1. Introduction

Voice morphing is the transition of one speech signal into another. Voice morphing technology enables a user to transform one person's speech pattern into another person's pattern with distinct characteristics. Image morphing, speech morphing preserve the shared characteristics of the starting and final signals, while generating a smooth transition between them. Speech morphing is analogous to image morphing. In image morphing the in-between images show one face smoothly changing its shape and texture until it turns into the target face. In Speech morphing one speech signal should smoothly change into another, keeping the shared characteristics of the starting and ending signals but smoothly changing the other properties. Pitch and formant information in each signal is extracted using the cepstral approach. Necessary processing to obtain the morphed speech signal include methods like Cross fading of envelope information, Dynamic Time Warping to match the major signal features (pitch) and Signal Re-estimation to convert the morphed speech signal back into the acoustic waveform.

A successful procedure for voice morphing requires a representation of the speech signal in a parametric space, using a suitable mathematical model that allows interpolation between the characteristics of the two speakers. In other words, for the speech characteristics of the source speaker's voice to change gradually to those of the target speaker, the pitch, duration and spectral parameters must be extracted from both speakers. Natural-sounding synthetic intermediates, with a new voice timbre, can then be produced. A key element in the morphing is the manipulation of the pitch information. If two signals with different pitches were simply cross-faded it is highly likely that two separate sounds will be heard. 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.

Figure below shows the general block diagram of the voice morphing system. A recognition and alignment module was added for synchronizing the user's voice with the target voice before the morphing is done. Before we can morph a particular file we have to supply information about the file to be morphed and the file recording itself (Target

Information and File Information). The system requires the phonetic transcription of the lyrics, the melody as MIDI data, and the actual recording to be used as the target audio data. Thus, a good impersonator of the person that originally spoke the speech has to be recorded. This recording has to be analyzed with SMS, segmented into "morphing units", and each unit labeled with the appropriate note and phonetic information of the file. This preparation stage is done semi-automatically, using a non-real time application developed for this task. The first module of the running system includes the real-time analysis and the recognition/alignment steps. Each analysis frame, with the appropriate parameterization, is associated with the phoneme of a specific moment of the song and thus with a target frame. Once a user frame is matched with a target frame, we morph those interpolating data from both frames and we synthesize the output sound. Only voiced phonemes are morphed and the user has control over which and by how much each parameter is interpolated. The frames belonging to unvoiced phonemes are left untouched thus always having the user's consonants.

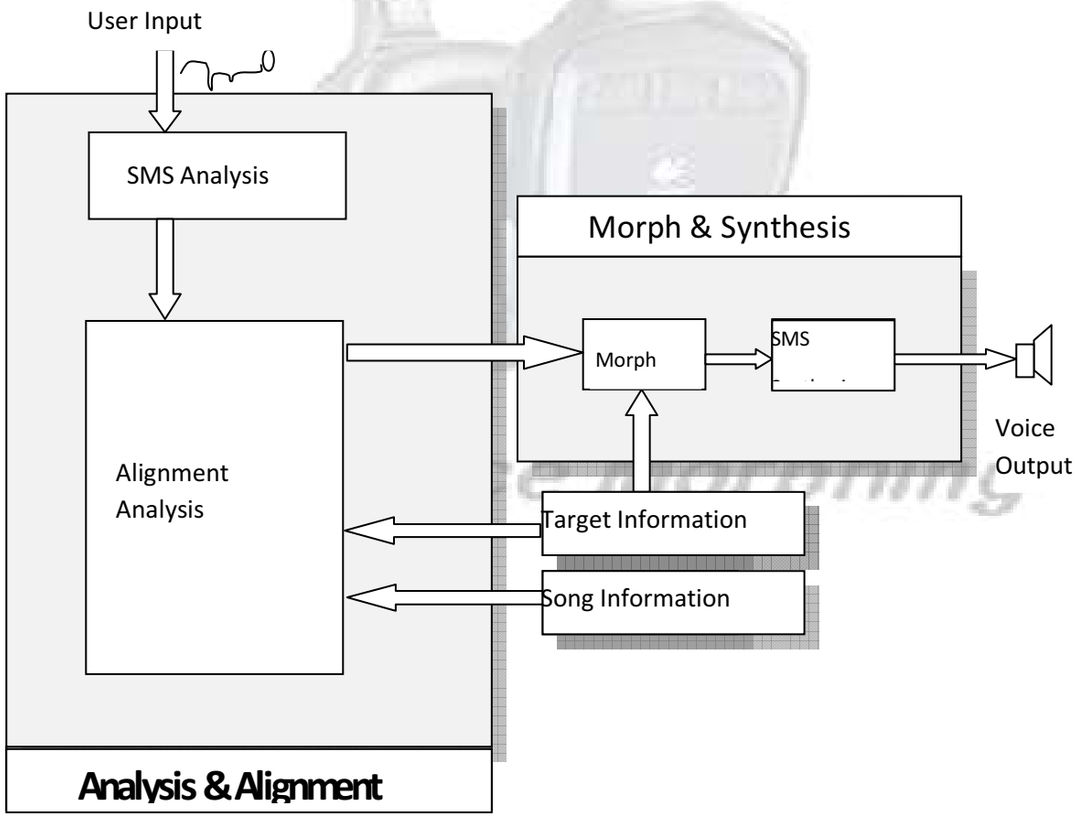

Fig 1.1: Block Diagram of Voice Morphing System

This report has been subdivided into nine chapters. The second chapter gives the details about history of voice morphing. The third chapter gives an idea of the various processes involved in voice morphing in a concise manner. A thorough analysis of the procedure used to accomplish morphing and the necessary theory involved is presented in an uncomplicated manner in the fourth chapter. Processes like pre processing, cepstral analysis, dynamic time warping and signal re-estimation are vividly described with necessary diagrams. The fifth chapter gives a deep insight into the actual morphing process. The conversion of the morphed signal into an acoustic waveform is dealt in detail in the sixth chapter. Chapter six summarizes the whole morphing process with the help of a block diagram. Chapter seven lists the conclusions that have been drawn from this project.

2. History

Many voice morphing techniques have been proposed since the original formulation of the voice morphing problem by Childers et al in 1985. Childers proposed solution involved a mapping of acoustical features from a source speaker to a target speaker. A year later, Shikano proposes to use vector quantization (VQ) techniques and codebook sentences. A few years later, in 1990, Abe introduced the idea of cross lingual voice conversion systems (CVCS) using bilingual subjects, and in 1991, Velbert rekindles the discussion by proposing personalized Text-to-Speech systems using the idea of Dynamic Frequency Warping (DFW). In 1995, Childers introduced the idea of Voice morphing based on the physiological model of glottal pulse and vocal tract, and Narendranath added Artificial Neural Networks (ANN) to the list of Voice Morphing techniques. By the end of the 90's, Arslan proposed a model using Line Spectral Frequencies for spectral envelope representation, which resulted in the STASC (Speaker Transformation Algorithm using Segmental Codebooks) algorithm, and Stylianou used Gaussian Mixture Models (GMMs) combined with Mel-Frequency Cepstral Coefficients (MFCCs) as an alternative to spectral envelope representation. In 2001, Toda proposed a combined spectral representation and voice conversion technique named STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weighted spectrum), which allowed the manipulation of spectral, acoustical and rhythmic parameters. A year later Turk proposed a variation of Arslan's STASC algorithm using the Discrete Wavelet Transform (DWT). Sundermann made a series of contributions since 2003. He has established the concept of text-independent voice conversion and has been the first to propose a text independent cross lingual voice conversion system that did not require bilingual subjects for training the system. He also brought up to the field of voice conversion a technique known as Vocal Tract Length Normalization (VTLN), which had been originally proposed in 1995 by Kamm etal in the context of speech recognition. More recent contributions by Rentzos, ye, Rao and Zhang have focused in probabilistic techniques, such as GMMs, codebook sentences, and technique such as ANN and DFW, among others.

3. Classification of Voice Morphing techniques

Voice morphing techniques may be classified according to the acoustical features used in the alternative representation of the signals, as well as according to the transformation techniques employed in conversion.

3.1 Representation Models

There are a few parameters that are usually computed for each frame, such as pitch (F0), energy (rms), and some representation of frequency content, which is fundamental both for classification and transformation of voice quality. Besides the Fourier spectrum and its envelope, voice morphing systems use many other representation models for a voice signal, such as:

Voice-based models: Vocal Tract Length Normalization (VTLN), Formant Frequencies, and Glottal Flow models.

Voice-based models are based on representations of human voice-producing mechanisms, using concepts such as glottal pulse, which is the raw signal produced by vocal folds, and vocal tract, which comprises the oral and nasal cavities, palate, tongue, jaw and lips, and is responsible for many timbral voice qualities.

Mixed Voice/Signal Models: Linear Prediction Coding (LPC), Line Spectral Frequencies (LSF), Cepstral Coefficients, and STRAIGHT (Speech Transformation & Representation using Adaptative Interpolation of weiGHTed spectrum).

Mixed voice/signal models are actually signal models that provide compact representations for the signals. Since they are largely used by the speech recognition community, they acquired many helpful voice-related interpretations. For instance, parts of the cepstrum are often related to formant regions (and thus to vocal tract contribution) or to the fundamental frequency, and LPC coefficients and LPC residuals can also sometimes be associated to vocal tract and glottal pulse (viewed in a subtractive synthesis context).

Signal-based models: Improved Power Spectrum Envelope (IPSE), Discrete Wavelet Transform (DWT), Harmonic plus Noise Model (HNM).

Purely signal-based models are based on general time-domain and frequency-domain signal representations, and are usually devoid of specific voice-related or phonetic related semantics.

Harmonic-pulse-noise model: The harmonic-plus-noise model is more specific than the others, but is especially useful in tracking voiced portions (e.g., vowels) of the signal.

Besides the usual linear frequency spacing that is common to Fourier-based methods, many of these techniques also allow the use of alternative frequency scales such as *BARK* and *MEL*.

3.2 Transformation Techniques

The transformation phase in voice morphing systems is concerned with every acoustic feature used in the representation of the voice signal. This includes pitch shifting and energy compensation, but also the transformation of frequency content in such a way that both timbral aspects and intelligibility are taken into account. Transformation techniques are intrinsically tied to representation models. Some of the common techniques are:

Statistical Techniques: Gaussian Mixture Models (GMM), Hidden Markov Models (HMM), MultiSpace Probability Distributions, Maximum Likelihood Estimators (MLE), Principal Component Analysis (PCA), Unit Selection (US), Frame Selection(FS), K-means, K-histograms. Cognitive Techniques: Artificial Neural Networks (ANN), Radial Basis Function Neural Networks (RBFNN), Classification and Regression Trees (CART), Topological Feature Mapping, and Generative Topographic Mapping.

Statistical techniques usually assume that data such as feature vectors or vocal parameters have a random component and may be reasonably described by means and standard deviations (Gaussian model), or that they evolve over time according to simple rules based on the recent past (Markov models).

Linear Algebra Techniques: Bilinear Models, Singular Value Decomposition (SVD), Weighted Linear Interpolations (WLI) and Perceptually Weighted Linear Transformations, and Linear Regression (LRE, LMR, MLLR).

Linear algebra techniques are based on geometrical interpretations of data, for instance in finding simplified models by orthogonal projection (linear regression), in obtaining convex combinations of input data (weighted interpolations), or in decomposing transformations into orthogonal components (SVD).

Signal Processing Techniques: Vector Quantization (VQ) and Codebook Sentences, Speaker Transformation Algorithm using Segmental Codebooks (STASC), and Frequency Warping (FW, DFW, and WFW).

Signal processing techniques define transformations based on time-domain or frequency-domain representations of the signal. These may try to encode a signal using a library of frequently found signal segments or code words, or to convert timbre-related voice qualities by modifying frequency scale representations (warping).

These techniques basically differ with respect to the way they look at data. For instance, Cognitive techniques are based on learning processes using abstract neuronal structures, and usually depend on a training phase (where both inputs and outputs are available). Frequently they are used for decision problems (where only 2 possible output values are available), for instance in speech recognition, where a separate network is trained for every specific phoneme or word or sentence that is going to be recognized.

4. An Introspection of the Morphing Process

Processes like cepstral analysis and the re-estimation of the morphed speech signal into an acoustic waveform involve much intricacy and challenge. This seminar digs deep into the basics of digital signal processing.

Voice morphing can be achieved by transforming the signal's representation from the acoustic waveform obtained by sampling of the analog signal, to another representation. To prepare the signal for the transformation, it is split into a number of 'frames' known as sections of the waveform. The transformation is then applied to each frame of the signal. This provides another way of viewing the signal information. The new representation which is also known as the frequency domain, describes the average energy present at each frequency band.

Further analysis enables two pieces of information to be obtained: pitch information and the overall envelope of the sound. As we know a key element in the morphing is the manipulation of the pitch information. If two signals with different pitches are simply cross-faded it is highly recommended that two separate sounds will be heard. This occurs because the signal will have two distinct pitches causing the auditory system to perceive two different objects. A successful morph must exhibit a smoothly changing pitch throughout. The pitch information of each sound is compared to provide the best match between the two signals' pitches. To do this match, the signals are stretched and compressed so that important sections of each signal match in time. The interpolation of the two sounds can then be performed which creates the intermediate sounds in the morph. The final stage is then to convert the frames back into a normal waveform.

After the morphing has been performed, the legacy of the earlier analysis becomes apparent. The conversion of the sound to a representation in which the pitch and spectral envelope can be separated loses some information. Therefore, this information has to be re-estimated for the morphed sound. This process obtains an acoustic waveform, which can then be stored or listened to. Figure below is the graphical presentation of morphing process.

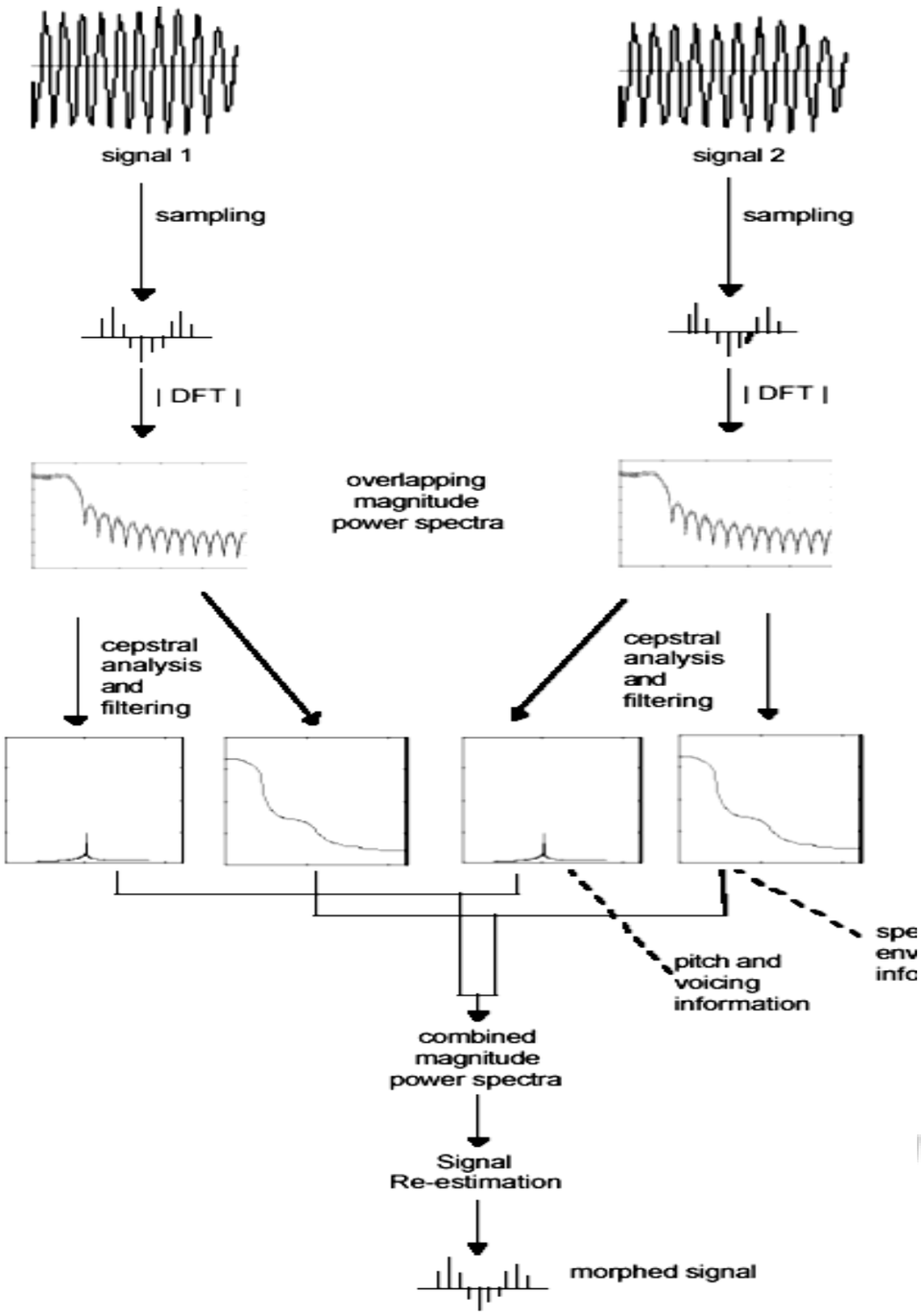

Fig 4.1: Schematic block diagram of the speech morphing process

5. Morphing Process: A Comprehensive Analysis

The algorithm to be used is shown in the simplified block diagram placed below. The algorithm contains a number of fundamental signal processing methods including sampling, the discrete Fourier transform and its inverse, cepstral analysis. However the main processes can be categorized as follows.

I. Preprocessing or representation conversion: This involves processes like signal acquisition in discrete form and windowing.

II. Cepstral analysis or Pitch and Envelope analysis: This process extracts the pitch and formants information in the speech signal.

III. Morphing which includes Warping and interpolation.

IV. Signal re-estimation.

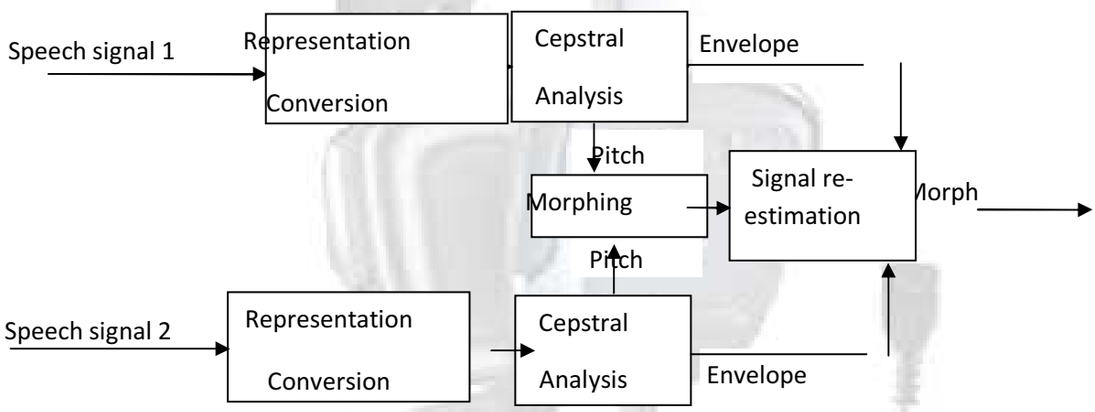

Fig: 5.1 Block diagram of the simplified speech morphing algorithm

5.1 Acoustics of speech production

Speech production can be viewed as a filtering operation in which a sound source excites a vocal tract filter. The source may be periodic, resulting in voiced speech, or noisy and a periodic, causing unvoiced speech. As a periodic signal, voiced speech has a spectra consisting of harmonics of the fundamental frequency of the vocal cord vibration; this frequency often abbreviated as F0, is the physical aspect of the speech signal corresponding to the perceived pitch. Thus pitch refers to the fundamental frequency of the vocal cord vibrations or the resulting periodicity in the speech signal. This F0 can be determined either

from the periodicity in the time domain or from the regularly spaced harmonics in the frequency domain.

The vocal tract can be modeled as an acoustic tube with resonances, called formants, and anti resonances. Moving certain structures in the vocal tract alters the shape of the acoustic tube, which in turn changes its frequency response. The filter amplifies energy at and near formant frequencies, while attenuating energy around anti resonant frequencies between the formants.

The common method used to extract pitch and formant frequencies is the spectral analysis. This method views speech as the output of a liner, time-varying system (vocal tract) excited by either quasiperiodic pulses or random noise. Since the speech signal is the result of convolving excitation and vocal tract sample response, separating or "de-convolving" the two components can be used. In general, de-convolution of the two signals is impossible, but it works for speech, because the two signals have quite different spectral characteristics. The de-convolution process transforms a product of two signals into a sum of two signals. If the resulting summed signals are sufficiently different spectrally, they may be separated by linear filtering. Now we present a comprehensive analysis of each of the processes involved in morphing with the aid of block diagrams wherever necessary.

5.2 Preprocessing

This section shall introduce the major concepts associated with processing a speech signal and transforming it to the new required representation to affect the morph. This process takes place for each of the signals involved with the morph.

5.2.1 Signal Acquisition

Before any processing can begin, the sound signal that is created by some real-world process has to be ported to the computer by some method. This is called sampling. A fundamental aspect of a digital signal (in this case sound) is that it is based on processing sequences of samples. When a natural process, such as a musical instrument, produces sound the signal produced is analog (continuous-time) because it is defined along a continuum of times. A discrete-time signal is represented by a sequence of numbers - the signal is only defined at discrete times. A digital signal is a special instance of a discrete-time signal - both time and amplitude are discrete. Each discrete representation of the signal is termed a sample.

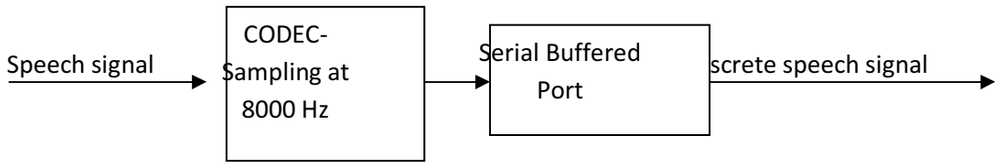

Fig 5.2: Signal acquisitions

The input speech signals are taken using MIC and CODEC. The analog speech signal is converted into the discrete form by the inbuilt CODEC TLC320AD535 present onboard and stored in the processor memory. This completes the signal acquisition phase.

5.2.2 Windowing

A DFT (Discrete Fourier Transformation) can only deal with a finite amount of information. Therefore, a long signal must be split up into a number of segments. These are called frames. Generally, speech signals are constantly changing and so the aim is to make the frame short enough to make the segment almost stationary and yet long enough to resolve consecutive pitch harmonics. Therefore, the length of such frames tends to be in the region of 25 to 75 milliseconds. There are a number of possible windows. A selection is:

The Hanning window

W (n) = 0.5 - 0.5 cos (2 π n /N) when 0<= n <= N,

=0 otherwise ——————→5.1

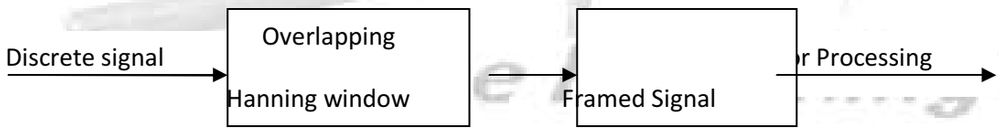

Fig 5.3: Windowing

The frequency-domain spectrum of the Hamming window is much smoother than that of the rectangular window and is commonly used in spectral analysis. The windowing function splits the signal into time-weighted frames.

However, it is not enough to merely process contiguous frames. When the frames are put back together, modulation in the signal becomes evident due to the windowing function.

As the weighting of the window is required, another means of overcoming the modulation must be found. A simple method is to use overlapping windows. To obtain a number of overlapping spectra, the window is shifted along the signal by a number of samples (no more than the window length) and the process is repeated. Simply put, it means that as one frame fades out, its successor fades in. It has the advantage that any discontinuities are smoothed out. However, it does increase the amount of processing required due to the increase in the number of frames produced.

5.3 Morphing

5.3.1 Matching and Warping: Background theory

Both signals will have a number of 'time-varying properties'. To create an effective morph, it is necessary to match one or more of these properties of each signal to those of the other signal in some way. The property of concern is the pitch of the signal - although other properties such as the amplitude could be used - and will have a number of features. It is almost certain that matching features do not occur at exactly the same point in each signal. Therefore, the feature must be moved to some point in between the position in the first sound and the second sound. In other words, to smoothly morph the pitch information, the pitch present in each signals needs to be matched and then the amplitude at each frequency cross-faded. To perform the pitch matching, a pitch contour for the entire signal is required. This is obtained by using the pitch peak location in each cepstral pitch slice.

Consider the simple case of two signals, each with two features occurring in different positions as shown in the figure below.

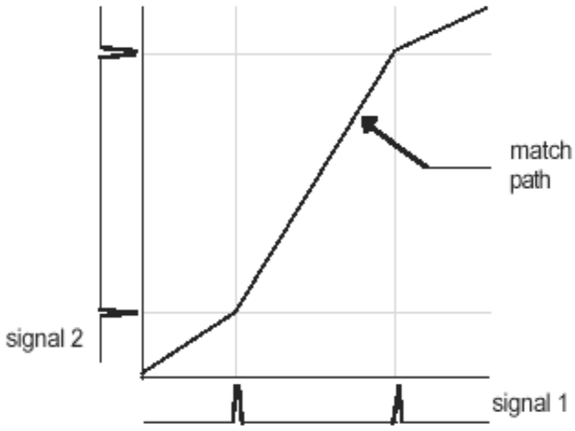

Fig 5.4: the match path between two signals with differently located features

The match path shows the amount of movement (or warping) required in order aligning corresponding features in time. Such a match path is obtained by Dynamic Time Warping (DTW).

5.3.2 Dynamic Time Warping

Speaker recognition and speech recognition are two important applications of speech processing. These applications are essentially pattern recognition problems, which is a large field in itself. Some Automatic Speech Recognition (ASR) systems employ time normalization. This is the process by which time-varying features within the words are brought into line. The current method is time-warping in which the time axis of the unknown word is non-uniformly distorted to match its features to those of the pattern word. The degree of discrepancy between the unknown word and the pattern – the amount of warping required to match the two words - can be used directly as a distance measure. Such time-warping algorithm is usually implemented by dynamic programming and is known as Dynamic Time Warping. Dynamic Time Warping (DTW) is used to find the best match between the features of the two sounds - in this case, their pitch. To create a successful morph, major features, which occur at generally the same time in each signal, ought to remain fixed and intermediate features should be moved or interpolated. DTW enables a match path to be created. This shows how each element in one signal corresponds to each element in the second signal.

In order to understand DTW, two concepts need to be dealt with:

Features: The information in each signal has to be represented in some manner.

Distances: some form of metric has to be used in order to obtain a match path. There are two types:

1. Local: a computational difference between a feature of one signal and a feature of the other.

2. Global: the overall computational difference between an entire signal and another signal of possibly different length.

Feature vectors are the means by which the signal is represented and are created at regular intervals throughout the signal. In this use of DTW, a path between two pitch contours is required. Therefore, each feature vector will be a single value. In other uses of DTW, however, such feature vectors could be large arrays of values. Since the feature vectors could possibly have multiple elements, a means of calculating the local distance is required. The distance measure between two feature vectors is calculated using the Euclidean distance metric. Therefore the local distance between feature vector *x* of signal 1 and feature vector *y* of signal 2 is given by,

$$d(x,y) = \sqrt{\sum_i (x_i - y_i)^2} \qquad \text{5.3}$$

As the pitch contours are single value feature vectors, this simplifies to,

$$d(x,y) = |x - y| \qquad \text{5.4}$$

The global distance is the overall difference between the two signals. Audio is a time- dependent process. For example, two audio sequences may have different durations and two sequences of the sound with the same duration are likely to differ in the middle due to differences in sound production rate. Therefore, to produce a global distance measure, time alignment must be performed - the matching of similar features and the stretching and compressing, in time, of others. Instead of considering every possible match path which would be very inefficient, a number of constraints are imposed upon the matching process.

5.3.3 The DTW Algorithm

The basic DTW algorithm is symmetrical - in other words, every frame in signals must be used. The constraints placed upon the matching process are:

- Matching paths cannot go backwards in time;
- Every frame in each signal must be used in a matching path;
- Local distance scores are combined by adding to give a global distance.

If D (i,j) is the global distance up to (i,j) and the local distance at (i,j) is given by d(i,j)

$$D(i,j) = \min\left[D(i-1,j-1), D(i-1,j), D(i,j-1)\right] + d(i,j) \qquad \text{5.5}$$

Computationally, the above equation is already in a form that could be recursively programmed. However, unless the language is optimized for recursion, this method can be slow even for relatively small pattern sizes. Another method, which is both quicker and requires less memory storage, uses two nested for loops. This method only needs two arrays that hold adjacent columns of the time-time matrix. In the following explanation, it is assumed that the array notation is of the form 0…N-1 for an array of length N.

The only directions in which the match path can move when at (i, j) in the time-time matrix are given in figure 3.8 below.

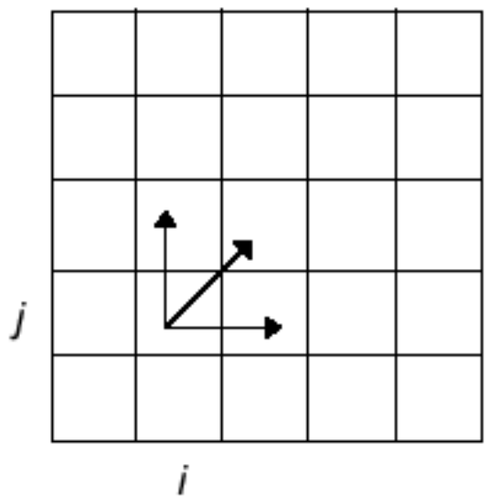

Fig 5.5: Time –Time matrix

The three possible directions in which the best match path may move from cell (i, j) in symmetric DTW.

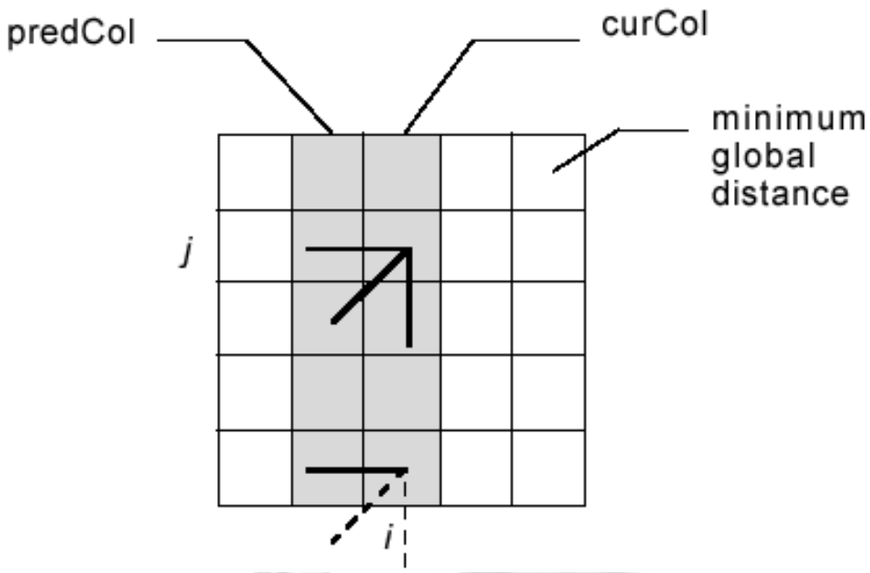

Fig 5.6: Minimum cost path

The cells at (i, j) and (i, 0) have different possible originator cells. The path to (i, 0) can only originate from (i-1, 0). However, the path to (i, j) can originate from the three standard locations as shown in the figure above.

The algorithm to find the least global cost is:

I. Calculate column 0 starting at the bottom most cell. The global cost to this cell is just its local cost. Then, the global cost for each successive cell is the local cost for that cell plus the global cost to the cell below it. This is called the predCol (predecessor column).

II. Calculate the global cost to the first cell of the next column (the curCol). This local cost for the cell plus the global cost to the bottom most cell of the previous column.

III. Calculate the global cost of the rest of the cells of curCol. For example, at (i,j) this is the local distance at (i,j) plus the minimum global cost at either (i-1,j), (i-1,j-1) or (i,j-1).

IV. curCol is assigned to predCol and repeat from step 2 until all columns have been calculated.

V. Global cost is the value stored in the top most cell of the last column.

However, in the case of audio morphing, it is not the minimum global distance itself, which is of interest but the path to achieve. In other words, a back trace array must be kept with entries in the array pointing to the preceding point in the path. Therefore, a second algorithm is required to extract the path.

The path has three different types of direction changes:

- Vertical
- Horizontal
- Diagonal

The back trace array will be of equal size to that of the time-time matrix. When the global distance to each cell, say (i, j), in the time-time matrix is calculated, its predecessor cell is known - it's the cell out of (i-1, j), (i-1, j-1) or (i, j-1) with the lowest global cost. Therefore, it is possible to record in the back trace array the predecessor cell using the following notation (for the cell (i, j)):

1) (i-1, j-1) -- Diagonal
2) (i-1, j) -- Horizontal
3) (i, j-1) -- Vertical

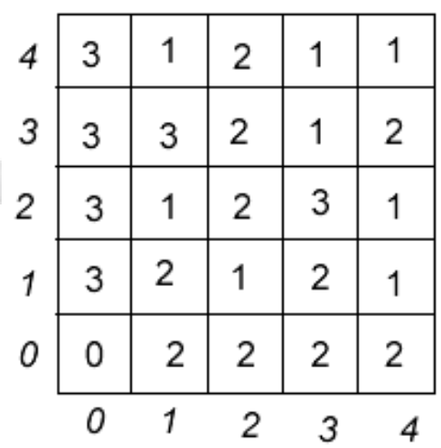

Fig 5.7: A sample back trace array with each cell containing a number, which represents the location of the predecessor cell in the lowest global path distance to that cell.

The path is calculated from the last position, in figure above this would be (4, 4). The first cell in the path is denoted by a zero in the back trace array and is always the cell (0, 0).

A final 2D array is required which gives a pair (signal1 vector, signal2 vector) for each step in the match path given a back trace array similar to that of figure above.

The pseudo code is:

Store the back trace indices for the top right cell.

Obtain the value in that cell - current Val.

While current Val is not 0

If current Val is 1 then reduce both indices by 1

If current Val is 2 then reduce the signal 1 index by 1

If current Val is 3 then reduce the signal 2 index by 2

Store the new indices at the beginning of the 2D array

Obtain the value in that cell – current Val

End.

Therefore, for the example in Figure above, the 2D array would be

signal 1	0	0	1	2	3	4
signal 2	0	1	2	2	3	4

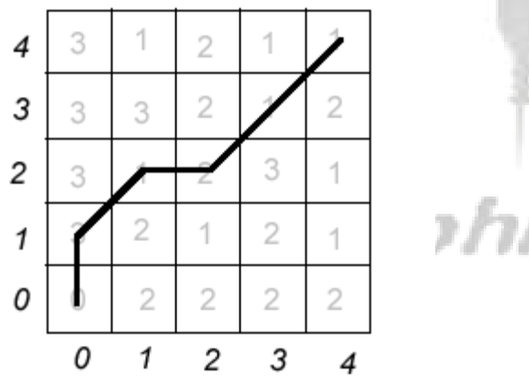

Fig 5.8: the sample back trace array with the calculated path overlaid

At this stage, we now have the match path between the pitches of the two signals and each signal in the appropriate form for manipulation. The next stage is to then produce the final morphed signal.

6. Morphing Stage

Now we shall move to a detailed account of how the morphing process is carried out. The overall aim is to make the smooth transition from signal 1 to signal 2. This is partially accomplished by the 2D array of the match path provided by the DTW. At this stage, it was decided exactly what form the morph would take. The implementation chosen was to perform the morph in the duration of the longest signal. In other words, the final morphed speech signal would have the duration of the longest signal. In order to accomplish this, the 2D array is interpolated to provide the desired duration.

However, one problem still remains: the interpolated pitch of each morph slice. If no interpolation were to occur then this would be equivalent to the warped cross-fade which would still be likely to result in a sound with two pitches. Therefore, a pitch in- between those of the first and second signals must be created. The precise properties of this manufactured pitch peak are governed by how far through the morph the process is. At the beginning of the morph, the pitch peak will take on more characteristics of the signal 1 pitch peak - peak value and peak location - than the signal 2 peak. Towards the end of the morph, the peak will bear more resemblance to that of the signal 2 peaks. The variable l is used to control the balance between signal 1 and signal 2. At the beginning of the morph, l has the value 0 and upon completion, l has the value 1. Consider the example in Figure 4.6. This diagram shows a sample cepstral slice with the pitch peak area highlighted. Figure 4.7 shows another sample cepstral slice, again with the same information highlighted. To illustrate the morph process, these two cepstral slices shall be used.

There are three stages:

1. Combination of the envelope information;

2. Combination of the pitch information residual - the pitch information excluding the pitch peak;

3. Combination of the pitch peak information.

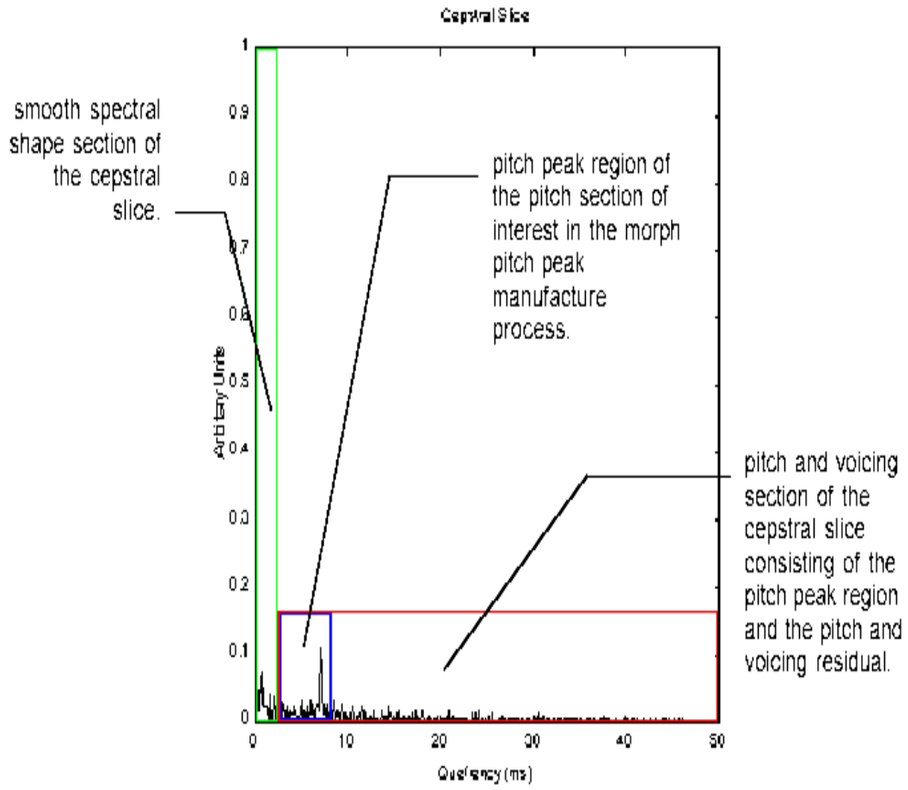

Fig 6.1: A sample cepstral slices with the three main areas of interest in the morphing process

6.1 Combination of the envelope information

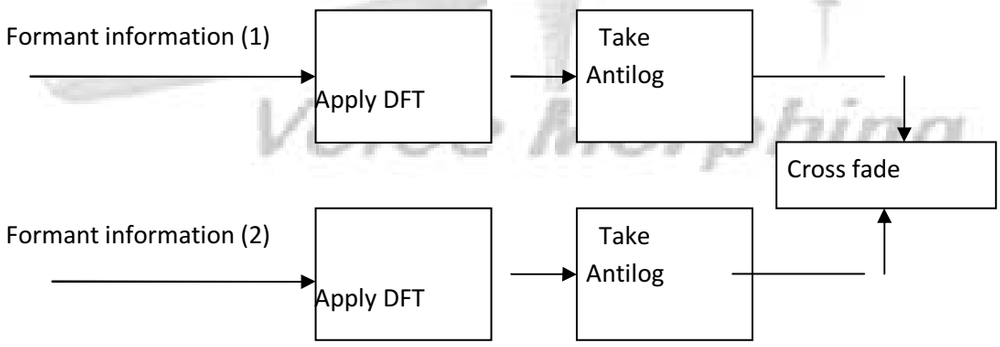

Fig 6.2: Cross fading of the formants.

We can say that that the best morphs are obtained when the envelope information is merely cross-faded, as opposed to employing any pre-warping of features, and so this approach is adopted here. In order to cross-fade any information in the cepstral domain, care has to be taken. Due to the properties of logarithms employed in the cepstral analysis stage, multiplication is transformed into addition. Therefore, if a cross-faded between the two envelopes were attempted, multiplication would in fact take place. Consequently, each envelope must be transformed back into the frequency domain (involving an inverse logarithm) before the cross-fade is performed. Once the envelopes have been successfully cross-faded according to the weighting determined by l, the morphed envelope is once again transformed back into the cepstral domain. This new cepstral slice forms the basis of the completed morph slice.

6.2 Combination of the pitch information residual

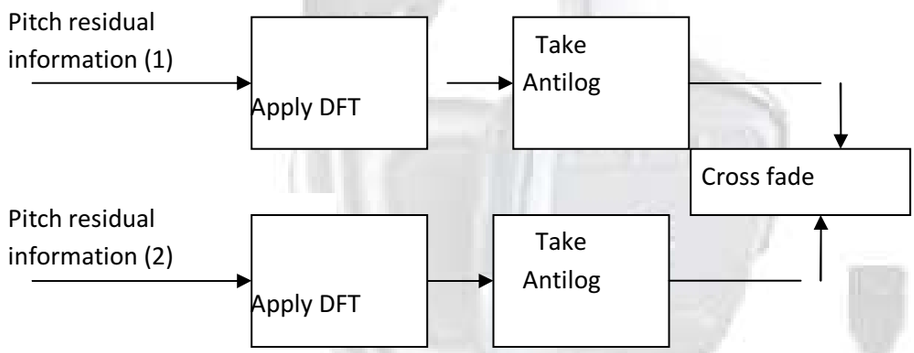

Fig: 6.3 Cross fading of the Pitch information.

The pitch information residual is the pitch information section of the cepstral slice with the pitch peak also removed by liftering. To produce the morphed residual, it is combined in a similar way to that of the envelope information: no further matching is performed. It is simply transformed back into the frequency domain and cross-faded with respect to l. Once the cross-fade has been performed, it is again transformed into the cepstral domain. The information is now combined with the new morph cepstral slice (currently containing envelope information). The only remaining part to be morphed is the pitch peak area.

6.3 Combination of the Pitch peak information

As stated above, in order to produce a satisfying morph, it must have just one pitch. This means that the morph slice must have a pitch peak, which has characteristics of both signal 1 and signal 2. Therefore, an artificial' peak needs to be generated to satisfy this requirement. The positions of the signal 1 and signal 2 pitch peaks are stored in an array (created during the pre-processing, above), which means that the desired pitch peak location can easily be calculated.

In order to manufacture the peak, the following process is performed,

- Each pitch peak area is liftered from its respective slice. Although the alignment of the pitch peaks will not match with respect to the cepstral slices, the pitch peak areas are liftered in such a way as to align the peaks with respect to the liftered area.
- The two liftered cepstral slices are then transformed back into the frequency domain where they can be cross-faded with respect to l. The cross-fade is then transformed back into the cepstral domain.
- The morphed pitch peak area is now placed at the appropriate point in the morph cepstral slice to complete the process.

The morphing process is now complete. The final series of morphed cepstral slices is transformed back in to the frequency domain. All that remains to be done is re-estimate the waveform.

7. Block Diagram for Morphing Process

The whole morphing process is summarized using the detailed block diagram shown below.

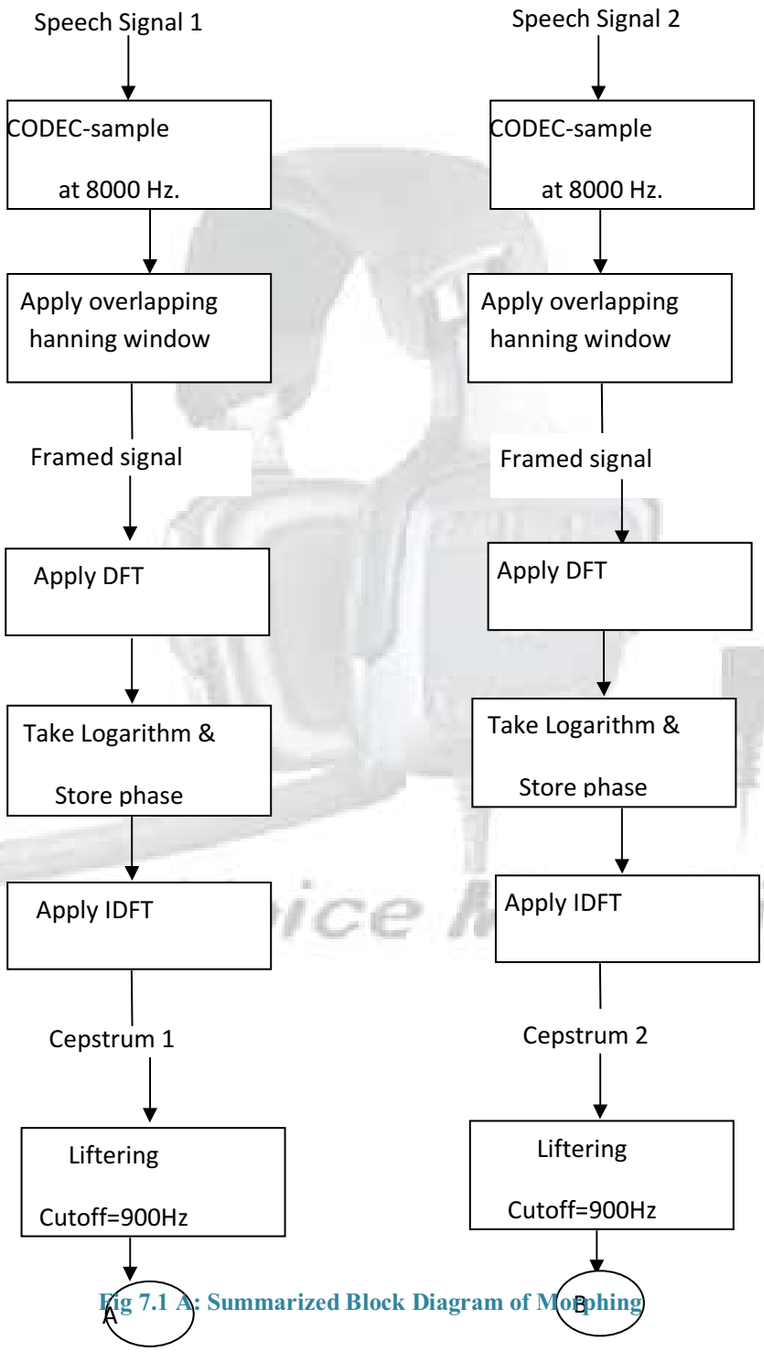

Fig 7.1 A: Summarized Block Diagram of Morphing

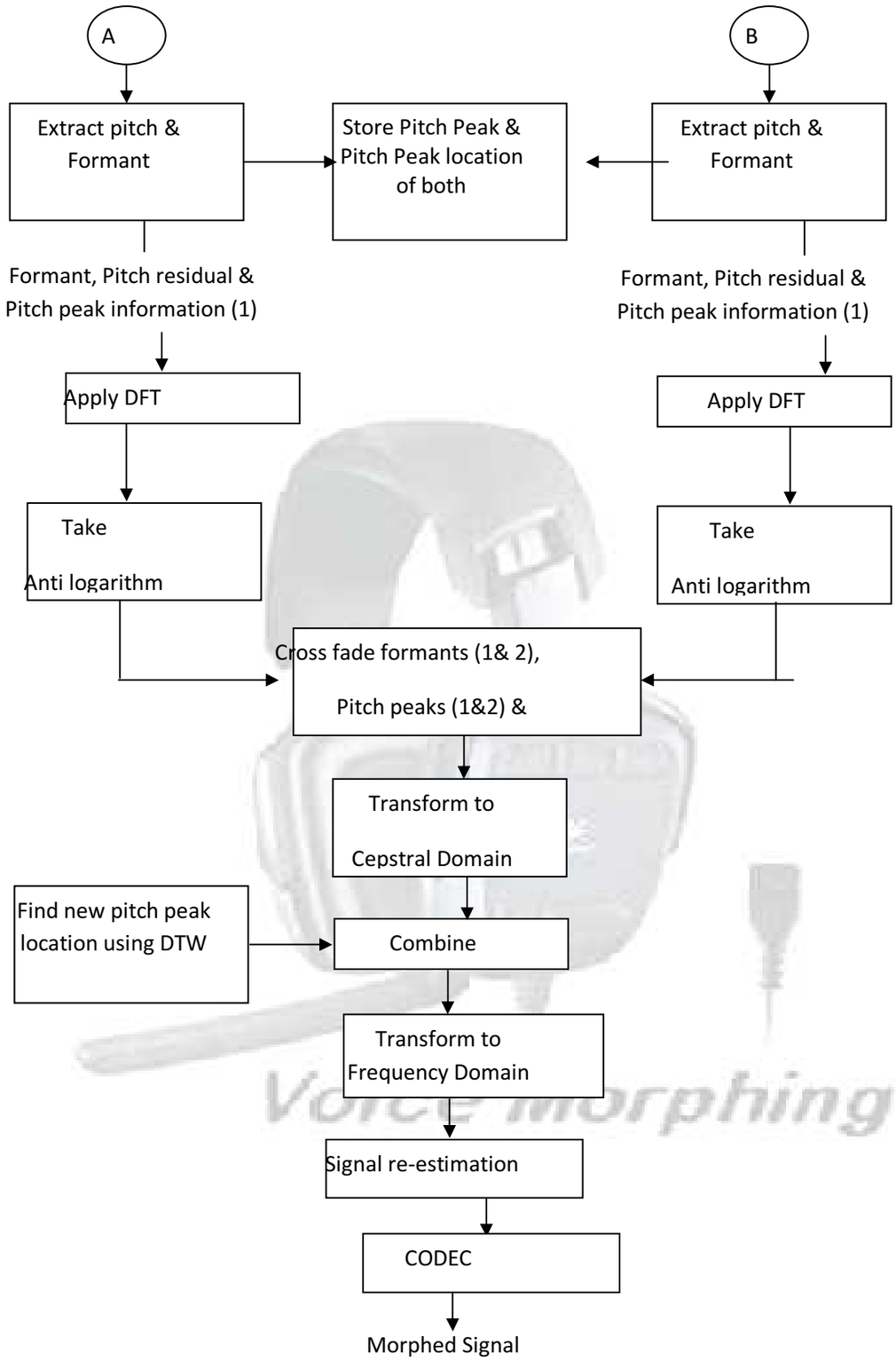

Fig 7.1 B: Summarized Block Diagram of Morphing

8. Steps For Morphing A File Using Software

Step 1: Open a file

- Click on Editor Tab on the module bar to open AV Wave Editor.
- Click Open file button the toolbar to import a file into AV wave editor.

Step 2: Play the file

- Play file by pressing Play button on the player panel.
- Remember to turn on the equalizer to enhance the sound of music.
- One can adjust all bands to their taste or choose an available preset by clicking the presets button.

Fig 8.1: Window of morphing software

- Select a part of vocals to morph by clicking & dragging the mouse.
- Press copy button from the edit menu.

Step 3: Extract the voice

- Press new file in the menu bar to open a new window, and then click on Paste (or ctrl + v).
- Open the effect library on the right of the window and choose any effect. Then select voice Extractor (Center Filtering Method).Its dialog box will appear.
- Press OK and wait a moment for the program to process voice morphing.

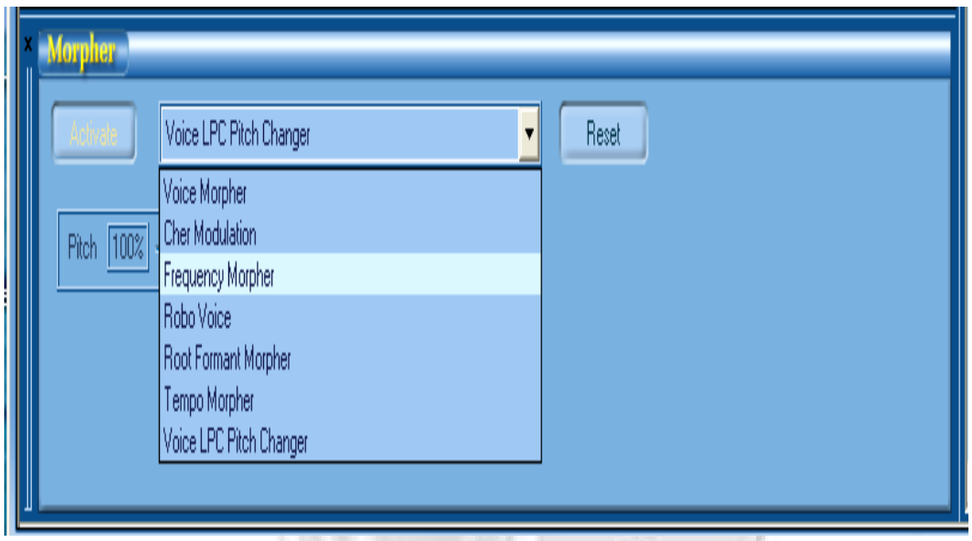

Fig 8.2: Different options to change effects

- Play back the song to see the result. Remember to choose none in the Effect Library so that the voice cannot be morphed twice.

Step 4: Morph voice

- Open the effect library on the right of the window and choose voice morphing Independently of channels quantity. Then select Voice Morphed. Its dialog box will appear.
- Adjust the pitch & timber to your taste to change voice of the selected part. Don't Forget to set the advanced tune at the low level so that the background sound cannot be distorted much.
- Press OK & wait a moment for the program to process voice morphing.
- Playback the file & see the result. Remember to choose none in the effect library so that the voice cannot be morphed twice.

- Ctrl + A then press Copy (or Ctrl + C).

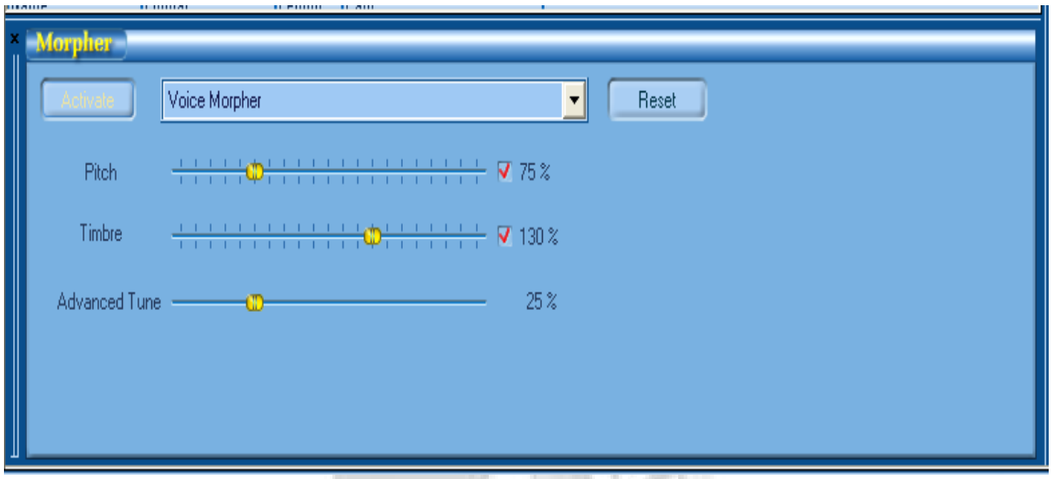

Fig 8.3: Voice Morpher

Step 5: Mix two voices

- Click on the window on the menu bar and back to the original file.
- Press paste mix on the toolbar to mix the two voices.
- Click file on the menu bar and choose a save to have the modified audio sample overwrite its original copy or save as to create a new file leaving the original file intact.

9. Results & Evaluation

In order to evaluate the performance of voice morphing system in terms of it perceptual effects an ABX-style preference test was performed, which is common practice for voice morphing evaluation tests. Figures below show some results of morphing on speech from male-to-male, female-to-female, female-to-male, and male-to-female pairs of speakers. In order to evaluate the performance of system in terms of it perceptual effects an ABX-style preference test was performed, which is common practice for voice morphing evaluation tests. Independent listeners were asked to judge whether an utterance X sounded closer to

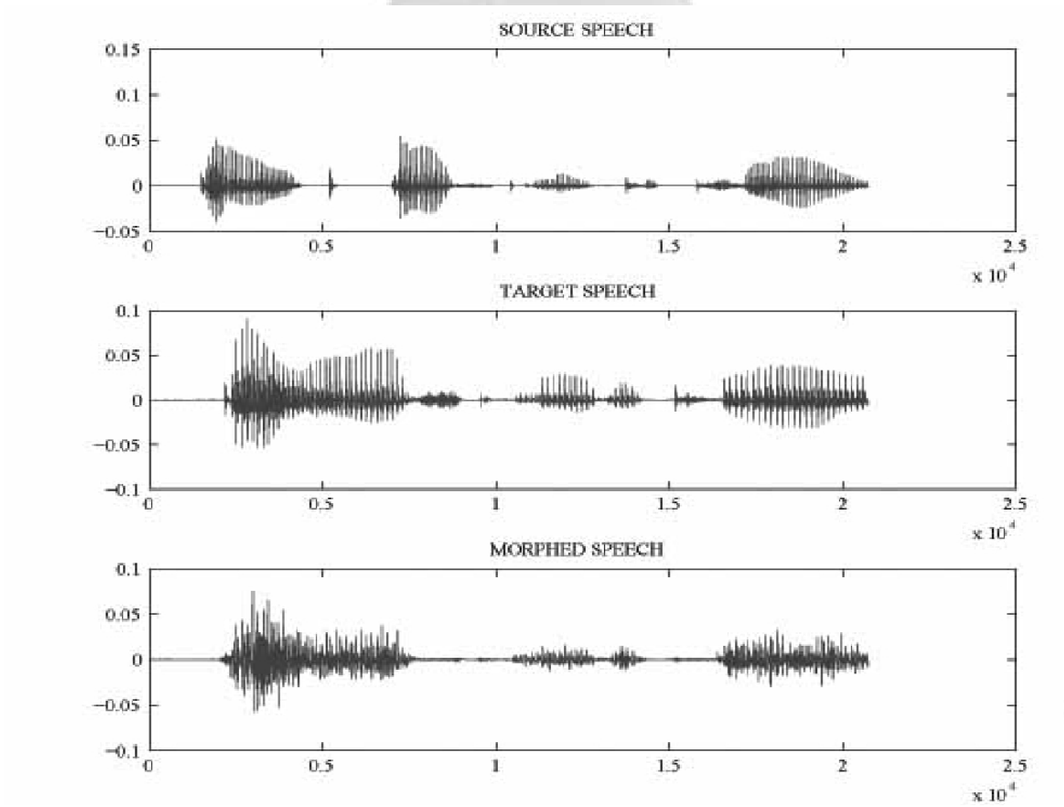

Fig 9.1: Source, Target and Morphed waveform for a male-to-male speaker speech transformation

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.

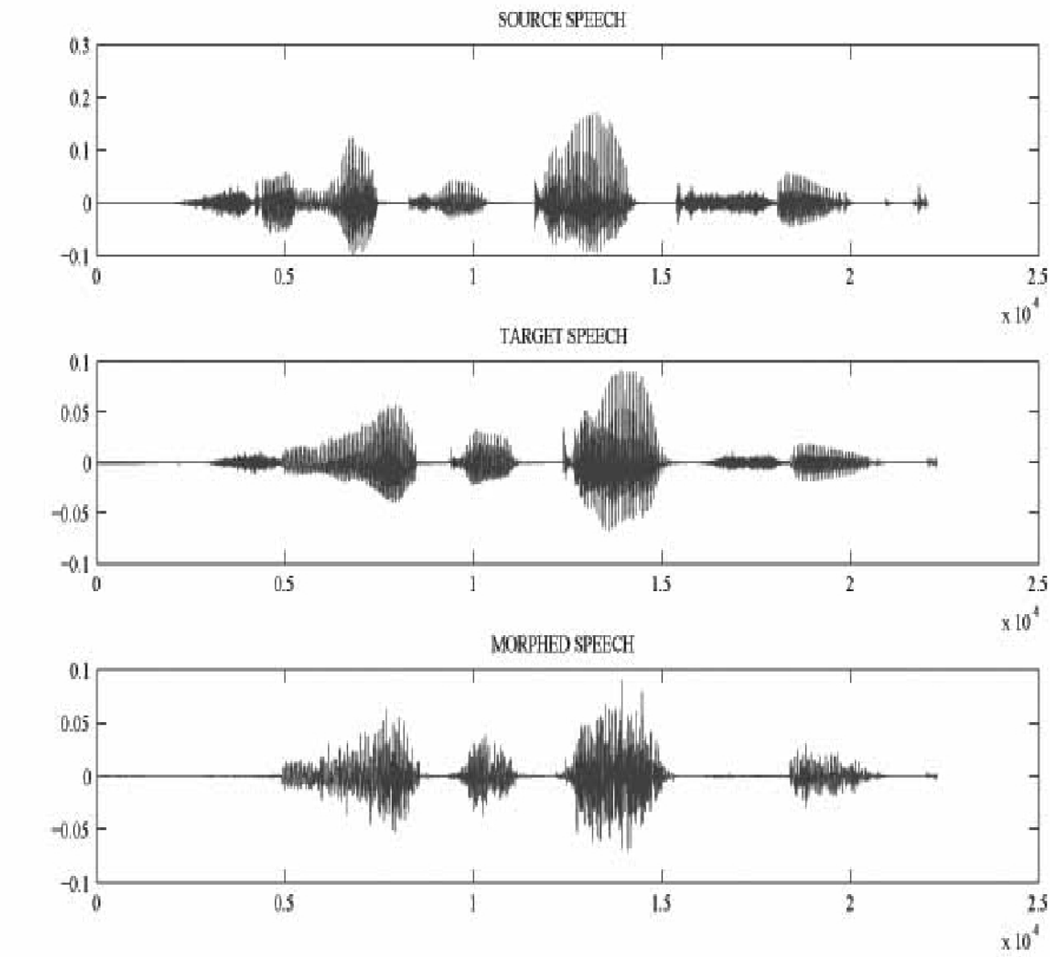

Fig 9.2: Source, Target and Morphed waveform for a female-to-female speaker speech transformation.

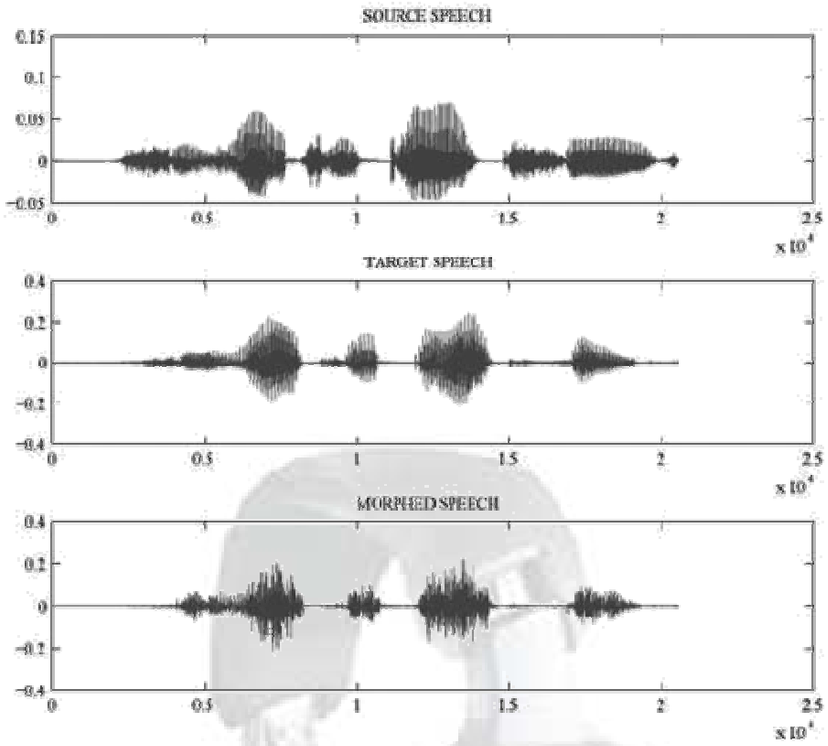

Fig 9.3: Source, Target and Morphed waveform for a female-to-male speaker speech transformation.

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 shown below gives the percentage of the converted utterances that were labeled as closer to the target speaker as well as the p-values.

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

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.

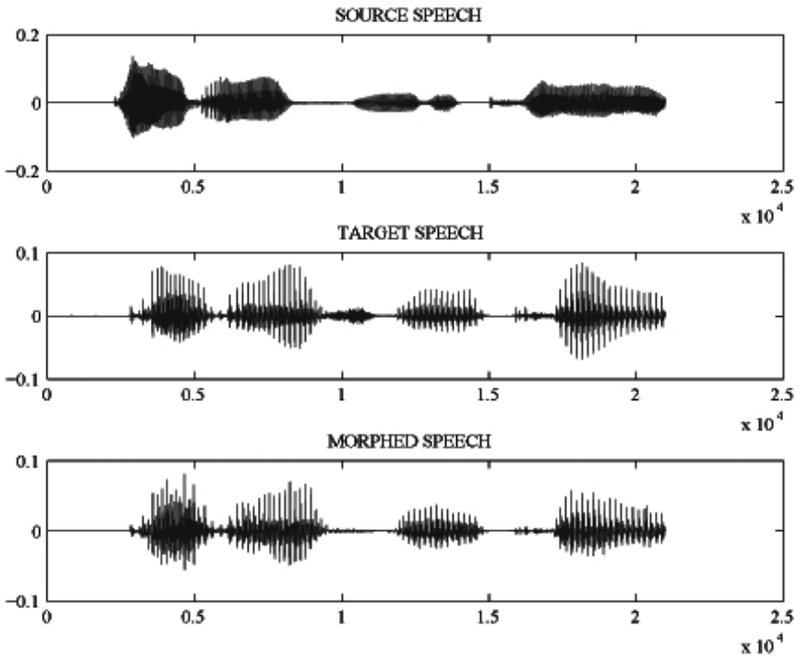

Fig 9.4: Source, Target and Morphed waveform for a male-to-female speaker speech transformation

10. Limitations And Challenges

There are many open problems in voice morphing.

Quality Issues: some of the problems that are perceived as a lack of quality are hissing noise, ringing tones, clicks and also timbral aspects that may be described as a synthetic or unnatural voice. For instance, large pitch shifts without formant correction may degrade quality (and even intelligibility) of the converted voice. These issues have been reported many times and are easily detected in subjective tests.

Similarity Issues: These are related to the timbral quality and vocal identity, mainly correlated to phonetic aspects of speech, although they may easily be confused with quality and prosody issues in experimental tests. In theory, a purely synthetic voice might be perceived as unnatural but similar in timbre to a target voice. Inter-gender voice conversion is particularly susceptible to this type of problem.

Evaluation Issues: Objective measures such as spectral distance or cepstral distortion may be uncorrelated to human perceptual measures, whereas subjective measures such as MOS or ABX may be useful if some sort of experimental benchmark is agreed upon.

Excessive Smoothing Issues: this is a technical issue caused by interpolation methods in the transformation phase, which degrade the spectrum by eliminating details and reducing the similarity of target and converted voices.

Over fitting Issues: This is a counterpart of the previous issue, and is caused by using excessive data in training and obtaining an excessively fine-grained transformation which might produce discontinuities between adjacent frames.

The system only morphs the voiced parts; the unvoiced consonants of the user are directly bypassed to the output. This is done because the morph engine deals better with voiced sounds and the results show that this restriction does not limit the quality of the impersonation. However, some audible artefacts may appear. One emerges from the fact the human voice organ produces all type of voice-unvoiced sounds and the 4 pitch-unpitch boundaries are, in most cases, uncertain. This makes the system sometimes fails in the boundaries of unvoiced-voiced transitions.

11. Advantages & Disadvantages of Voice Morphing

Advantages of Voice Morphing:

- Voice morphing is using technology to change the voice of one person to sound like the voice of another.
- A good system can shift the pitch to make a male voice sound like a female while maintaining the same timing and pronunciation.
- It is used for a variety of purposes like producing cartoons where one person does the voices of a number of characters.
- It can also be used for sinister purposes to disguise ones voice to keep from being identified.
- Allows speech model to be duplicated and an exact copy of a person's voice.
- It's a Powerful combat zone weapon.
- Used to pull out the useful information.

Disadvantages of Voice Morphing:

- There are lots of normalization problems.
- Some applications require extensive sound libraries.
- Different languages require different phonetics.
- It is very seldom complete.

Advantages and disadvantages would depend on what it is being used for.

12. Conclusions

The approach separates the sounds into two forms: spectral envelope information and pitch and voicing information. These can then be independently modified. The morph is generated by splitting each sound into two forms: a pitch representation and an envelope representation. The pitch peaks are then obtained from the pitch spectrograms to create a pitch contour for each sound. Dynamic Time Warping of these contours aligns the sounds with respect to their pitches. At each corresponding frame, the pitch, voicing and envelope information are separately morphed to produce a final morphed frame. These frames are then converted back into a time domain waveform using the signal re-estimation algorithm.

There are a number of areas in which further work should be carried out in order to improve the technique described and extend the field of speech morphing in general. The time required to generate a morph is dominated by the signal re-estimation process. Even a small number of iterations take a significant amount of time even to re-estimate signals of approximately one second duration. Although in speech morphing, an inevitable loss of quality due to manipulation occurs and so less iteration are required, an improved re-estimation algorithm is required.

References

[1] http://www.audio4fun.com

[2] http://www2.maths.ox.ac.uk/ads/Publications/WSEAS_orphanid.pdf

[3] http://mi.eng.cam.ac.uk/~hy216/VoiceMorphingPrj.html

[4] http://www.ime.usp.br/~mqz/SMC2010_Voice.pdf

[5] http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.32.4102&rep=rep1&type=pdf

[6] http://infosearch4u.blogspot.com/2008/01/voice-morphing.html

[7] M. Abe, S. Nakamura, K. Shikano & H. Kuwabara: Voice conversion the vector quantization. IEEE Proceedings of the IEEE ICASSP, 1998, 565–568.

[8] L. Arslan: Speaker transformation algorithm using segmental codebooks (stasc).Speech Communication 28, 1999, 211–226.

[9] L. Arslan and D. Talkin: Voice conversion by codebook mapping of line spectral frequencies and excitation spectrum. Proc. EUROSPEECH, 1997, 1347–1350.

[10] C.M. Bishop: Neural Networks for Pattern Recognition. Clarendon Press, Oxford, 1997.