] Zou Cairong[1]

([1] Key Laboratory of Underwater Acoustic Signal Processing of Ministry of Education, Southeast University, Nanjing 210096, China)
([2] School of Physics and Electronic Engineering, Xuzhou Normal University, Xuzhou 221116, China)
([3] School of Communication Engineering, Nanjing Institute of Technology, Nanjing 211167, China)

Abstract: In order to improve the performance of voice conversion, the fundamental frequency (F0) transformation methods are investigated, and an efficient F0 transformation algorithm is proposed. First, unlike the traditional linear transformation methods, the relationships between F0s and spectral parameters are explored. In each component of the Gaussian mixture model (GMM), the F0s are predicted from the converted spectral parameters using the support vector regression (SVR) method. Then, in order to reduce the over-smoothing caused by the statistical average of the GMM, a mixed transformation method combining SVR with the traditional mean-variance linear (MVL) conversion is presented. Meanwhile, the adaptive median filter, prevalent in image processing, is adopted to solve the discontinuity problem caused by the frame-wise transformation. Objective and subjective experiments are carried out to evaluate the performance of the proposed method. The results demonstrate that the proposed method outperforms the traditional F0 transformation methods in terms of the similarity and the quality.

Key words: F0 prediction; support vector regression; mean-variance linear conversion; adaptive median filter
doi: 10.3969/j.issn.1003-7985.2012.02.002

Voice conversion is a technique converting the speech spoken by a source speaker so that it sounds as if it was spoken by a target speaker. It has many practical applications, such as flexible text-to-speech synthesis, identity disguise, help to speech-impaired people, and low bit rate communication, etc.

As it is known to all, the speech signal can be considered a consequence of many factors, among which, spectral characteristics, prosodic features, and speaking styles contribute greatly to the speaker individuality. In the past decades, many efforts have been made for the transformation of the spectral parameters, and great progresses have been made. From the state-of-the-art references, the Gaussian mixture model (GMM) method[1-2] has proved to be the most prevalent and well-known, and it is chosen for the spectral transformation in this paper. Meanwhile, the studies on prosody transformation are few. However, prosodic features, especially, the F0s are also key factors in the manifestation of the speaker individuality.

The context of the paper is focused on the transformation of F0s. There are two kinds of mainstream F0 transformation approaches[3]. One is the linear transformation on a frame-by-frame basis, including the mean-variance linear (MVL) method, the *N*-th order polynomial method, and the GMM method. The other is the transformation of the whole F0 contours using the codebook methods, which attempts to impart the entire contours onto the source speech from the target references. Although the latter one can show better performance in most cases, it is very difficult to perform as it greatly depends on lexical and paralinguistic factors[4]. So the first method is investigated and improved.

In this paper, a novel F0 transformation method is proposed. Being different from the traditional conversion methods, the F0s are predicted from the converted spectral parameters using the SVR method in each component of the GMM. In order to reduce over-smoothing, a mixed F0 prediction and MVL method is also presented. Meanwhile, the conversion is carried out on a frame-by-frame basis, ignoring the correlations between the adjacent frames, and it will cause the discontinuities in the converted F0s. So an adaptive median filter is adopted to reduce this problem. Finally, experiments are carried out to confirm the efficiency of the proposed approach.

1 Spectral Conversion

The GMM method[2] is chosen for the spectral conversion. Let $\boldsymbol{x} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T\}$ and $\boldsymbol{y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_T\}$ denote the sequences of the spectral parameters of the source speaker and the target speaker, respectively, where $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ are d-dimensional vectors, and T is the number of frames. The dynamic time warping (DTW) technique is employed to align the spectral parameters of the source

Received 2012-02-19.

Biographies: Song Peng (1983—), male, graduate; Zhao Li (corresponding author), male, doctor, professor, zhaoli@seu.edu.cn.

Foundation items: The National Natural Science Foundation of China (No. 60975017), the Natural Science Foundation of Guangdong Province (No. 10252800001000001), the Natural Science Foundation of Higher Education Institutions of Jiangsu Province (No. 10KJB510005).
Citation: Song Peng, Jin Yun, Bao Yongqiang, et al. Efficient fundamental frequency transformation for voice conversion [J]. Journal of Southeast University (English Edition), 2012, 28(2): 140 - 144. [doi: 10.3969/j.issn.1003-7985.2012.02.002]

speaker to the counterparts of the target one. The distribution of $\begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix}$ is described by a GMM as

$$p(\boldsymbol{x},\boldsymbol{y}) = \sum_{m=1}^{M} \alpha_m N\left(\begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix}, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m \right) \tag{1}$$

where α_m is the prior probability of $\begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix}$ belonging to the m-th component, and it satisfies $\sum_{m=1}^{M} \alpha_m = 1$. $\boldsymbol{\mu}_m$ and $\boldsymbol{\Sigma}_m$ are the mean and covariance matrices, respectively. The unknown parameters $(\alpha_m, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$ can be estimated by the expectation maximization (EM) algorithm, and $\boldsymbol{\mu}_m$ and $\boldsymbol{\Sigma}_m$ are divided into blocks corresponding to the components of the source and target speakers, which take the forms as

$$\boldsymbol{\mu}_m = \begin{bmatrix} \boldsymbol{\mu}_m^{x} \\ \boldsymbol{\mu}_m^{y} \end{bmatrix}, \quad \boldsymbol{\Sigma}_m = \begin{bmatrix} \boldsymbol{\Sigma}_m^{xx} & \boldsymbol{\Sigma}_m^{xy} \\ \boldsymbol{\Sigma}_m^{yx} & \boldsymbol{\Sigma}_m^{yy} \end{bmatrix} \tag{2}$$

The transformation between $\boldsymbol{x}$ and $\boldsymbol{y}$ is given by

$$\hat{\boldsymbol{y}} = E(\boldsymbol{y} \mid \boldsymbol{x}) = \sum_{m=1}^{M} p(m \mid \boldsymbol{x}) \left[\boldsymbol{\mu}_m^{y} + \frac{\boldsymbol{\Sigma}_m^{yx}}{\boldsymbol{\Sigma}_m^{xx}} (\boldsymbol{x} - \boldsymbol{\mu}_m^{x}) \right] \tag{3}$$

$$p(m \mid \boldsymbol{x}) = \frac{\alpha_m N(\boldsymbol{x}, \boldsymbol{\mu}_m^{x}, \boldsymbol{\Sigma}_m^{xx})}{\sum_{k=1}^{M} \alpha_k N(\boldsymbol{x}, \boldsymbol{\mu}_k^{x}, \boldsymbol{\Sigma}_k^{xx})} \tag{4}$$

where $p(m \mid \boldsymbol{x})$ is the conditional probability of $\boldsymbol{x}$ belonging to the m-th component.

2 F0 Transformation

2.1 Baseline method

One simple and popular F0 transformation method is the mean-variance linear (MVL) conversion[3]. Let f_x and f_y denote the F0s of the source and target speakers, respectively. The underlying assumption of this approach is that f_x and f_y belong to a GMM, and the conversion function takes the form as

$$\hat{f}_{y_{\mathrm{MVL}}} = \mu_y + \frac{\sigma_y}{\sigma_x}(f_x - \mu_x) \tag{5}$$

where μ_x and μ_y are the means; σ_x and σ_y are the standard deviations of f_x and f_y, respectively. In this method, the global means and ranges of the F0s are converted, while preserving the shapes of the F0 contours of the source speaker. Another approach is based on the GMM[4]. The F0s of the source and target speakers are modeled in a GMM, and the transformation function has a similar form as shown in Eq. (3).

2.2 Proposed transformation method

2.2.1 F0 prediction using SVR

In the previous section, the linear F0 transformation methods are introduced, which have been proved to achieve satisfactory results. But there is still much room for improvement. Ref. [5] proves that there exist some relationships between the F0s and the spectral trajectory. In this paper, a novel F0 transformation method is proposed. Unlike the traditional methods, the F0s are predicted from the converted spectral parameters.

In the training phase, we assume that the F0s and the spectral parameters of the target speaker are correlated by the support vector regression (SVR) method [6-7]. Given the training set $\{(\boldsymbol{y}_1, f_{y_1}), (\boldsymbol{y}_2, f_{y_2}), \ldots, (\boldsymbol{y}_T, f_{y_T})\}$, the regression function can be given by

$$\hat{f}_{y_{\mathrm{SVR}}}(\boldsymbol{y}) = F(\boldsymbol{y}) = \langle \boldsymbol{w}, \boldsymbol{y} \rangle + b \tag{6}$$

Introducing the slack variables ξ_i and ξ_i^*, Eq. (6) can be resolved by

$$\begin{aligned} &\min\left[\frac{1}{2} \| \boldsymbol{w} \|^2 + C \sum_{i=1}^{T} (\xi_i + \xi_i^*) \right] \\ &f_{y_i} - \langle \boldsymbol{w}, \boldsymbol{y}_i \rangle - b \leqslant \varepsilon + \xi_i \\ &\text{s.t. } \langle \boldsymbol{w}, \boldsymbol{y}_i \rangle + b - f_{y_i} \leqslant \varepsilon + \xi_i^* \\ &\xi_i, \xi_i^* \geqslant 0 \end{aligned} \tag{7}$$

where $C > 0$ is a constant, and ε is a penalty variable. Utilizing the Lagrange function and the dual form, we can obtain the dual optimization problem as

$$\begin{aligned} &\max\left[-\frac{1}{2} \sum_{i,j=1}^{T} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) \langle \boldsymbol{y}_i, \boldsymbol{y}_j \rangle - \right. \\ &\left. \varepsilon \sum_{i=1}^{T} (\alpha_i + \alpha_i^*) + \sum_{i=1}^{T} f_{y_i} (\alpha_i - \alpha_i^*) \right] \\ &\text{s.t.} \quad \sum_{i=1}^{T} (\alpha_i - \alpha_i^*) = 0 \\ &\alpha_i, \alpha_i^* \in [0, C] \end{aligned} \tag{8}$$

The unknown parameters $\boldsymbol{w}$ and b can be computed, and the regression function is modified as

$$F(\boldsymbol{y}_i) = \sum_{i=1}^{T} (\alpha_i^* - \alpha_i) K(\boldsymbol{y}_i, \boldsymbol{y}) + b \tag{9}$$

where $K(\boldsymbol{y}_i, \boldsymbol{y})$ is a kernel function, and the common and popular radial basis function (RBF) is adopted as the kernel function, which takes the form as

$$K(\boldsymbol{y}_i, \boldsymbol{y}_j) = \exp\left(-\frac{\| \boldsymbol{y}_i - \boldsymbol{y}_j \|^2}{2\sigma^2} \right) \tag{10}$$

As it is well-known to all, it is unlikely to use one single transformation for all the data. So the SVR method is combined with the GMM, and multiple local regressions are made to overcome this problem. Similar to the spectral conversion methods, the predicted value of $\hat{f}_{y_{\mathrm{SVR}}}$ is cal-

culated by a weighted sum of local $\hat{f}_{y,m}$, which is given by

$$\hat{f}_{y_{\mathrm{SVR}}} = \sum_{m=1}^{M} \beta_m \hat{f}_{y,m} \tag{11}$$

In each component of the GMM, the local values of F0s $\hat{f}_{y,m}$ are predicted from the local spectral parameters of the converted speech, and the value of β_m is equal to $p(m \mid \boldsymbol{x})$ (the formula of which is shown by Eq. (4)).

2.2.2 Combining F0 prediction and MVL methods

One of the main problems of the proposed method is the over-smoothing caused by the statistical average of the GMM. In order to solve this problem, a hybrid solution is proposed. Specifically, we combine the F0 prediction with the MVL conversion method, and the F0 values of the converted speech can be seen as a weighted sum of F0 values obtained from the two methods,

$$\hat{f}_y = \lambda \hat{f}_{y_{\mathrm{SVR}}} + (1-\lambda)\hat{f}_{y_{\mathrm{MVL}}} \tag{12}$$

where λ is a weighting coefficient, and $0 \leqslant \lambda \leqslant 1$. The objective evaluation method provided in section 3.2 is employed to determine the optimal value.

2.3 Post-processing by adaptive median filter

Currently, the F0 transformation is carried out on a frame-by-frame basis, which ignores the relationships between the neighboring frames, and it will introduce the discontinuities in the converted F0s. In order to solve this problem, the delta values of the F0s are considered[7]. An RAMF (ranked-order based adaptive median filter) technique previously applied to image processing is employed in this paper[8]. Unlike the median filter, the RAMF can effectively remove the unrepresentative F0 values, while keeping the detailed information. Assuming that W is a rectangular filtering window, f_{cur} is the value of a current point, and $f_{\min}$, $f_{\max}$ and f_{med} are the minimum, maximum and median values of the points in the filtering window, the RMAF can be regarded as a two-level structure: level A and level B.

Level A

$$A_1 = f_{\mathrm{med}} - f_{\min}$$
$$A_2 = f_{\mathrm{med}} - f_{\max}$$

If $A_1 > 0$ and $A_2 < 0$, then turn to level B, or increase the size of W to repeat level A until the size of W exceeds the maximum set value, then f_{cur} is used as the output.

Level B

$$B_1 = f_{\mathrm{cur}} - f_{\min}$$
$$B_2 = f_{\mathrm{cur}} - f_{\max}$$

If $B_1 > 0$ and $B_2 < 0$, then f_{cur} is used as the output, or f_{med} is adopted as the output.

3 Experimental Results and Discussion

3.1 Experimental preparation

The experiments are carried out on the CMU ARCTIC dataset[9]. The subsets of one US male (RMS) and two US females (SLT and CLB) are employed, respectively. 200 parallel utterances of each speaker are provided for the material, 100 of which are used for training, while the other 100 are for testing. Two kinds of F0 conversion strategies are designed to evaluate the performance of the proposed method. They are SLT-to-RMS (female-to-male), and SLT-to-CLB (female-to-female), respectively. The GMM method is used for spectral conversion, and the number of GMM components is optimized as 16. The STRAIGHT method[10] is chosen for analysis and synthesis of the speech signal. The 16th LSFs (linear spectral frequencies) are extracted to represent the spectral trajectory, and the F0s are processed in the log domain.

Both objective and subjective experiments are conducted. The Pearson product-moment correlation coefficient (PPMCC) is used for the objective experiment, while the ABX and mean opinion score (MOS) methods are adopted for the subjective evaluation. The MVL method, the GMM method, the proposed F0 prediction method(FP), the hybrid FP and MVL method(FP + MVL), and the proposed method considering RAMF post-processing (FP + MVL + RAMF) are compared. The weighting coefficient λ is optimized as 0.7 for the experiments, and eight experienced listeners are hired to make the subjective tests.

3.2 Objective evaluation

Correlation is a common method to evaluate the performance of the converted F0s objectively[4], the correlation coefficient r is a measure of the correlation between the converted and target F0s, with values between -1 and 1. It is obvious that the highest value of r is 1, which demonstrates the converted F0s are closest to the target ones. The results are shown in Tab. 1. It can be found that the FP method can outperform the baseline methods such as the MVL and the GMM. Although the values of r of the hybrid FP and MVL methods are a little lower than those of the FP method, which are caused by the inaccurate prediction of the MVL method. The subjective tests in next section will show that it can efficiently improve the quality of the converted speech. Meanwhile, introducing the RAMF strategy can enhance the performance to

Tab. 1 Correlation results of different methods

Methods	Female-to-female	Female-to-male
MVL	0.652	0.589
GMM	0.673	0.612
FP	0.675	0.617
FP + MVL	0.674	0.615
FP + MVL + RAMF	0.677	0.619

some extent. It can also be found that the correlations of the intra-gender are stronger than those of the inter-gender.

3.3 Subjective evaluation

For the subjective experiments, an ABX test is conducted to evaluate the perceptual performance. The target speech is X, and A and B are either source speech or the converted speech using different F0 transformation methods. The listeners are asked to choose whether A or B is closer to X, or choose N/A demonstrating that A and B are equal. No prior information is given to the listeners. Tab. 2 and Tab. 3 summarize the results of the proposed method compared with the baseline methods. It is obvious that the proposed strategy is superior to the baseline MVL and GMM methods. Compared with the MVL method, the proposed method shows clear superiority, while compared with the GMM method, the results of the proposed method are also satisfactory, although the values of N/A are a little higher.

Tab. 2 Comparison of proposed method and MVL method %

Strategies	MVL	Proposed method	N/A
Female-to-female	31	62	7
Female-to-male	19	75	6

Tab. 3 Comparison of proposed method and GMM method %

Strategies	GMM	Proposed method	N/A
Female-to-female	29	53	18
Female-to-male	32	59	9

The MOS test is also carried out to evaluate the quality of the converted speech. The listeners are asked to rate the quality on a scale of five values between one for "unsatisfactory" and five for "excellent". The converted speech using different F0 transformation methods is chosen to make pairs with the target speech. The results are depicted in Fig. 1. The confidence interval is set as 95%,

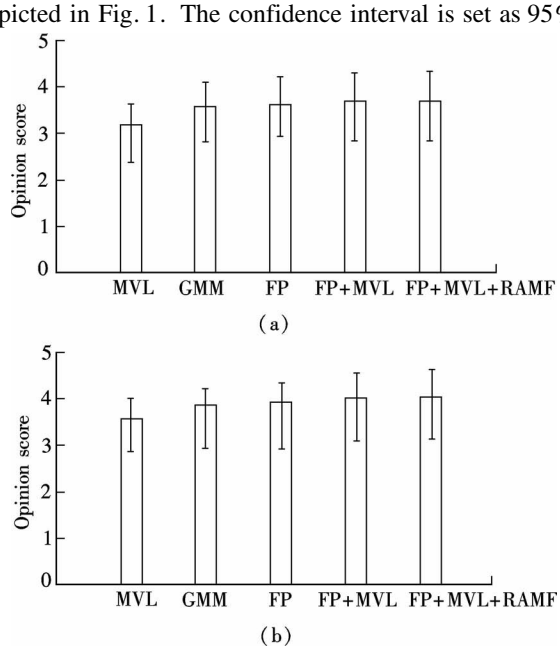

Fig. 1 Results of MOS test. (a) Female-to-female; (b) Female-to-male

and the average mean opinion scores are given for each method. We can find that the proposed method employing RAMF can greatly improve the quality compared with the baseline methods.

4 Conclusion

A novel F0 transformation method is proposed in this paper. The F0s of the converted speech are predicted from the spectral parameters using the SVR method, and the SVR is performed in each component of the GMM to increase the accuracy of prediction. In order to efficiently reduce the over-smoothing, a hybrid method combining the SVR prediction and MVL is also proposed. Meanwhile, an adaptive median filter is also employed to reduce the discontinuities in the converted F0s. Experimental results demonstrate that the proposed method outperforms the traditional MVL and GMM methods.

References

[1] Stylianou Y, Cappé O, Moulines E. Continuous probabilistic transform for voice conversion [J]. *IEEE Transactions on Speech and Audio Processing*, 1998, **6**(2):131-142.

[2] Kain A, Macon M W. Spectral voice conversion for text-to-speech synthesis [C]//*International Conference on Acoustics, Speech, and Signal Processing*. Seattle, USA, 1998: 285-288.

[3] Inanoglu Z. Transforming pitch in a voice conversion framework [D]. Cambridge, UK: St. Edmund's College of the University of Cambridge, 2003: 28-32.

[4] Wu Z Z, Kinnunen T, Chng E S, et al. Text-independent F0 transformation with non-parallel data for voice conversion [C]//11*th Annual Conference of the International Speech Communication Association*. Makuhari, Japan, 2010: 1732-1735.

[5] Shao X, Milner B. Pitch prediction from MFCC vectors for speech reconstruction [C]//*Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*. Montreal, Canada, 2004: 97-100.

[6] Basak D, Pal S, Patranabis D C. Support vector regression [J]. *Neural Information Processing—Letters and Reviews*, 2007, **11**(10): 203-224.

[7] Song P, Bao Y Q, Zhao L, et al. Voice conversion using support vector regression [J]. *Electronics Letters*, 2011, **47**(18): 1045-1046.

[8] Hwang H, Haddad R A. Adaptive median filters: new algorithms and results [J]. *IEEE Transactions on Image Processing*, 1995, **4**(4): 499-502.

[9] Kominek J, Black A W. The CMU Arctic speech databases [C]//*Proceedings of the 5th ISCA Speech Synthesis Workshop*. Pittsburgh, USA, 2004: 223-224.

[10] Kawahara H, Masuda-Katsuse I, de Cheveigné A. Restructuring speech representation using pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds [J]. *Speech Communication*, 1999, **27**(3): 187-207.

用于语音转换的有效基音频率转换算法

宋 鹏[1] 金 赟[1,2] 包永强[3] 赵 力[1] 邹采荣[1]

([1] 东南大学水声信号处理教育部重点实验室,南京 210096)
([2] 徐州师范大学物理与电子工程学院,徐州 221116)
([3] 南京工程学院通信工程学院,南京 211167)

摘要:为了改善语音转换的性能,对基音频率转换方法进行了研究,并提出了一种有效的转换算法. 首先,不同于传统的线性变换方法,对基音频率和频谱特征的内在关系进行了分析,在 GMM 中的每一分量,基音频率通过 SVR 方法从转换后的频谱特征预测得到. 然后,为了缓解 GMM 统计平均带来的过平滑问题,将传统的均值-方差转换方法和 SVR 方法相结合. 同时,引入广泛应用于图像处理的自适应中值滤波来解决由基于帧转换引起的不连续问题. 通过主客观评价方法对转换后的语音质量进行了测试,结果表明:该方法无论在语音的相似度还是转换语音的质量上,都取得了比传统方法更好的效果.

关键词:基音频率预测;支持向量回归;均值-方差线性转换;自适应中值滤波
中图分类号:TN912. 3