Wavelet-based Voice Morphing

ORPHANIDOU C.,
Oxford Centre for Industrial and Applied Mathematics
Mathematical Institute, University of Oxford
Oxford OX1 3LB, UK
orphanid@maths.ox.ac.uk

MOROZ I.M.
Oxford Centre for Industrial and Applied Mathematics
Mathematical Institute, University of Oxford
Oxford OX1 3LB, UK
moroz@maths.ox.ac.uk

ROBERTS S.J.
Pattern Analysis and Machine Learning Research Group
Information Engineering, University of Oxford
Oxford OX1 3PJ, UK
sjrob@robots.ox.ac.uk

Abstract: - This paper presents a new multi-scale voice morphing algorithm. This algorithm enables a user to transform one person's speech pattern into another person's pattern with distinct characteristics, giving it a new identity, while preserving the original content. The voice morphing algorithm performs the morphing at different subbands by using the theory of wavelets and models the spectral conversion using the theory of Radial Basis Function Neural Networks. The results obtained on the TIMIT speech database demonstrate effective transformation of the speaker identity.

Key-Words: - Voice Morphing, Wavelets, Radial Basis Function Neural Networks.

1 Introduction

Voice morphing technology enables a user to transform one person's speech pattern into another person's pattern with distinct characteristics, giving it a new identity while preserving the original content. Many ongoing projects will benefit from the development of a successful voice morphing technology: text-to-speech (TTS) adaptation with new voices being created at a much lower cost than the currently existing systems; broadcasting applications with appropriate voices being reproduced without the original speaker being present; voice editing applications with undesirable utterances being replaced with the desired ones; internet voice applications with e-mail readers and screen readers for the blind as well as computer and video game applications with game heroes speaking with desired voices.

A complete voice morphing system incorporates a voice conversion algorithm, the necessary tools for pre- and post-processing, as well as analysis and testing. The processing tools include waveform editing, duration scaling as well as other necessary enhancements so that the resulting speech is of the highest quality and is perceived as the target speaker.

In the last few years many papers have addressed the issue of voice morphing using different signal processing techniques [1-6]. Most methods developed were single-scale methods based on the interpolation of speech parameters and modeling of the speech signals using formant frequencies [7], Linear Prediction Coding Cepstrum coefficients [9], Line Spectral Frequencies [8] and harmonic-plus-noise model parameters [2]. Other methods are based on mixed time- and frequency-domain methods to alter the pitch, duration and spectral features. The methods suffer from absence of detailed information during the extraction of formant coefficients and the excitation signal which results in the limitation on accurate estimation of parameters as well as distortion caused during synthesis of target speech.

Wavelet methods have been used extensively in many areas of signal processing in order to discover informative representations of non-stationary signals. Wavelets are able to encode the approximate shape of a signal in terms of its inner products with a set of dilated and translated wavelet atoms and have been used extensively in speech analysis [9-10] but not in the area of voice morphing. Turk and Arslan [12] introduced the Discrete Wavelet Transform to voice conversion and got some encouraging results.

2. Main Features of proposed method

Our proposed model uses the theory of wavelets as a means of extracting the speech features followed by Radial Basis Function Neural Networks (RBFNN) for modeling the conversion.

2.1 Wavelet Decomposition

Subband decomposition is implemented using the Discrete Wavelet Transform (DWT).

Wavelets are a class of functions that possess compact support and form a basis for all finite energy signals. They are able to capture the non-stationary spectral characteristics of a signal by decomposing it over a set of atoms which are localized in both time and frequency. The DWT uses the set of dyadic scales and translates of the mother wavelet to form an orthonormal basis for signal analysis.

In wavelet decomposition of a signal, the signal is split using high-pass and low-pass filters into an approximation and a detail. The approximation is then itself split again into an approximation and a detail. This process is repeated until no further splitting is possible or until a specified level is reached. Fig. 1 shows a diagram of a wavelet decomposition tree [13]. The DWT provides a good signal processing tool as it guarantees perfect reconstruction and prevents aliasing when appropriate filter pairs are used.

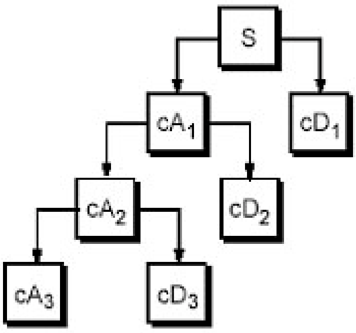

Fig. 1: Example of a wavelet decomposition tree. The original signal *S* is split into an approximation cA_1 and a detail cD_1. The approximation is then itself split into an approximation and a detail and so on.

Decomposing a signal into *k* levels of decomposition therefore results in *k+1* sets of coefficients at different frequency resolutions, *k* levels of detail and 1 level of approximation coefficients.

2.2 Radial Basis Function Neural Networks

A Radial Basis Function Neural Network (RBFNN) can be considered as a mapping $\Re^d \rightarrow \Re^c$ where *d* is the dimension of the input space and *c* the dimension of the output space. For a *d*-dimensional input vector x, the basic form of the mapping is

$$y_k(x) = \sum_{j=1}^{M} w_{kj}\phi_j(x) + w_{k0} \qquad (1)$$

where ϕ_j is the j^{th} radial basis function (RBF) and w_{k0} is the bias term which can be absorbed into the summation by including an extra RBF ϕ_0 whose activation is set to 1 and ϕ_j is usually of the form

$$\phi_j(x) = \exp\left[-\frac{1}{2}(x-\mu_j)^T \Sigma_j (x-\mu_j)\right] \qquad (2)$$

where μ_j is the vector determining its centre and Σ_j is the covariance matrix associated to σ_j, the width of the basis function [14].

The learning process of the RBFNN is done in two different stages. The first stage of the process is unsupervised as the input data set $\{x^n\}$ alone is used to determine the parameters of the basis functions. The selection is done by estimating a Gaussian Mixture Model (GMM) probability distribution from the input data using Expectation-Maximization [16]. The widths of the RBFs are calculated from the mean distance of each centre to its closest neighbor. The second stage of the learning process is supervised as both input and output data are required. Optimization is done by a classic least squares approach. Considering the RBFNN mapping defined in (1) (and absorbing the bias parameter into the weights) we now have

$$y_k(x) = \sum_{j=0}^{M} w_{kj}\phi_j(x) \qquad (4)$$

where ϕ_0 is an extra RBF with activation value fixed at 1. Writing this in matrix notation

$$y(x) = W\varphi \qquad (5)$$

where $W = (w_{kj})$ and $\varphi = (\varphi_j)$. The weights are then optimized by minimization of the sum-of-squares error function

$$E = \frac{1}{2}\sum_n \sum_k \{y_k(x^n) - t_k^n\}^2 \qquad (6)$$

where t_k^n is the target value for output unit k when the network is presented with the input vector x^n . The weights are then determined by the linear equations [14]

$$\Phi^T \Phi W^T = \Phi^T T \qquad (7)$$

where $(T)_{nk} = t_k^n$ and $(\Phi)_{nj} = \phi_j(x^n)$. This can be solved by

$$W^T = \Phi^{\dagger} T \qquad (8)$$

where $\Phi^{\dagger}$ denotes the *pseudo-inverse* of Φ. The second layer weights are then found by fast, linear matrix inversion techniques [14].

3 Proposed model

Voice morphing is performed in two steps: training and transformation. The training data consist of repetitions of the same phonemes uttered by both source and target speakers. The utterances are phonetically rich (i.e. the frequency of occurrence of different phonemes is proportionate to their frequency of occurrence in the english language) and are normalized to zero mean and unit variance. The source and target training data is divided into frames of 128 samples and the data is randomly divided into training and validation sets. A 5-level wavelet decomposition is then performed to the source and target training data. The wavelet basis used is the one which gives the lowest *reconstruction error* given by

$$E_R = \sqrt{\frac{\sum_{m=1}^{M}(y(m) - y^*(m))^2}{\sum_{m=1}^{M} y(m)^2}} \qquad (9)$$

where M is the number of points in the speech signal, $m = 1, \cdots, M$ is the index of each sample, $y(m)$ is the original signal and $y^*(m)$ is the signal reconstructed from the wavelet coefficients using the Inverse Discrete Wavelet Transform (IDWT). The wavelet basis used was the Coiflet 5 for male-to-female and male-to-male speakers morphing and the Biorthogonal 6.8 for female-to-male and female-to-female speakers morphing as they gave the smallest reconstruction error.

In the transformation stage, the wavelet coefficients are calculated at different subbands. In order to reduce the number of parameters used and reduce the complexity of the RBFNN mapping, the coefficients at the two levels of highest frequencies are set to zero. At each of the remaining 4 levels, the wavelet coefficients are normalized to zero mean and unit variance (by using the coefficients' statistics) and a mapping is learned using the RBFNN model using frames of coefficients as input vectors. The best RBFNN on each level is chosen by minimizing the error on the validation data. The complexity of the RBFNN depends on the size of the training data which varies depending on the frequency of occurrence of each phoneme in the available speech corpus. More than 20 RBF centers were rarely needed in order to optimize the network. Test data from the source speaker are then pre-processed and split into the same number of subbands as the training data and the wavelet coefficients are projected through the trained network in order to produce the morphed wavelet coefficients. The morphed coefficients are then un-normalized i.e. they are given the statistics (mean and variance) of the wavelet coefficients of the target speaker and then used to reconstruct the target speaker's speech signal. Post-processing again includes amplitude editing so that the morphed signal has the statistics of the target speaker. The process is repeated for all different phonemes of interest which are then put together in order to create the desired text. Fig. 2 shows a diagram of the proposed model.

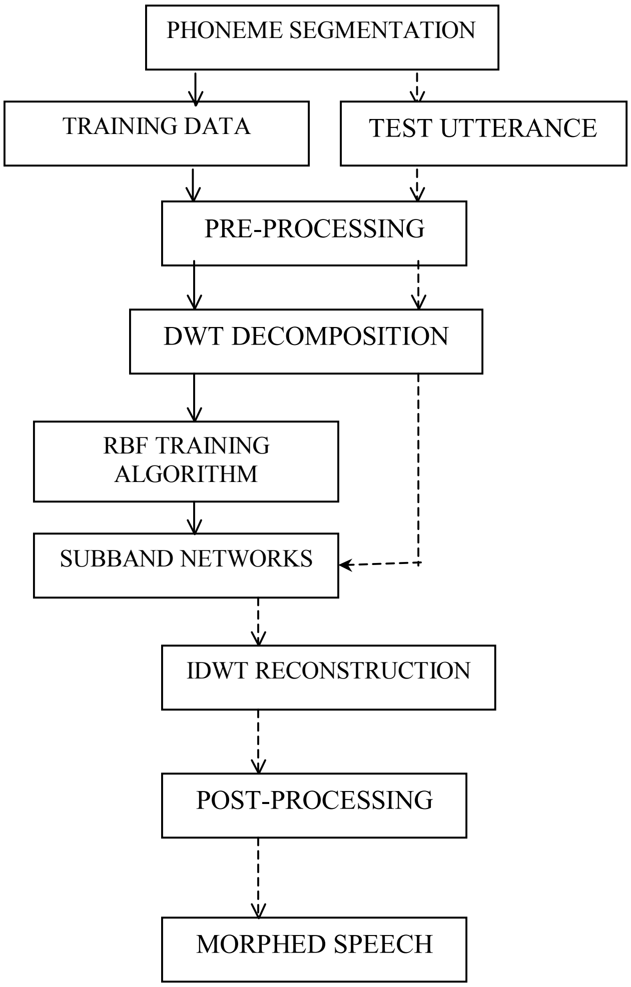

Fig 2. Proposed model

4 Results and Evaluation

Figures 3, 4, 5 and 6 show some results of our morphing on speech from male-to-male, female-to-female, female-to-male, and male-to-female pairs of speakers. The sentences uttered are taken from the TIMIT database [15]. In order to evaluate the performance of our system in terms of it perceptual effects an ABX-style preference test was performed, which is common practice for voice morphing evaluation tests [3, 5]. Independent listeners were asked to judge whether an utterance X sounded closer to utterance A or B in terms of speaker identity, where X was the converted speech and A and B were the source and target speech, respectively. The ABX-style test performed is a variation of the standard ABX test since the sound X is not actually spoken by either speaker A or B, it is a new sound and the listeners need to identify which of the two sounds it *resembles*. Also, utterances A and B were presented to the listeners in random order. In total, 16 utterances were tested which consisted of 4 male-to-male, 4 female-to-female, 4 female-to-male and 4 male-to-female source-target combinations. All utterances were taken from the TIMIT database. 11 independent listeners took part in the testing. Each listener was presented with the 16 different triads of sounds (source, target and converted speech, the first two in random order) and had only one chance of deciding whether sound X sounds like A or B. It is assumed that there is no correlation between the decisions made by the same person and that all 176 resulting decisions are independent. Since the probability of a listener recognizing the morphed speech as the target speaker is 0.5, the results were verified statistically by testing the null hypothesis that the probability of recognizing the target speaker is 0.5 versus the alternative hypothesis that the probability is greater than 0.5. The measure of interest is the *p*-value associated with the test i.e. probability that the observed results would be obtained if the null hypothesis was true i.e. if the probability of recognizing the target speaker was 0.5. Table 1 gives the percentage of the converted utterances that were labeled as closer to the target speaker as well as the p-values.

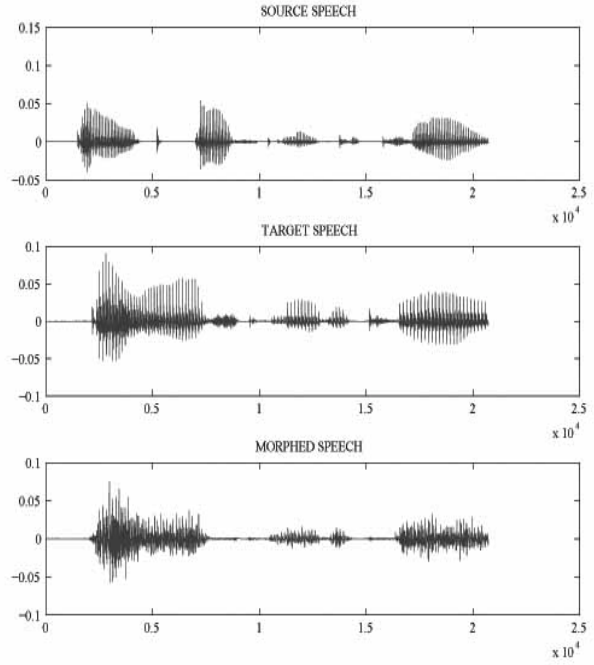

Fig 3. Source, Target and Morphed sentence waveform for a male-to-male speaker speech transformation of the utterance “Don’t ask me to carry”.

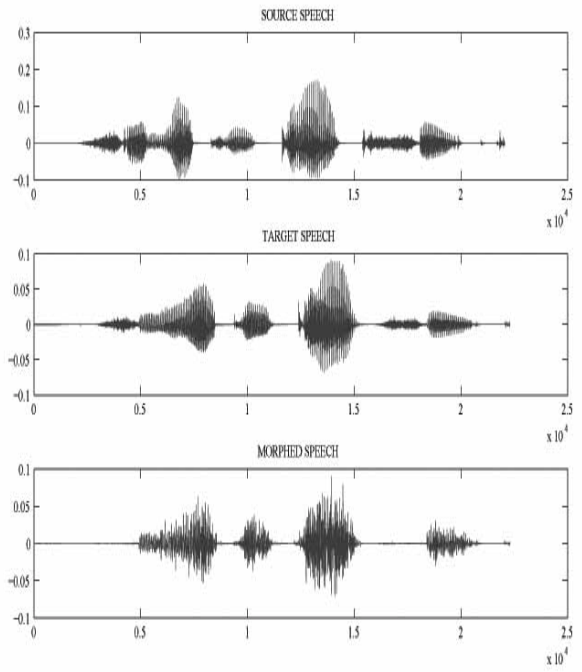

Fig 4. Source, Target and Morphed sentence waveform for a female-to-female speaker speech transformation of the utterance “Don’t ask me to carry”.

Source-Target	% success	p-value
Male-to-Male	79.5	0.0001
Female-to Female	77.3	0.0003
Male-to-Female	86.3	0.00004
Female-to-Male	88.6	0.00001

Table 1. Percentage of “successful” labeling and associated *p*-values.

The *p*-values obtained are considered statistically insignificant, it is therefore evident that the null hypothesis is rejected and the alternative hypothesis is valid i.e. the converted speech is successfully recognized as the target speaker.

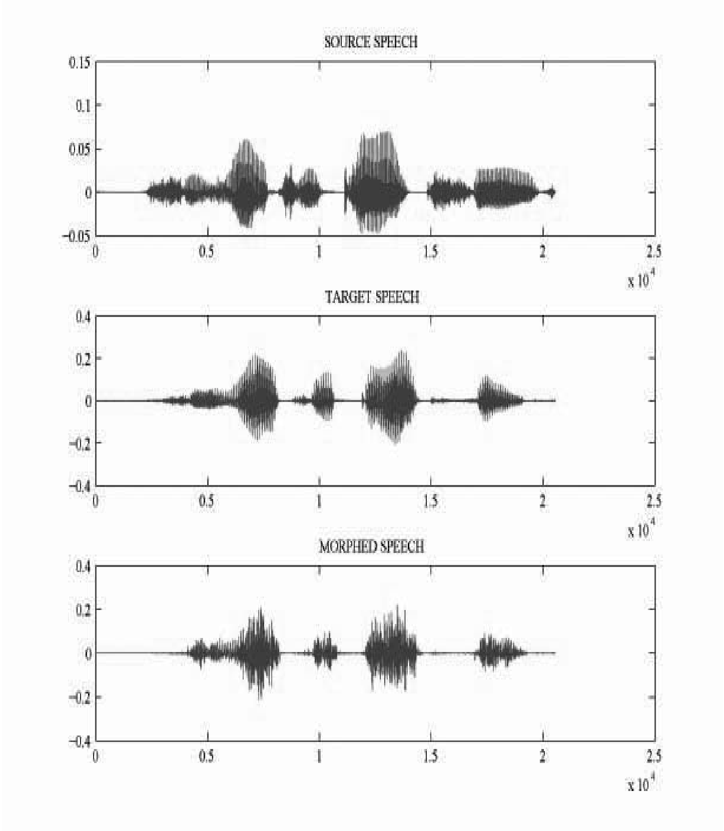

Fig 5: Source, Target and Morphed sentence waveform for a female-to-male speaker speech transformation of the utterance "She had your dark suit".

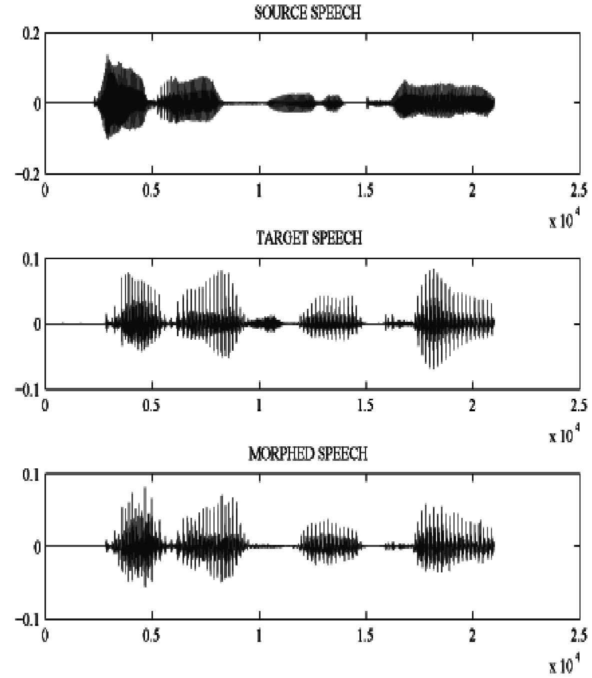

Fig 6. Source, Target and Morphed sentence waveform for a male-to-female speaker speech transformation of the utterance "She had your dark suit".

5 Conclusion

A voice morphing system was presented which extracts the voice characteristics by means of wavelet decomposition and then uses the theory of RBFNN for morphing at the different levels of decomposition. The experimental results show that the conversion is successful in terms of producing speech that can be recognized as the target speaker although the speech signals sounded muffled. Furthermore, it was observed that most distortion occurred at the unvoiced parts of the signals. The muffled effect as well as the distortion could be due to the removal of some of the highest frequency components of the speech signals therefore methods of including all frequency levels to the morphing will be a subject of future work.

References:

[1] Drioli C., Radial basis function networks for conversion of sound speech spectra, *EURASIP Journal on Applied Signal Processing*, Vol. 2001, No.1, 2001, pp. 36-40.

[2] Valbret H. , Voice transformation using PSOLA technique , *Speech Communication,* Vol .11, No 2-3, 1992, pp. 175-187.

[3] Arslan L., Speaker transformation algorithm using segmental codebooks, *Speech Communication,* No. 28, 1999, pp. 211-226.

[4] Arslan L. and Talkin D, Voice Conversion by Codebook Mapping of Line Spectral Frequencies and Excitation Spectrum, *Proceedings of Eurospeech*, 1997, pp. 1347-1350.

[5] Stylianou Y., Cappe O. and Moulines E., Statistical Methods for Voice Quality Transformation, *Proceedings of Eurospeech*, 1995, pp. 447-450.

[6] Orphanidou C., Moroz I.M and Roberts S.J. Voice Morphing Using the Generative Topographic Mapping, Proceedings of CCCT '03, Vol. I, pp. 222-225.

[7]Abe M., Nakamura S., Shikano K. and Kuwabara H., Voice conversion through vector quantization, *Proceedings of the ICASSP*, 1988, pp. 655-658.

[8] Kain A. and Macon M., Spectral Voice Conversion for text-to-speech synthesis, *Proceedings of the IEEE ICASSP*, 1998, pp. 285-288.

[9] Furui S., Research on individuality features in speech waves and automatic speaker recognition techniques, *Speech Communication*, Vol. 5, No. 2, 1986, pp. 183-197.

[10] Barros A.K., Rutkowski T., Itakura F., Estimation of speech embedded in a reverberant and noisy environment by independent component analysis and wavelets, IEEE Transactions in Neural Networks, Vol. 13, No. 4, 2002, pp. 888-893.

[11]Ercelebi E., Second generation wavelet transform-based pitch period estimation and voiced/unvoiced decision for speech signals. *Applied Acoustics*, Vol. 64, No. 1, 2003, pp 25-41.

[12] Turk O., Arslan L.M., Subband based voice conversion, Proceedings ICSLP, 2002, pp. 289-293.

[13] Misiti M., Misiti Y., Oppenheim G., Poggi J. M., *Wavelet Toolbox User's Guide*, The MathWorks, 1997.

[14] Bishop C. M., *Neural Networks for Pattern Recognition*, Clarendon Press, 1995.

[15] Garofolo J.S., Lamel L.F., Fisher W.M., Fiscus J.G., Pallet D.S., Dahlgren N.L., *TIMIT Acoustic-Phonetic Continuous Speech Corpus*, US Department of Commerce, 1993.

[16] Dempster A.P., Laird N.M. and Rubin D.B. *Maximum likelihood from incomplete data via the EM algorithm*. J.R. Statistical Society, Vol. 39(B), pp. 1-38, 1977.