Non-linear Pitch Modification in Voice Conversion Using Artificial Neural Networks

Bajibabu Bollepalli, Jonas Beskow, and Joakim Gustafson

Department of Speech, Music and Hearing, KTH, Sweden

Abstract. Majority of the current voice conversion methods do not focus on the modelling local variations of pitch contour, but only on linear modification of the pitch values, based on means and standard deviations. However, a significant amount of speaker related information is also present in pitch contour. In this paper we propose a non-linear pitch modification method for mapping the pitch contours of the source speaker according to the target speaker pitch contours. This work is done within the framework of Artificial Neural Networks (ANNs) based voice conversion. The pitch contours are represented with Discrete Cosine Transform (DCT) coefficients at the segmental level. The results evaluated using subjective and objective measures confirm that the proposed method performed better in mimicking the target speaker's speaking style when compared to the linear modification method.

1 Introduction

The aim of a *voice conversion* system is to transform the utterance of an arbitrary speaker, referred to as source speaker, to sound as if spoken by a specific speaker, referred to as target speaker. Listeners perceive the source speaker's speech as if uttered by the target speaker. Voice conversion can also be referred to as *voice transformation* or *voice morphing*. Since past two decades voice conversion has been an active research topic in the area of speech synthesis [1], [2], [3], [4]. Applications like text-to-speech (TTS), speech-to-speech translation, mimicry generation and human-machine interaction systems are greatly benefited by having a voice conversion module.

In the literature, majority of voice conversion techniques focused mainly on the modification of short-term spectral features [5], [6]. However, prosodic features, such as pitch contour and speaking rhythm, also contain important cues of speaker identity. In [8] it was shown that pure prosody alone can be used, to an extent, to recognize speakers that are familiar to us. To build a good quality voice conversion system, it needs to modify the prosodic features along with the spectral features. The pitch contour is one of the most important prosodic features related to speaker identity.

The most common method for pitch contour transform is:

$$log(f_0^t) = \frac{log(f_0^s) - \mu^s_{logf_0}}{\sigma^s_{logf_0}} * \sigma^t_{logf_0} + \mu^t_{logf_0} \quad (1)$$

T. Drugman and T. Dutoit (Eds.): NOLISP 2013, LNAI 7911, pp. 97–103, 2013.

where f_0^s, f_0^t represent the pitch values at frame level, and $\mu_{logf_0}^s$, $\sigma_{logf_0}^s$, $\mu_{logf_0}^t$, and $\sigma_{logf_0}^t$ represent the mean and standard deviation of the pitch values in log domain for the source and target speakers, respectively. In this paper, we refer to this method as linear transformation. The local shapes of the pitch contour segments are not modelled and transformed in the linear transformation method. To capture the local dynamics of the pitch contour, we proposed a non-linear transformation method using artificial neural networks (ANNs). The pitch contours over the voiced segments are represented by their discrete cosine transform (DCT) coefficients.

There are some studies which have used the DCT for parametric representation of pitch contour and its modelling [9], [10], [11]. In [9], it is shown that the use of DCT for analysis and synthesis of pitch contours is beneficial. In [10], DCT is used to model the pitch contours of syllables for conversion of neutral speech into expressive speech using Gaussian mixture models (GMM). In [11], DCT representation is used for modelling and transformation of prosodic information in a voice conversion system using a code book generated by classification and regression trees (CART) methods. The work presented in this paper is different from [11] in the following aspects:

1. The proposed method does not use any linguistic information for pitch contour modification.
2. The proposed method uses ANNs to model the non-linear mapping between the pitch contours of source and target speakers.
3. The proposed method, represents the pitch contour of a voiced segment using two sets of parameters. One set represents the statistics, and another set represents the fine variations of a pitch contour.

This paper is organised as follows: Section 2 describes the database, feature extraction and parametrization of the pitch contour. Section 3, outlines the ANN based voice conversion system. The experimental results obtained using both subjective and objective tests are presented in Section 4. Section 5 gives a summary of the work.

2 Database and Feature Extraction

The experiments are carried out on the CMU ARCTIC database consisting of utterances recorded by seven speakers. Each speaker has recorded a set of 1132 phonetically balanced utterances, same for all speakers. ARCTIC database contains the utterances of SLT (US Female), CLB (US Female), BDL (US Male), RMS (US Male), JMK (Canadian Male), AWB (Scottish Male), and KSP (Indian Male).

To extract the features from a given speech signal we used a high quality analysis tool called STRAIGHT vocoder [12]. The features were extracted for every 5ms of speech. Features are: 1) mel-cepstral coefficients (MCEPs), 2)band aperiodicity coefficients (BAPs) and 3) fundamental frequency (pitch contour). All these three features were used for voice conversion. Section 2.1 explains about the parametrization of pitch contour.

2.1 Parametrization of Pitch Contour

The proposed pitch contour model is defined on a voiced segment basis. For voiced speech, the pitch contour varies slowly and continuously over time. It is therefore well modelled by using DCT, an orthogonal transform. One advantage of DCT representation is that the mean square error between two linearly time-aligned pitch contours can be simply estimated from the mean square error between coefficients. The following steps explains the parametrization of a pitch contour:

1. Derive the pitch contours from the utterances spoken by the source speaker.
2. Segment the pitch contour with respect to the voiced segments present in the utterance.
3. Consider only if the duration of each voiced segment is $\geq$ 50ms. If the duration is less than 50ms then use the linear transformation to transform the pitch values.
4. Map the pitch contour of each voiced segment onto equivalent rectangular bandwidth (ERB) scale [7] using Equation 2.

$$F0_{ERB} = \log_{10}(0.00437 * F0 + 1) \tag{2}$$

5. Compute the DCT coefficients for each voiced segment using Equation 3.

$$c_n = \sum_{i=0}^{M-1} F0(i) \cos(\frac{\pi}{M} n(i + \frac{1}{2})) \tag{3}$$

   where pitch contour $F0$ of length M is decomposed into N DCT coefficients $[c_0, c_1, c_2, c_3, ...c_{N-1}]$. The first coefficient represents the mean value and remaining DCT coefficients represents the variations in pitch contour such as those due to syllable stress.
6. Each segment is represented by two sets of parameters. They are

$$F0_{shape} = [c_1, c_2, c_3, ...c_{N-1}] \text{ and } F0_{limits} = [c_0, var_{F0}, max_{F0}, min_{F0}, log(dur)] \tag{4}$$

   Where $F0_{shape}$ and $F0_{limits}$ represents the local variations and the constraints in a pitch contour. $[c_0, c_1, c_2, c_3, ...c_{N-1}]$ are the DCT coefficients and var_{F0}, max_{F0}, min_{F0}, and $log(dur)$ are variance, maximum value, minimum value, and logrithm of duration of a pitch contour, respectively.

3 Voice Conversion Using ANNs

Figure 3, shows the block diagram of both training and transformation process in a voice conversion system. In this work, we used the parallel utterances to build a mapping function between source and target speakers. Even though both speakers speak the same utterances they still differ in the durations. To align

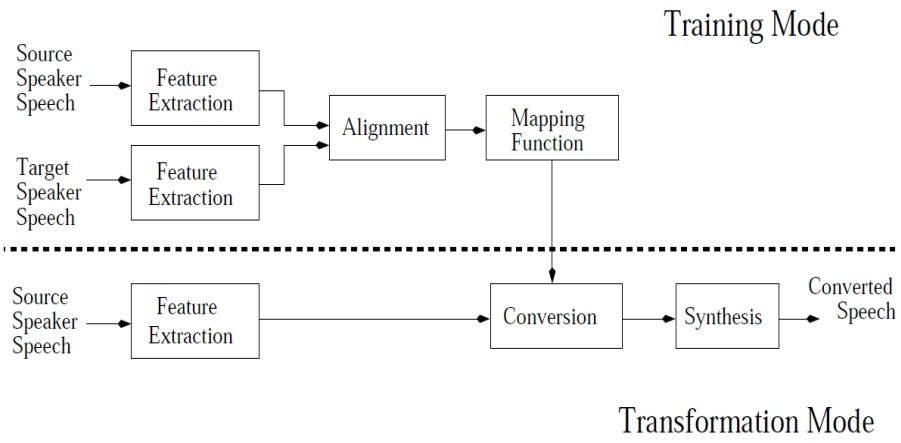

Fig. 1. A block diagram of voice conversion system

the feature vectors of source speaker with respect to target speaker we use the dynamic time warping (DTW) method. It enables us to build a mapping function at frame-level.

For mapping the acoustic features between the source and target speakers, various models have been explored in literature. These models are specific to the kind of features used for mapping. For instance, GMMs [3], vector quantization (VQ) [1] and ANNs [4] are widely used for mapping the vocal tract characteristics. The changes in the vocal tract shape for different speakers are highly non-linear, therefore to model these non-linearities, it is required to capture the non-linear relations present in the patterns. Hence, to capture the non-linear relations between acoustic features, we use a neural network based model (multilayer feed forward neural networks) for mapping the MCEPs, BAPs and pitch contour coefficients.

During the process of training, acoustic features of the source and target speakers are given as input-output to the network. The network learns from these two data sets and tries to capture a non-linear mapping function based on minimum mean square error. A generalized back propagation learning [13] is used to adjust the weights of the neural network so as to minimize the mean squared error between the desired and the actual output values. Selection of initial weights, architectures of ANNs, learning rate, momentum and number of iterations are some of the optimization parameters in training. Once the training is complete, we get a weight matrix that represents the mapping function between the acoustic features of the given source and target speakers. Such a weight matrix can be used to predict acoustic features of the target speaker from acoustic features of the source speaker.

Different network structures can be possible by varying the number of hidden layers and the number of nodes in each of the hidden layer. In [14] it is shown that four layer network is optimal for mapping the vocal tract characteristics of the source speaker to the target speaker. Therefore, we consider the four layer

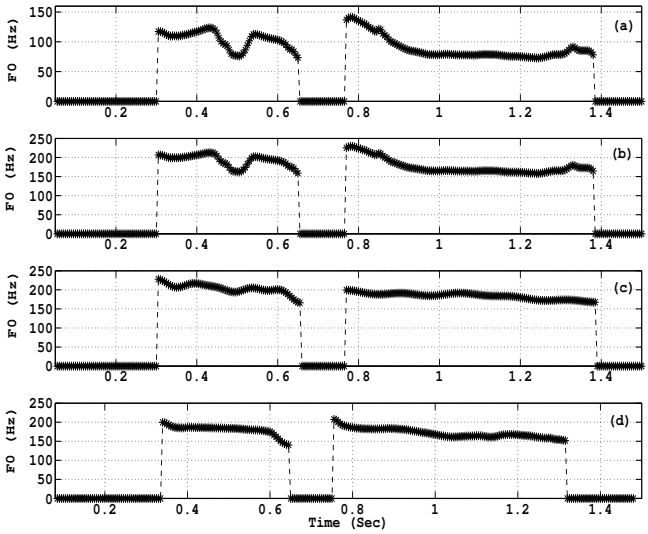

Fig. 2. Conversion of pitch contour from source speaker to target speaker. (a) original source speaker pitch contour, (b) linear modification of source speaker pitch contour, (c) non-linear modification of source speaker pitch contour and (d) original target speaker pitch contour.

networks with architectures $40L - 80N - 80N - 40L$, $21L - 42N - 42N - 21L$, $9L - 18N - 18N - 9L$ and $5L - 10N - 10N - 5L$ for mapping the features MCEPs, BAPs, $F0_{shape}$ and $F0_{limits}$, respectively. The first and fourth layers are input-output layers with linear units (L) and have the same dimension as that of input-output acoustic features. The second layer (first-hidden layer) and third layer (second-hidden layer) have non-linear nodes (N), which help in capturing the non-linear relationship that may exist between the input-output features.

4 Experiments and Results

As described in Section 2, from ARCTIC database we picked one male speaker (RMS) and one female speaker (SLT) for our experiments. For each speaker, we considered 80 parallel utterances for training and a separate set of 32 utterances for testing. We extracted acoustic features, MCEPs of dimension 40, BAPs of dimension 21, and 10 DCT coefficients for every 5ms of speech. Given these features for training, they are aligned using dynamic time warping to obtain paired feature vectors as explained in Section 3. We build a separate mapping function for spectral, band aperiodicity and pitch contour transformations. After the mapping functions are trained, we use the test sentences of the source speaker to predict the acoustic features of the target speaker. The pitch contour is constructed back by using the IDCT on predicted features. An instance of converted

pitch contour from source speaker (RMS) to target speaker (SLT) is illustrated in Figure 3. From Figure 3.(b), we can observe that linear modification of pitch contour is not able to model the local variations of the target speaker, whereas in Figure 3.(c) the non-linear method is able to model the local variations of the target speaker. Please note that here we have used the same durations of the source speaker.

Table 1. RMSE (in Hz) between target and converted contours with linear and non-linear transformation methods

Speaker pair	Linear modification	Non-linear modification
RMS-to-SLT	18.28	14.36
SLT-to-RMS	15.92	12.50

In order to evaluate the performance of the proposed method, we estimate the root mean square error (RMSE) between target and converted pitch contours of test set. The RMSE is calculated after the durations of predicted contours normalized with respect to actual contours of target speaker. It can be seen from Table 1 that the non-linear transformation method performed better than linear method.

Table 2. Speaker similarity score

Speaker pair	Linear modification	Non-linear modification
RMS-to-SLT	3	3.3
SLT-to-RMS	2.55	3.1

An informal perceptual test was also conducted with 10 transformed speech signals randomly chosen for both conversion pair and presented to 10 listeners. We have used the STRAIGHT vocoder to synthesize the transformed speech signals. The subjects were asked to compare similarity of the transformed speech signals with respect to original target speaker speech signals. The ratings were given on a scale of 1-5, with 5 for excellent match and 1 for not-at-all match. The scores are shown in Table 2. It can be observed from Table 2, that non-linear modification performs better than linear modification in perceptual tests as well.

5 Conclusion

A non-linear pitch modification method was proposed for mapping the pitch contours of the source speaker according to the target speaker pitch contours. In this method, pitch contour was compressed to a few coefficients using DCT. A four layer ANN model was used for modelling the non-linear patterns of a pitch contour between the source and target speaker. The results showed that both objective and subjective scores gave very clear preference for the proposed method in mimicking the target speaker's speaking style when compared to the linear modification method.

References

1. Abe, M., Nakamura, S., Shikano, K., Kuwabara, H.: Voice conversion through vector quantization. In: Proc. of ICASSP, New York, USA, pp. 655–658 (April 1988)
2. Stylianou, Y., Cappe, O., Moulines, E.: Continuous probabilistic transform for voice conversion. IEEE Transactions on Speech and Audio Processing 6(2), 131–142 (1998)
3. Ohtani, Y., Toda, T., Saruwatari, H., Shikano, K.: Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation. In: Proc. of INTERSPEECH, Pittsburgh, USA, pp. 2266–2269 (September 2006)
4. Bollepalli, B., Black, A.W., Prahallad, K.: Modeling a noisy-channel for voice conversion using articulatory features. In: Proc. of INTERSPEECH, Portland, USA (August 2012)
5. Dutoit, T., Holzapfel, A., Jottrand, M., Moinet, A., Perez, J., Stylianou, Y.: Towards a voice conversion system based on frame selection. In: Proc. of ICASSP, pp. 513–516 (2007)
6. Stylianou, Y.: Voice transformation: A survey. In: Proc. of ICASSP, pp. 3585–3588 (2009)
7. Smith, J.O., Abel, J.S.: Bark and ERB bilinear transforms. IEEE Transactions on Speech and Audio Processing 7(6), 697–708 (1999)
8. Helander, E., Nurminen, J.: On the importance of pure prosody in the perception of speaker identity. In: Proc. of INTERSPEECH, pp. 2665–2668 (2007)
9. Teutenberg, J., Watson, C., Riddle, P.: Modeling and synthesizing F0 contours with the discrete cosine transform. In: Proc. of ICASSP, pp. 3973–3976 (2008)
10. Veaux, C., Rodet, X.: Intonation conversion from neutral to expressive speech. In: INTERSPEECH, pp. 2765–2768 (2011)
11. Helander, E., Nurminen, J.: A Novel method for prosody prediction in voice conversion. In: Proc. of ICASSP, pp. IV-509–IV-512 (2007)
12. Kawahara, H., Masuda-Katsuse, I., Cheveigne, A.: 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 27, 187–207 (1999)
13. Haykin, S.: Neural networks: A comprehensive foundation. Prentice-Hall Inc., NJ (1999)
14. Desai, S., Black, A.W., Yegnanarayana, B., Prahallad, K.: Spectral mapping using artificial neural networks for voice conversion. IEEE Trans. Audio, Speech and Language Processing 18(5), 954–964 (2010)