Voice Transformation by Mapping the Features at Syllable Level

K. Sreenivasa Rao[1], R.H. Laskar[2], and Shashidhar G. Koolagudi[1]

[1] School of Information Technology, IIT Kharagpur, Kharagpur-721302,
West Bengal, India
ksrao@iitkgp.ac.in, koolagudi@yahoo.com

[2] Department of Electrical Engineering, NIT Silchar, Silchar, Assam, India
laskar_r@nits.ac.in

Abstract. Voice transformation involves modifying the source speaker voice to target speaker voice. Voice characteristics of a speaker depends on the shape of the glottal pulse (source characteristics), shape of the vocal tract system (system characteristics) and the long term features (prosody or supra-segmental) of the speech signal produced by the speaker. In this paper we proposed the mapping functions to transform the vocal tract characteristics and intonation characteristics from source speaker to target speaker. Mapping functions are developed by the features extracted from syllable level. The shape of the vocal tract system is characterized by linear prediction coefficients, and the mapping function is realized by a five layer feedforward neural network. Mapping of the intonation characteristics (pitch contour) is provided by associating the code books derived from the pitch contours of the source and target speakers. The proposed mapping functions are used in voice transformation task. The target speaker's speech is synthesized and evaluated using listening tests. The results of the listening tests indicate that the proposed voice transformation provides better mapping of the voice characteristics compared to the earlier method proposed by the author. The original and the synthesized speech signals obtained using mapping functions are available for listening at *http://shilloi.iitg.ernet.in/∼ksrao/result.html*

1 Introduction

Voice transformation is generally performed in two steps. In the first step, the training stage, a set of speech feature parameters of both the source and target speakers are extracted and appropriate mapping rules that transform the parameters of the source speaker on to those of the target speaker are generated. In the second step, the transformation stage, the features of the source signal are transformed using mapping rules developed in the training stage so that the synthesized speech possesses the personality of the target speaker [2007a].

To implement voice transformation, two problems need to be considered: what features are extracted from the underlying speech signals, and how to modify these features in such a way so that the transformed speech signals mimic target

A. Ghosh, R.K. De, and S.K. Pal (Eds.): PReMI 2007, LNCS 4815, pp. 479–486, 2007.

speakers voices. The first problem can be solved by identifying the speaker-specific features from the given speech signals. It is known that the shape of the glottal pulse, vocal tract transfer function and the prosodic features are uniquely characterize the speaker [2001a]. Feature parameters representing the vocal tract transfer function have been widely used in voice transformation. They include formant frequencies, Linear Prediction Coeffients (LPC), cepstrum and Line Spectrum Pair (LSP) coefficients [1995, 1999a, 1996]. In our previous work, we carried out the voice conversion by modifying the formant frequencies, pitch contour, duration patterns and energy profile by fixed scale factors [2006a]. As a result, the desired speaker characteristics are not much perceived in the synthesized speech. Therefore in this paper we are proposing the mapping functions, which accurately model the relationships between source and target speaker voices.

For mapping the speaker-specific features between source and target speakers, various models have been explored in the literature. These models are specific to the kind of features used for mapping. For instance, Gaussian Mixture Model (GMM) and Vector Quantization (VQ) are widely used for mapping the vocal tract characteristics [2001b, 1998a]. Scatter plots, GMMs and linear models are used for mapping the prosodic features [2003, 1998b]. In this work, we used neural network model for mapping the vocal tract characteristics, and code books for mapping the intonation characteristics. The main reason for exploring neural network models for mapping the vocal tract characteristics is that, it captures the nonlinear relations present in the patterns. The changes in the vocal tract shape corresponds to different speakers is highly nonlinear, therefore to model these nonlinearities, we have chosen neural network model in this work.

For mapping the intonation patterns, there exists different methods with varying complexity. Mapping functions derived by linear, cubic and GMM models fail to predict the local variations at syllable and word levels [2003]. By using the code books consisting of intonation patterns at the utterance level can provide mapping to some extent at global level [2003]. But there is an error between the predicted pitch contour and the target contour with respect to local rise-falls in the intonation patterns. These rise-falls characterize the stress patterns present in the utterance, which are basically depend on the nature of the syllable (i.e., the basic constituents (consonants and vowels) of the syllable), and the linguistic context (nature of the preceding and the following syllables) associated to the syllable.

For Indian context, syllables are the most suitable basic units for the analysis and synthesis of speech. Syllables implicitly capture the duration patterns, shape of the vocal tract and coarticulation effects [2007b]. Eventhough the global characteristics of the intonation patterns depend on the nature of the utterance, but the local variations (rise-falls) depend on the nature of the individual syllables. Therefore for mapping the intonation patterns from source speaker to target speaker, the segments of the pitch contour derived from the syllables are used in this paper. Similarly for mapping the shape of the vocal tract system between the speakers the time-aligned linear prediction coefficients derived from the syllables are used.

The rest of the paper is organized as follows: In section 2, we discuss about the mapping of vocal tract system characteristics using feedforward neural network. Mapping of intonation characteristics between source and target speakers by using code books is discussed in section 3. Synthesis and evaluation of target speaker's voice is given in section 4. The final section contains the summary of the paper, conclusions derived from the work, and some future extensions to this work.

2 Mapping of the Vocal Tract System Characteristics

The basic shape of the vocal tract can be characterized by the gross envelope of the Linear Prediction (LP) spectrum. LP spectrum can be roughly represented by a set of resonant frequencies (formants) and their bandwidths. In our previous work, we used the formants and their bandwidths for characterizing the shape of the vocal tract, and further derived a gross relationship between formant frequencies and average pitch. This method provides the poor estimate of the shape of the vocal tract system for the desired speaker. Therefore in this paper we explored feedforward neural network (FFNN) model for capturing the relationship between the vocal tract system characteristics of the source and target speakers. Here LPCs are used to represent the shape of the vocal tract system. The LPCs are the design parameters of the LP filter, which models the vocal tract system accurately.

For deriving the mapping function using neural network, the network has to be trained with the LPCs derived from the spoken utterances of source and target speakers. For this study, we prepared the text transcription for 100 English sentences. These 100 sentences are recorded by 2 male and 2 female speakers. The duration of the sentence is varying between 3-5 secs. Each sentence has roughly about 15-20 syllables. The recorded speech files are segmented into syllables.

To capture the relationship between the vocal tract shapes of source and target speakers, we need to feed the time-aligned vocal tract features of source and target speakers at the input and output of the neural network respectively. In this work, we used Dynamic Time Warping (DTW) to derive the time-aligned vocal tract features. The procedure for deriving the time-aligned LPCs is given in Table 1. The neural network model used in this work is a five layer feedforward neural network, and it is shown in the Fig. 1.

Here the FFNN model is expected to capture the functional relationship between the input and output feature vectors of the given training data. The mapping function is between the 10-dimensional input vector and the 10-dimensional output. The 10-dimensional input and output vectors correspond to the time-aligned frame LPCs. Several network structures are explored in this study. The (empirically arrived) final structure of the network is $10L$ $20N$ $5N$ $20N$ $10L$, where L denotes a linear unit, and N denotes a nonlinear unit. The integer value indicates the number of units used in that layer. The nonlinear units use $tanh(s)$ as the activation function, where s is the activation value of that unit. The back-propagation learning algorithm is used for adjusting the weights of the network to minimize the mean squared error for each pair of time aligned LPCs [1999b].

Table 1. Steps for deriving the time-aligned LPCs

1. Derive the syllable pair from the utterance pair spoken by the pair of speakers.
2. Preemphasize the speech segments corresponding to the syllable pair.
3. Compute LPCs with 10^{th} order LP analysis, with a frame size of 20 ms and a frame shift of 5 ms.
4. Apply DTW on the syllable pair (with a frame size of 20 ms and a frame shift of 5 ms) and derive the matching frames.
5. Select the LPCs corresponding to the matching frames of the syllable pair and they turned to be pairs of time-aligned LPCs.
6. Same process (steps 1-5) is repeated for other syllable pairs.

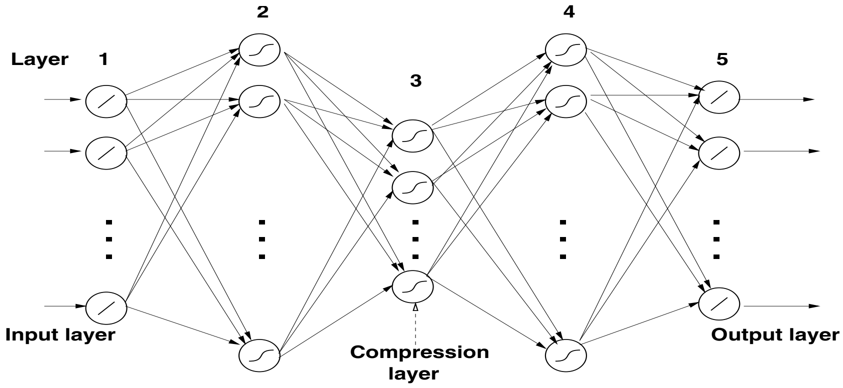

Fig. 1. A five layer FFNN model

A separate model is developed for each pair of the speakers. There are about 76000 matched frame LPCs present in the database, 80% of them are used for training the models. After the training phase, the weights in the network represents the mapping function between input and output. The performance of the model can be evaluated by both subjective and objective measures. Subjective evaluation consists of synthesizing the speech for the desired speaker using the LPCs derived from the neural network model, and conducting the listening tests to asses the desired speaker characteristics present in the synthesized speech. Objective measures consists of spectral distance between the predicted LPCs and actual LPCs. In this paper subjective evaluation is performed, and it is illustrated in Section 4.

3 Mapping of the Intonation Characteristics

For mapping the intonation characteristics between the source and target speakers, code books are prepared using the pitch contours derived from the syllables. For preparing the code books, all the feature vectors should have the same length.

Since the syllables in the database have varying durations, we used resampling technique to obtain the fixed length pitch contour for each syllable. In this study, we explored different lengths (6, 8, 10, 12 and 14), finally the optimum value is found to be 10. The database consists of 1783 syllables, and 90% of them are used for preparing the code books. The code books are prepared separately for source and target speakers using vector quantization [1980]. The entries of the code book represents the mean vectors (centroids) of the clusters formed in the 10-dimensional space. Size of the code book indicates the number of clusters considered for the analysis. In this study we explored different code book sizes (8, 16, 32 and 64), among them 32-size code book shows the better representation. Validation of the code books is performed using the test data.

3.1 Mapping of Code Book Entries

After preparing the code books for source and target speakers, the next step is to derive the mapping function between the entries of the code books. This will be carried out as follows:

1. The resized (10-dimensional) pitch contours of the source speaker which are used for preparing the code book are partitioned according to the entries of the source speaker code book (mean vectors or the cluster centroids). That is the syllable based pitch contours of the source speaker are divided into 32 groups corresponds to the entries of the source speaker code book.

2. For each group of pitch contours of the source speaker, determine the corresponding pitch contours of the target speaker and label them with the same identity as that of the source speaker.

3. The pitch contours of the target speaker belongs to group 1 are projected onto the entries of the target speaker code book, which corresponds to the mean vectors or centroids of the target vector space. The histogram distribution of the projected vectors with respect to the code book entries of the target speaker is determined, and is used to derive the sequence of weights for the sequence of mean vectors present in the target speaker code book.

Now it provides the mapping for the first entry of the source code book as the summation of the weighted code book entries of the target speaker, where the corresponding weights are derived using step 3. Similarly for other entries of the source code book, the corresponding weight vectors are determined. Each weight vector indicates the sequence of weights corresponds to the sequence of code book entries of the target speaker. The overall procedure for the transformation of source speaker pitch contour to target speaker pitch contour is given in Table 2.

4 Synthesis and Analysis of Target Speaker Voice

In the previous sections we described the methods for mapping the vocal tract characteristics and intonation patterns between source and target speakers. In this section, we discuss about the synthesis of target speaker's speech from source speaker's speech by using mapping functions, and then evaluating the presence

Table 2. Steps for transforming the pitch contour

1. Derive the pitch contour from the utterance spoken by the source speaker.
2. Segment the pitch contour with respect to the syllables present in the utterance.
3. For each segment of the pitch contour, determine the closest code book entry of the source speaker.
4. By using the weight vector (sequence of weights corresponds to the sequence of code book entries of the target speaker) corresponding to the code book entry of the source speaker, generate the pitch contour by the summation of the weighted code book entries of the target speaker.
5. Pitch contour for the target speaker at the utterance level is derived by concatenating the syllable level pitch contours derived from step 4.

of desired speaker characteristics in the synthesized speech. The transformation of source speaker speech to target speaker speech is performed as follows:

1. The LPCs representing the vocal tract shape and the pitch contour representing the intonation pattern of the source speaker are derived.

2. The LPCs corresponding to the target speaker are derived from the output of the 5-layer FFNN model, by giving the LPCs of the source speaker speech utterance as input to the FFNN model.

3. The pitch contour of the target speaker is obtained by concatenating the syllable level pitch contours, which are in turn derived from the syllable level pitch contours of the source speaker by using the code books.

Once the pitch contour for the target speaker is available, the pitch contour of the source speaker's speech utterance is replaced by the target speaker's pitch contour. In this work the desired pitch modification is carried out in linear prediction residual domain. The basic reason for choosing the residual domain for modification is that the successive samples in the LP residual are less correlated compared to the samples in the speech signal [2006b]. Therefore the residual manipulation is likely to introduce less distortion in the speech signal synthesized by using the modified LP residual. The details of the pitch modification method are discussed in our previous work [2006a, 2006b]. Finally, target speaker speech is synthesized by exciting the time varying filter representing the target speaker LPCs with the modified LP residual according to the target speaker pitch contour.

The performance of the mapping functions can be evaluated by using subjective and objective measures. In this work the basic goal is the voice transformation, therefore the mapping functions are evaluated by using perceptual tests (i.e., by conducting listening tests). In this work four separate mapping functions are developed for transforming the speaker voices: (1) male to female (M1-F1), (2) female to male (F2-M2), (3) male to male (M1-M2) and (4) female to female (F1-F2). Here M1, M2, F1 and F2 represents male and female speaker voices present in the database. For each case five utterances are synthesized using their associated mapping functions. Listening tests are conducted to assess the desired (target) speaker characteristics present in the synthesized speech. The recorded speech utterances of the target speaker correspond to the synthesized speech utterances are made available to the listeners to judge the relative performance. Twenty students are participated in conducting these tests. Each of the synthesized speech utterance is played to the listener, after playing its original recorded utterance. The subjects were asked to give their opinion score on a 5-point scale. The rating 5 indicates the excellent match between the original target speaker speech and the synthesized speech (i.e., synthesized speech is close to the original speech of the target speaker). The rating 1 indicates very poor match between the original and synthesized utterances, and the other ratings indicate different levels of deviation between 1 and 5. Each listener has to give the opinion scores for each of the five utterances in all the four cases (altogether 20 scores) mentioned above. The mean opinion scores (MOS) for male to female (M1-F1), female to male (F2-M2), male to male (M1-M2) and female to female (F1-F2) transformations are found to be 3.61, 3.94, 3.12 and 2.93, respectively. The obtained MOS indicate that the transformation is effective, if the source and target speakers are from different genders. The basic reason for this variation in MOS with respect to gender is that, the variation in the shapes of the vocal tract and intonation patterns may be large in the case of source and target speakers belongs to different genders. Since the listener is exposed to source and target speakers voices (original) as well as transformed voice with respect to target speaker, he or she may observe wide transformation in the case of male to female or female to male voice conversions compared to the other cases (male to male or female to female). Hence their feeling is reflected in the judgement. While comparing the performance of the voice transformation by the proposed method with the previous work (using formant modification and modification of prosody by a fixed factor), the MOS shows the superiority of the present method. The synthesized speech utterances of the target speaker derived from the proposed method and from the previous work are available for listening at *http://shilloi.iitg.ernet.in/∼ksrao/result.html*

5 Summary and Conclusions

In this paper, a five layer FFNN model was proposed for mapping the vocal tract characteristics between source and target speakers. In the present work LPCs were used for representing the shape of the vocal tract. The final structure of the FFNN model was arrived at empirically. A mapping function between source and target pitch contour code books was derived in the form of weight vectors for transforming the intonation patterns between source and target speakers. The target speaker's speech was synthesized by using the parameters derived from the mapping functions correspond to vocal tract system characteristics and intonation characteristics. Subjective tests were conducted to evaluate the target speaker characteristics present in the synthesized speech. From the perceptual tests, it was found that the voice transformation is more effective, if the source and target speakers belongs to different genders. Subjective evaluation also indicated that the developed voice conversion system has improved, compared to its earlier version proposed by the author.

The performance of each mapping function has to be evaluated separately for analyzing the results of the subjective tests. Investigating the performance of mapping functions using objective measures may give some directions for further improvement in the present system. In this paper, we have not included the mapping functions correspond to source characteristics (shape of the glottal pulse), duration patterns and energy patterns (intensity profile) between souce and target speakers. The overall performance of the voice conversion system may be improved by including the above mapping functions, which are not used in the present study.

References

1. Lee, K.-S.: Statistical approach for voice personality transformation. IEEE Trans. Audio, Speech, and Language processing 15, 641–651 (2007)
2. Yegnanarayana, B., Reddy, K.S., Kishore, S.P.: Source and system features for speaker recognition using AANN models. In: Proc. ICASSP, Salt lake city, Utah, USA, pp. 409–412 (May 2001)
3. Narendranadh, M., Murthy, H.A., Rajendran, S., Yegnanarayana, B.: Transformation of formants for voice conversion using artificial neural networks. Speech Communication 16, 206–216 (1995)
4. Arslan, L.M.: Speaker transformation algorithm using segmental code books (STASC). Speech Communication 28, 211–226 (1999)
5. Lee, K.S., Youn, D.H., Cha, I.W.: A new voice personality transformation based on both linear and non-linear prediction analysis. In: Proc. ICSLP, pp. 1401–1404 (1996)
6. Rao, K.S., Yegnanarayana, B.: Voice conversion by prosody and vocal tract modification. In: Proc. Int. Conf. Information Technology, pp. 111–116 (December 2006)
7. Toda, T., Saruwatari, H., Shikano, K.: Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum. In: Proc. ICASSP, vol. 2, pp. 841–844 (May 2001)
8. Abe, M., Nakanura, S., Shikano, K., Kuwabara, H.: Voice conversion through vector quantization. In: Proc. ICASSP, pp. 655–658 (May 1998)
9. Inanoglu, Z.: Transforming pitch in a voice conversion framework, M.Phil thesis, St.Edmund's College University of Cambridge (July 2003)
10. Stylianou, Y., Cappe, Y., Moulines, E.: Continuous probabilistic transform for voice conversion. IEEE Trans. Speech and Audio Processing 6, 131–142 (1998)
11. Rao, K.S., Yegnanarayana, B.: Modeling durations of syllables using neural networks, Computer Speech and Language, pp. 282–295 (April 2007)
12. Haykin, S.: Neural Networks: A Comprehensive Foundation. Prentice-Hall Inc., New Jersey (1999)
13. Linde, Y., Buzo, A., Gray, R.M.: An algorithm for vector quantizer design. IEEE Trans. Commn. 28(1), 84–95 (1980)
14. Rao, K.S., Yegnanarayana, B.: Prosody modification using instants of significant excitation. IEEE Trans. Audio, Speech and Language Processing 14, 972–980 (2006)