Voice Morphing System for Impersonating in Karaoke Applications

Pedro Cano, Alex Loscos, Jordi Bonada, Maarten de Boer, Xavier Serra
Audiovisual Institute, Pompeu Fabra University
Rambla 31, 08002 Barcelona, Spain
{pcano, aloscos, jboni, mdeboer, xserra}@iua.upf.es
http://www.iua.upf.es

Abstract

In this paper we present a real-time system for morphing two voices in the context of a karaoke application. As the user sings a pre-established song, his pitch, timbre, vibrato and articulation can be modified to resemble those of a pre-recorded and pre-analyzed recording of the same melody sang by another person. The underlying analysis/synthesis technique is based on SMS, to which many changes have been done to better adapt it to the singing voice and the real-time constrains of the system. Also a recognition and alignment module has been added for the needed synchronization of the user's voice with the target's voice before the morph is done. There is room for improvements in every single module of the system, but the techniques presented have proved to be valid and capable of musically useful results.

1. Introduction

With different names, and using different signal processing techniques, the idea of audio morphing is well known in the Computer Music community (Serra, 1994; Tellman, Haken, Holloway, 1995; Osaka, 1995; Slaney, Covell, Lassiter. 1996; Settel, Lippe, 1996). The main goal of the developed audio morphing methods is the smooth transformation from one sound to another, thus, the combination of two sounds to create a new sound with an intermediate timbre. Most of these methods are based on the interpolation of sounds parameterizations resulting from analysis/synthesis techniques, such as the Short-time Fourier Transform (STFT), Linear Predictive Coding (LPC) or Sinusoidal Models.

In this paper we present a very particular case of audio morphing. What we want is to be able to morph, in real-time, a 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.

The obvious use of our technique is in Karaoke applications. In such a situation it is very common for the user to want to impersonate the singer that originally sang the song. Our system is capable to do that automatically.

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.

Next we present the overall system functionality, then we discuss the basic techniques used and finally we comment on the results obtained. In another paper (Cano, Loscos, Bonada, M. de Boer, Serra, 2000) the actual software implementation is discussed.

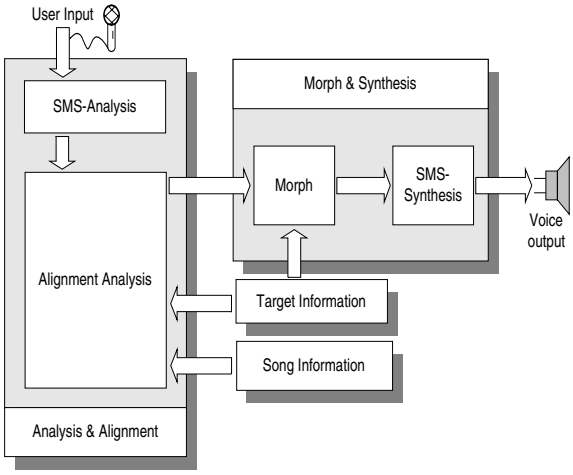

Figure 1. System block diagram.

2. The Voice Morphing System

Figure 1 shows the general block diagram of the voice impersonator system. The underlying analysis/synthesis technique is SMS (Serra, 1997) 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 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. The recognition/alignment algorithm is based on traditional speech recognition technology, that is, Hidden Markov Models (HMM) that were adapted to the singing voice (Loscos, Cano, Bonada, 1999).

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.

3. 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) (Serra, 1997). 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 (Cano, 1998). 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 (Cano, Loscos, 1999).

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 (Serra, Bonada, 1998). 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 (Childers, 1994).

4. 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.

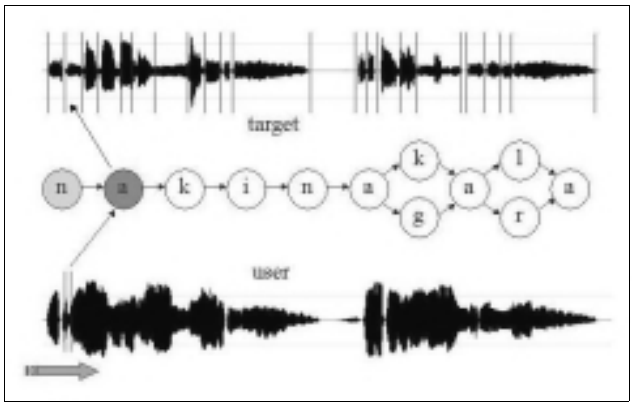

Figure 2. Recognition and matching of morphable units.

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.

Besides knowing the lyrics, music information is also available. The user is singing along with the music and hopefully according to a tempo and melody already specified in the score of the song. We thus also know the time at which a phoneme is supposed to be sung, its approximate duration, its associated pitch, etc. All this information is used to improve the performance of the recognizer and also to allow resynchronization, for example in the case that the singer skips a part of the song.

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 (Loscos, Cano, Bonada, 1999).

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.

5. 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 3.

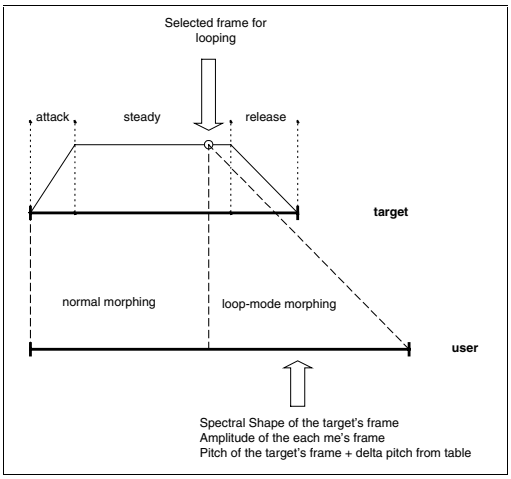

Figure 3. Loop synthesis diagram.

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 tables, 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.

6. Experiments

The singing voice impersonator has been implemented on a PC platform (Cano, Loscos, Bonada, de Boer, Serra, 2000). To check the feasibility of the real-time technology presented, that is the SMS based morph engine and the recognizer, the target data used was a complete song, as shown in Figure 2, instead of a database of target excerpts. Thus the search for the most appropriate morphing frame pairs becomes a simple process.

Another simplification is that 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

pitch-unpitch boundaries are, in most cases, uncertain. This makes the system sometimes fails in the boundaries of unvoiced-voiced transitions. The other problem appears when the interpolation factor for the spectral shape parameter is set to be around 50%. Since the shapes are linearly interpolated, the morphed spectral shape is too smoothed and looses the timbre character of the original voices. This is currently being solved by working on a more complex model for the spectral shape that takes into account the formants to do the interpolation.

The HMM phonetic models were trained with a limited singing voice database. It is a fact that the recognition step works better when the user singer has been used to train the database. We believe that taking into account context and using non-discrete symbol probability distributions would bring better results but they require bigger databases.

The system as a whole produces quite high quality sound. The delay between the sound input and the final sound output in the running system is less than 30 milliseconds. This delay is just good enough to make the user have the feeling of being in control of the output sound.

7. Conclusions

In this paper we have presented a singing voice morphing system. Obviously, there is room for improvements in every single module of the system, but the techniques presented have proved to be valid and capable of musically useful results.

The final purpose of the system was to make an average singer sing any song like any professional singer. In fact, we would like the system to morph the user's voice with qualities of several singers, for instance a mixture timbre of *Sinatra* and *Tom Jones* and the horseness of Tom Waits. However, at this point, and due to the limitations of our system, we need a clear recording of *Tom Jones*, or whomever the user wants to impersonate, singing the song. It is not easy to have this kind of popular professional singer to record songs for us and so in this project, we used professional impersonators' recordings. However it is clear that is by no means efficient, not only because of technical issues like memory requirements, but also due to the cost of having professional singers recording every song of the system. To allow the user to sing any song with the voice and expression of whomever he wants without having a professional singer singing each song for each possible timbre, we will need a model for every desired target voice and also every type of singing style. In order to achieve this, techniques to perform the match of the phonemes, considering style and musical context, must be incorporated into the system. One approach for the case of the saxophone has been studied (Arcos, Lopez de Mantaras, Serra, 1997).

8. Acknowledgments

We would like to acknowledge the support from Yamaha Corporation and the contribution to this research of the other members of the Music Technology Group of the Audiovisual Institute.

References

Arcos, J. LL., R. Lopez de Mántaras, X. Serra. 1997. "Saxex: a Case-Based Reasoning System for Generating Expressive Musical Performances". *Proceedings of the ICMC 1997.*

Cano, P. 1998. "Fundamental Frequency Estimation in the SMS Analysis". *Proceedings of the Digital Audio Effects Workshop*, 1998.

Childers, D.G. 1994. "Measuring and Modeling Vocal Source-Tract Interaction". *IEEE Transactions on Biomedical Engineering 1994.*

Cano, P., A. Loscos. 1999. *Singing Voice Morphing System based on SMS. UPC,* 1999.

Cano, P., A. Loscos, J. Bonada, M. de Boer, X. Serra. 2000. "Singing Voice Impersonator Application for PC". *Proceedings of the ICMC 2000.*

Loscos, A., P. Cano, J. Bonada. 1999. "Low-Delay Singing Voice Alignment to text". *Proceedings of the ICMC* 1999.

Osaka, N. 1995. "Timbre Interpolation of sounds using a sinusoidal model", *Proceedings of the ICMC* 1995.

Serra, X. 1994. "Sound hybridization techniques based on a deterministic plus stochastic decomposition model", *Proceedings of the ICMC 1994.*

Serra, X. 1997. "Musical Sound Modeling with Sinusoids plus Noise". G. D. Poli and others (eds.), *Musical Signal Processing*, Swets & Zeitlinger Publishers, 1997.

Serra, X., J. Bonada. 1998. "Sound Transformations on the SMS High Level Attributes". *Proceedings of 98 Digital Audio Effects Workshop*, Barcelona 1998.

Settel, Z., C. Lippe. 1996. "Real-Time Audio Morphing", 7th International Symposium on Electronic Art, 1996.

Slaney, M., M. Covell, B. Lassiter. 1996. "Automatic audio morphing", *Proc. IEEE Int. Conf. Acosut. Speech Signal Process*. 2, 1001-1004 (1996).

Tellman, E., L. Haken, B. Holloway. 1995. "Timbre Morphing of Sounds with Unequal Number of Features", *J. Audio Eng. Soc.,* 43:9 1995.