Available online at www.elixirpublishers.com (Elixir International Journal)

Computer Science and Engineering

Elixir Comp. Sci. & Engg. 54 (2013) 12420-12424

ISSN: 2229-712X

Cepstral approach in voice morphing

Radhika Karthikeyan and G.Ayyappan

Department of Information Technology, Bharath University, Department of MCA, Bharath University.

ARTICLE INFO

Article history:

Received: 1 November 2012;

Received in revised form:

27 December 2012;

Accepted: 5 January 2013;

Keywords

Voice morphing,

Speech morphing,

Cepstral,

Speech signal.

ABSTRACT

Voice morphing means the transition of one speech signal into another. Like image morphing, speech morphing aims to 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 all show one face smoothly changing its shape and texture until it turns into the target face. It is this feature that a speech morph should possess. One speech signal should smoothly change into another, keeping the shared characteristics of the starting and ending signals but smoothly changing the other properties.

The major properties of concern as far as a speech signal is concerned are its pitch and envelope information. These two reside in a convolved form in a speech signal. Hence some efficient method for extracting each of these is necessary. We have adopted an uncomplicated approach namely cepstral analysis to do the same. 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.

© 2013 Elixir All rights reserved.

Introduction

Voice morphing, which is also referred to as voice transformation and voice conversion, is a technique for modifying a source speaker's speech to sound as if it was spoken by some designated target speaker. There are many applications of voice morphing including customizing voices for text to speech (TTS) systems, transforming voice-overs in adverts and films to sound like that of a well-known celebrity, and enhancing the speech of impaired speakers such as laryngectomees. Two key requirements of many of these applications are that firstly they should not rely on large amounts of parallel training data where both speakers recite identical texts, and secondly, the high audio quality of the source should be preserved in the transformed speech. The core process in a voice morphing system is the transformation of the spectral envelope of the source speaker to match that of the target speaker and various approaches have been proposed for doing this such as codebook mapping, formant mapping, and linear transformations. Codebook mapping, however, typically leads to discontinuities in the transformed speech. Although some discontinuities can be resolved by some form of interpolation technique, the conversion approach can still suffer from a lack of robustness as well as degraded quality. On the other hand, formant mapping is prone to formant tracking errors. Hence, transformation-based approaches are now the most popular. In particular, the continuous probabilistic transformation approach introduced by Stylianou provides the baseline for modern systems. In this approach, a Gaussian mixture model (GMM) is used to classify each incoming speech frame, and a set of linear transformations weighted by the continuous GMM probabilities are applied to give a smoothly varying target output. The linear transformations are typically estimated from time-aligned parallel training data using least mean squares. More recently, Kain has proposed a variant of this method in which the GMM classification is based on a joint density model. However, like the original Stylianou approach, it still relies on parallel training data. Although the requirement for parallel training data is often acceptable, there are applications which require voice transformation for nonparallel training data. Examples can be found in the entertainment and media industries where recordings of unknown speakers need to be transformed to sound like well-known personalities. Further uses are envisaged in applications where the provision of parallel data is impossible such

As when the source and target speaker speak different languages. Although interpolated linear transforms are effective in transforming speaker identity, the direct transformation of successive source speech frames to yield the required target speech will result in a number artifacts. The reasons for this are as follows. First, the reduced dimensionality of the spectral vector used to represent the spectral envelope and the averaging effect of the linear transformation result in formant broadening and a loss of spectral detail. Second, unnatural phase dispersion in the target speech can lead to audible artifacts and this effect is aggravated when pitch and duration are modified. Third, unvoiced sounds have very high variance and are typically not transformed. However, in that case, residual voicing from the source is carried over to the target speech resulting in a disconcerting background whispering effect .To achieve high quality of voice conversion, include a spectral refinement approach to compensate the spectral distortion, a phase prediction method for natural phase coupling and an unvoiced sounds transformation scheme. Each of these techniques is assessed individually and the overall performance of the complete solution evaluated using listening tests. Overall it is found that the enhancements significantly improve.

Tele:

E-mail addresses: radhika_karthikeyan@yahoo.com

© 2013 Elixir All rights reserved

Transform-Based Voice Morphing System

Overall Framework

Transform-based voice morphing technology converts the speaker identity by modifying the parameters of an acoustic representation of the speech signal. It normally includes two parts, the training procedure and the transformation procedure. The training procedure operates on examples of speech from the source and the target speakers. The input speech examples are first analyzed to extract the spectral parameters that represent the speaker identity. Usually these parameters encode the short-term acoustic features, such as the spectrum shape and the formant structure. After the feature extraction, a conversion function is trained to capture the relationship between the source parameters and the corresponding target parameters. In the transformation procedure, the new spectral parameters are obtained by applying the trained conversion functions to the source parameters. Finally, the morphed speech is synthesized from the converted parameters. There are three interdependent issues that must be decided before building a voice morphing system. First, a mathematical model must be chosen which allows the speech signal to be manipulated and regenerated with minimum distortion. Previous research suggests that the sinusoidal model is a good candidate since, in principle at least, this model can support modifications to both the prosody and the spectral characteristics of the source signal without inducing significant artifacts However, in practice, conversion quality is always compromised by phase incoherency in the regenerated signal, and to minimize this problem, a pitch synchronous sinusoidal model is used in our system .Second, the acoustic features which enable humans to identify speakers must be extracted and coded. These features should be independent of the message and the environment so that whatever and wherever the source speaker speaks, his/her voice characteristics can be successfully transformed to sound like the target speaker. Clearly the changes applied to these features must be capable of straightforward realization by the speech model. Third, the type of conversion function and the method of training and applying the conversion function must be decided

Spectral Parameters

As indicated above, the overall shape of the spectral envelope provides an effective representation of the vocal tract characteristics of the speaker and the formant structure of voiced sounds. Generally, there are several ways to estimate the spectral envelope, such as using linear predictive coding (LPC), cepstral coefficients, and line spectral frequencies (LSF). The main steps in estimating the LSF envelope for each speech frame are as follows.

1. Use the amplitudes of the harmonics determined by the pitch synchronous sinusoidal model to represent the magnitude spectrum. K is determined by the fundamental frequency, its value can typically range from 50 to 200.
2. Resample the magnitude spectrum non uniformly according to the bark scale frequency warping using cubic spline interpolation.
3. Compute the LPC coefficients by applying the Levinson-Durbin algorithm to the autocorrelation sequence of the warped power spectrum.
4. Convert the LPC coefficients to LSF.
5. In order to maintain adequate encoding of the formant structure, LSF spectral vectors with an order of p=15 were used throughout our voice conversion experiments.

Linear Transforms

We now turn to the key problem of finding an appropriate conversion function to transform the spectral parameters. Assume that the training data contains two sets of spectral vectors X and Y which, respectively, encode the speech of the source speaker and the target speaker A straightforward method to convert the source vectors is to use a linear transform. In the general case, the linear transformation of a p dimensional vector x is represented by a p*(p+1) dimensional matrix W applied to the extended vector x=[x',1]'. Since there are a wide variety speech sounds, a single global transform is not sufficient to capture the variability in human speech. Therefore, a commonly used technique is to classify the speech sounds into classes using a statistical classifier such as a Gaussian mixture model (GMM) and then apply a class-specific transform. However, in practice, the selection of a single N transform from a finite set of transformations can lead to discontinuities in the output signal. In addition, the selected transform may not be appropriate for source vectors that fall in the overlap area between classes. Hence, inorder to generate more robust transformations, a soft classification is preferred in which all N transformations contribute to the conversion of the source vector. The contribution degree of each transformation matrix depends on the degree to which that source vector belongs to the corresponding speech class.

Least Square Error Estimation: When parallel training data is available, the transformation matrices can be estimated directly using the least square error (LSE) criterion. In this case, the source and target vectors are time aligned such that each source training vector xi corresponds to a target training vector yi For ease of manipulation, the general form of the interpolated transformation in (2) can be rewritten compactly as

$$\mathcal{F}(\mathbf{x}) = \left[W_1 \vdots W_2 \vdots \cdots \vdots W_M\right] \begin{pmatrix} \lambda_1(\bar{\mathbf{x}})\bar{\mathbf{x}} \\ \cdots \\ \lambda_2(\bar{\mathbf{x}})\bar{\mathbf{x}} \\ \cdots \\ \vdots \\ \cdots \\ \lambda_M(\bar{\mathbf{x}})\bar{\mathbf{x}} \end{pmatrix} = \bar{\mathbf{W}}\Lambda(\mathbf{x})$$

where

$$\mathbf{W} = \left[W_1 \vdots W_2 \vdots \cdots \vdots W_N\right]_{P\times(N\times(P+1))}$$

and

$$\mathbf{\Lambda}(\mathbf{x}) = \begin{pmatrix} \lambda_1(\mathbf{x})\bar{\mathbf{x}} \\ \cdots \\ \lambda_2(\mathbf{x})\bar{\mathbf{x}} \\ \cdots \\ \vdots \\ \cdots \\ \lambda_N(\mathbf{x})\bar{\mathbf{x}} \end{pmatrix}_{(N\times(P+1))\times 1}.$$

The accurate alignment of source and target vectors in the training set is crucial for a robust estimation of the transformation matrices. Normally, a dynamic time warping

(DTW) algorithm is used to obtain the required time alignment where the local cost function is the spectral distance between source and target vectors. However, the alignment obtained using this method will sometimes be distorted when the source and target speakers are very different, this is especially a problem in cross gender transformation.

Maximum Likelihood Estimation: As noted in the introduction, the provision of parallel training data is not always feasible and hence it would be useful if the required transformation matrices could be estimated from nonparallel data. The form of suggests that, analogous to the use of transforms for adaptation in speech recognition, maximum likelihood (ML) should provide a framework for doing this. To estimate multiple transforms using this scheme, a source GMM is used to assign the source vectors to classes via as in the LSE estimation scheme. A transform matrix is then estimated separately for each class using the above ML scheme applied to just the data for that class. Though it is theoretically possible to estimate multiple transforms using soft classification, in practice, matrices will become too large to invert. Hence, the simpler hard classification approach is used here. As with the least mean squares method using parallel data, performance is greatly improved if subphone segment boundaries can be accurately determined in the source data using the target HMM and "forced alignment" recognition mode. This enables the set of Gaussians evaluated for each source frame to be limited to just those associated with the HMM state corresponding to the associated subphone. This does, of course, require that the orthography of the source utterances be known. Similarly, knowing the orthography of the target training data makes training the target HMM simpler and more effective.

System Enhancement

The converted speech produced by the baseline system described above will often contain artifacts. This section discusses these artifacts in more detail and describes the solutions developed to mitigate them.

Phase Prediction

As is well known, the spectral magnitude and phase of human speech are highly correlated. In the baseline system, when only spectral magnitudes are modified and the original phase is preserved, a harsh quality is introduced into the converted speech. However, to simultaneously model the magnitude and phase and then convert them both via a single unified transform is extremely difficult. A GMM model is first trained to cluster the target spectral envelopes coded via LSF coefficients into M classes (C1,……..,CM). For each target envelope v we have a set of posterior probabilities. This can be regarded as another form of representation of the spectral shape. A set of template signal T = [T1,……., TM] can be estimated by minimising the waveform shape prediction error.

Spectral Refinement

Although the formant structure of the source speech has been transformed to match the target, the spectral detail has been lost as a result of reducing the dimensionality of the envelope representation during the transform. Another clearly visible effect is the broadening of the spectral peaks caused, at least in part, by the averaging effect of the estimation method. All these degradations lead to muffled effects in the converted speech. To solve this problem, a straightforward idea is to reintroduce the lost spectral details to the converted envelopes. A spectral residual prediction approach has been developed to do this based on the residual codebook method, where the codebook is trained using a GMM model. After the residual codebook is obtained, the spectral residual needed to compensate each converted spectral envelope can be predicted straightforwardly based on the posterior probabilities.

Transforming Unvoiced Sounds

Many unvoiced sounds, have some vocal tract coloring and simply copying the source to the target affects the converted speech characteristics, especially in cross gender conversion. A typical effect is the perception of another speaker whispering behind the target speaker. Since most unvoiced sounds have no obvious vocal tract structure and cannot be regarded as short-term stationary signals, their spectral envelopes show large variations. Therefore, it is not effective to convert them using the same solution as for voiced sounds. However randomly deleting, replicating and concatenating segments of the same unvoiced sound does not induce significant artifacts. This observation suggests a possible solution based on unit selection and concatenation to transform unvoiced sounds.

Realtime Voice Morphing

In real time voice morphing what we want is to be able to morph, in real-time user singing a melody with the voice of another singer. It results in an "impersonating" system with which the user can morph his/her voice attributes, such as pitch, timbre, vibrato and articulation, with the ones from a prerecorded target singer. The user is able to control the degree of morphing, thus being able to choose the level of "impersonation" that he/she wants to accomplish. In our particular implementation we are using as the target voice a recording of the complete song to be morphed. A more useful system would use a database of excerpts of the target voice, thus choosing the appropriate target segment at each particular time in the morphing process. In order to incorporate to the user's voice the corresponding characteristics of the "target" voice, the system has to first recognize what the user is singing(phonemes and notes), finding the same sounds in the target voice (i.e. synchronizing the sounds), then interpolate the selected voice attributes, and finally generate the output morphed voice. All this has to be accomplished in real-time.

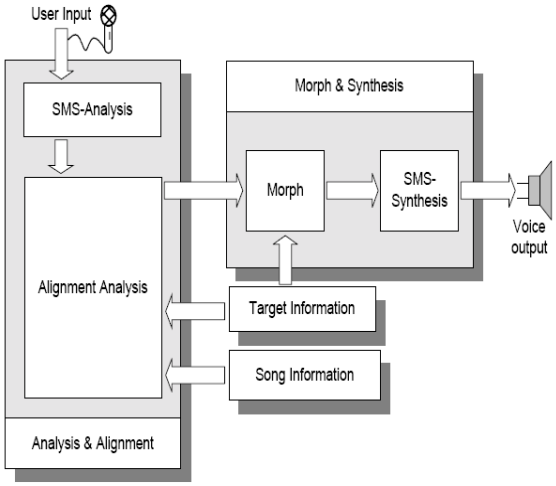

Fig 4.1 System block diagram

The Voice Morphing System

Figure shows the general block diagram of the voice impersonator system. The underlying analysis/synthesis technique is SMS to which many changes have been done to better adapt it to the singing voice and to the real-time constrains of the application. Also 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 song we have to supply information about the song to be morphed and the song recording itself (Target Information and Song 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 singer that originally sang the song 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 song. 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 realtime 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. The recognition/alignment algorithm is based on traditionalspeech recognition technology, that is, Hidden Markov Models (HMM) that were adapted to the singing voice. Once a user frame is matched with a target frame, we morph them 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.

Voice analysis/synthesis using SMS

The traditional SMS analysis output is a collection of frequency and amplitude values that represent the partials of the sound (sinusoidal component), and either filter coefficients with a gain value or spectral magnitudes and phases representing the residual sound (non sinusoidal component. Several modifications have been done to the main SMS procedures to adapt them to the requirements of the impersonator system. A major improvement to SMS has been the real-time implementation of the whole analysis/synthesis process, with a processing latency of less than 30 milliseconds and tuned to the particular case of the singing voice. This has required many optimizations in the analysis part, especially in the fundamental frequency detection algorithm. These improvements were mainly done in the pitch candidate's search process, in the peak selection process, in the fundamental frequency tracking process, and in the implementation of a voiced-unvoiced gate. Another important set of improvements to SMS relate to the incorporation of a higher-level analysis step that extracts the parameters that are most meaningful to be morphed. Attributes that are important to be able to interpolate between the user's voice and the target's voice in a karaoke application include spectral shape, fundamental frequency, amplitude and residual signal. Others, such as pitch micro variations, vibrato, spectral tilt, or harmonicity, are also relevant for various steps in the morphing process or to perform other sound transformation that are done in parallel to the morphing. For example, transforming some of these attributes we can achieve voice effects such as Tom Waits hoarseness.

Phonetic recognition/alignment

This part of the system is responsible for recognizing the phoneme that is being uttered by the user and also its musical context so that a similar segment can be chosen from the target information. There is a huge amount of research in the field of speech recognition. The recognition systems work reasonably well when tested in the well-controlled environment of the laboratory. However, phoneme recognition rates decay miserably when the conditions are adverse. In our case, we need a speaker independent system capable of working in a bar with a lot of noise, loud music being played and not very-high quality microphones. Moreover the system deals with singing voice, which has never been worked on and for which there are no available databases. It has to work also with very low delay, we cannot wait for a phoneme to be finished before we recognize it and we have to assign a phoneme to each frame.

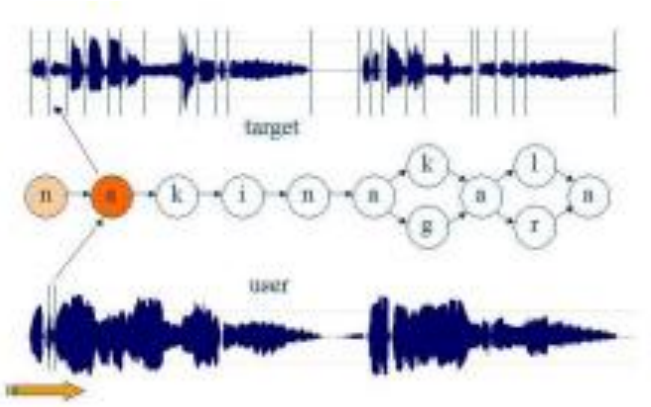

Fig 4.2 Recognition and matching of morphable units

This would be a rather impossible/impractical problem if it was not for the fact that we know the words beforehand, the lyrics of the song. This reduces a big portion of the search problem: all the possible paths are restricted to just one string of phonemes, with several possible pronunciations. Then the problem reduces to locating the phoneme

in the lyrics and placing the start and end points. We have incorporated a speech recognizer based on phoneme-base discrete HMM's that handles musical information and that is able to work with very low delay. The details of the recognition system can be found in another paper of our group. The recognizer is also used in the preparation of the target audio data, to fragment the recording into morphable units (phonemes) and to label them with the phonetic transcription and the musical context. This is done out of real-time for a better performance.

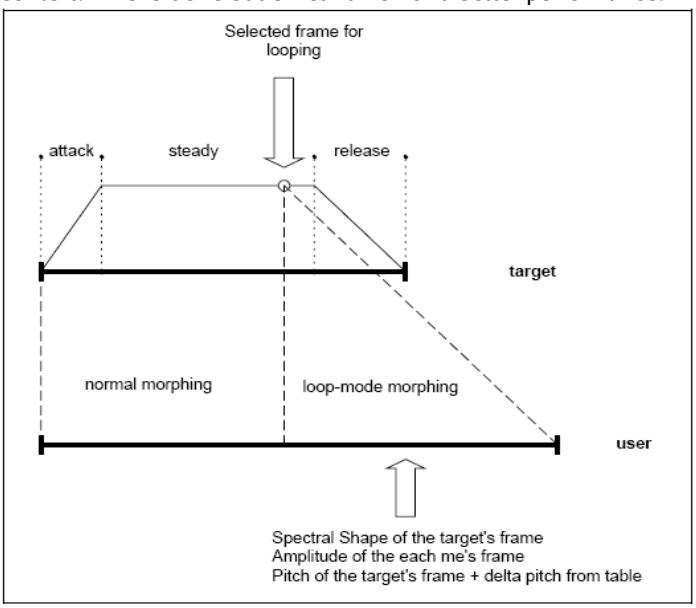

Fig 4.3 Loop synthesis diagram

Morphing

Depending on the phoneme the user is singing, a unit from the target is selected. Each frame from the user is morphed with a different frame from the target, advancing sequentially in time. Then the user has the choice to interpolate the different parameters extracted at the analysis stage, such as amplitude, fundamental frequency, spectral shape, residual signal, etc. In general the amplitude will not be interpolated, thus always using the amplitude from the user and the unvoiced phonemes will also not be morphed, thus always using the consonants from the user. This will give the user the feeling of being in control. In most cases the durations of the user and target phonemes to be morphed will be different. If a given user's phoneme is shorter

than the one from the target the system will simply skip the remaining part of the target phoneme and go directly to the articulation portion. In the case when the user sings a longer phoneme than the one present in the target data the system enters in the loop mode. Each voiced phoneme of the target has a loop point. frame, marked in the preprocessing, non-real time stage. The system uses this frame to loop-synthesis in case the user sings beyond that point in the phoneme. Once we reach this frame in the target, the rest of the frames of the user will be interpolated with that same frame until the user ends the phoneme. This process is shown in Figure

The frame used as a loop frame requires a good spectral shape and, if possible, a pitch very close to the note that corresponds to that phoneme. Since we keep a constant spectral shape, we have to do something to make the synthesis sound natural. The way we do it is by using some “natural” templates obtained from the analysis of a longer phoneme that are then used to generate more target frames to morph with out of the loop frame. One feature that adds naturalness is pitch variations of a steady state note sung by the same target. These delta pitches are kept in a look up table whose first access is random and then we just read consecutive values. We keep two tbles, one with variations of steady pitch and another one with vibrato to generate target frames. Once all the chosen parameters have been interpolated in a given frame they are added back to the basic SMS frame of the user. The synthesis is done with the standard synthesis procedures of SMS.

References

1. Quality-enhanced Voice Morphing using Maximum Likelihood Transformations Hui Ye and Steve Young
2. High quality Voice Morphing Hui Ye and Steve Young
3. www.wikipedia.org