Developing a Timbre Morphing Package: The Process of Automatic Feature Identification

Ciarán J. Hope
Dept. of Entertainment Studies and Performing Arts UCLA

Abstract

The objective of this paper is to investigate the accuracy of the process known as Automatic Feature Identification (AFI), when attempting to morph two or more tones together to create a ‘mongrel’ sound in the process known as timbre morphing. Morphing is examined using the processes endemic to the morphing package called *Mongrel* which concentrates on AFI and enhanced feature identification. The automatic processes are examined, and the resulting morphs from the process are tested on a sample base of musical and non- musically trained listeners to check for accuracy of representation in the sounds derived.

1 INTRODUCTION

The term *Morphing* derives from the Greek word *Morph* which refers to form. Morphing, as such, is the technique of blending one form into another. In fact, timbre morphing can be defined as the process of combining two or more sounds to create a new sound of intermediate timbre (Holloway, Tellman and Haken 1995). Previous research investigated ‘audio’ morphing through the representation of sound in a multi- dimensional space (Slaney, Covell and Lassiter 1996), developing a new approach based on separate spectrograms to encode the pitch and broad spectral shapes of the sound, focusing on correctly identifying pitches. This work concluded that better representations, matching techniques and more natural sounding interpolation schemes - especially perceptually optimal interpolation functions - should be developed.

Another practical approach to Timbre Morphing was developed by the CERL Sound Group at the University of Illinois (Walker and Fitz 1996). They developed a package called *Lemur* , which outputs a file containing a sequential list of frames, each describing a small portion of the sound. The time-frequency analysis method that Lemur uses to generate these frames is effective, but not optimal in its representation of a sound due to both temporal and spectral smearing.

The work being analysed in this paper, derives from a software package called *Mongrel* which was developed at the Audio Laboratory in Trinity College Dublin. Mongrel takes two tones and cross-breeds them to generate a mongrel tone which contains characteristics of both parent tones. The motivation behind this development is to aid contemporary electro-acoustic composers in their exploration of new timbres in their compositional process. Two of the main differences between this work and that already complete by other teams, was the use of improved time-frequency analysis representations and a simplified software interface which allows a composer to generate new sounds as they please.

To achieve the objective of morphing two tones into a third new tone, the tones must be processed correctly to ensure accurate representation. The accepted and logical method for causing maximum contribution on the part of both parent tones is to identify certain *features* in the contributing spectrograms, and morph these characteristic features into one another. In the Mongrel morphing package, all signal manipulation and morphing is implemented automatically, keeping user interaction to a minimum. The user simply chooses the two tones to morph and how much of each tone is required in the new sound. This paper sets out to analyse the effectiveness of this implementation by analysing the methodology of the automatic processes and through subject testing with the resultant morphed hybrid sounds.

2 TIMBRE REPRESENTATION

Two sounds of the same pitch and loudness may have recognisably different qualities: for instance the sounds of string instruments versus reed instruments in the orchestra. These distinguishing qualities of sound are called *timbre* (defined as that part of a sound which is neither loudness nor pitch) and are sometimes compared to visible colour. Despite this negative definition timbre is one of the primal attributes by which we characterise sounds. Initially, researchers believed the classical view that the *spectrum* of a sound was the essential ingredient for the reproduction of that sound. However, the importance of the temporal aspects of a sound located in *timbre space* were first noted by Stumpf (Stumpf 1926) in 1910 when he observed that instrument recognition was impaired by removing the onset stage of a tone - the first main temporal section of a musical tone. Thus, to allow the temporal development of spectral components in a musical sound to be observed, it is necessary to combine time-amplitude and frequency-amplitude representations. When the signal is observed using a time-frequency representation it is imperative that these components can be effectively quantified.

The CERL sound group define a *feature* as being “any portion of the sound that is important in the morphing process” (Walker and Fitz 1996). In order to facilitate a user-friendly package, a primary requirement is the use of Automatic Feature Identification. During the morphing process, *spectral features* - which include the fundamental and harmonics -and *transient* or *temporal features* such as the onset, steady state and decay, which are known to contribute perceptually to what a listener hears, need to be identified. If any one particular harmonic (itself a *spectral feature* of a complex sound), for example, is taken and analysed, the temporal fluctuations associated with it can be observed. Figure 1 shows an example of a harmonic from a clarinet spectrum. The diagram subdivides the tone into three temporal divisions: Onset, Steady State, and Decay. The significance of each feature varies according to instrument, as well as with duration and pitch of the tone.

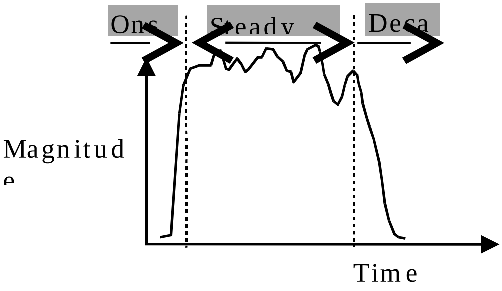

Figure 1 – The First Harmonic of a Clarinet ‘A’, 440Hz.

The *Short-Time Fourier Transform* (STFT) or the *Spectrogram* (Furlong and Hope 1997) is a joint time-frequency representation which enables us to observe the temporal and spectral characteristics at any point in this timbre space. However, the STFT has an intrinsic problem in that it introduces both spectral and temporal smearing. As such, its use as an analysis tool for feature identification is somewhat questionable. The *Wigner Distribution* (WD) has however, been examined as an optional time-frequency analysis method for musical signals, instead of the STFT as it contributes less spectral smearing than the spectrogram, and no temporal smearing (Furlong and Hope 1997). Therefore it is a more accurate representation to use for the purpose of musical synthesis.

3 THE PROCESS OF CREATING A MONGREL TONE

Say a composer chooses a violin 'starting' timbre and a flute 'finishing' timbre and manipulates various important (spectral and temporal) features so that one can move smoothly between them. This manipulation opens up new parts of timbre space which nevertheless are not entirely unknown to us in that we would at least be operating between the known boundaries of 'violin' and 'flute', or whatever other timbres had been chosen at the onset. Thus morphing would provide the composer with access to a much richer *timbre palette* than before in that previously unexplored regions of timbre space could be made available for creative manipulation. By basing such morphing tools on the WD rather than the STFT, representational distortions would be minimised as a result of no temporal smearing and less spectral smearing in the WD, ensuring that detailed spectral and temporal features of real instrument tones would be made clearer for the purposes of computational analysis and synthesis.

The specific feature identification that is being used in Mongrel concentrates on a number of pre-defined perceptually significant features, *i.e.* the harmonics, their onset, steady state, and decay envelopes. These were chosen as the foundation for an algorithm encompassing *Automatic-Feature-Identification* (AFI) which would choose the important features of a tone and number them accordingly without user intervention. This algorithm is called an *Automatic-Feature-Identification-Algorithm* (AFIA).

Figure 2 shows a diagrammatic version of the operation of the morphing algorithm used in Mongrel. Both main design characteristics are built into the design of the algorithm. Both tones, are processed in a *representation coder* or *TF-distribution Generator*, which takes the time-signal, and performs Fourier Analysis on the signal, resulting in a joint time-frequency representation of the tone. After AFI, and an *automatic melt process* (AMP), the resulting representation is decoded, and returned as a time-signal of the hybrid tone.

3.1 Spectral and Temporal Feature Identification

The issues involved in 'error free' feature detection that require care include the spectral identification of partials, and the temporal identification of salient points, such as the onset and decay of harmonics. Once the features have been identified, *matching* is necessary so that we know which features of the first original sound correspond to features of the second. Consider say, the decay time of a piano tone, which far exceeds that of a violin tone. The features - the decay of all the harmonics - are temporal features and would be numbered T_1, T_2, T_3...T_n as the first *n* temporal features in each respective representation. These temporal features then require *time- warping* to correctly match their respective duration's.

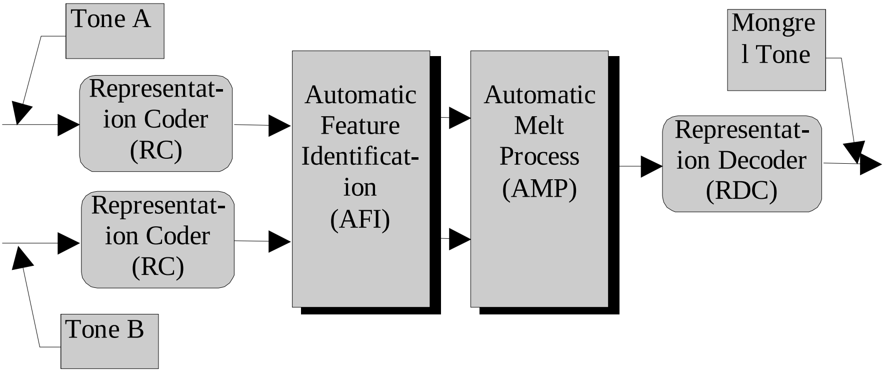
Tone A
Representat- ion Coder (RC)
Representat- ion Coder (RC)
Tone B
Automatic Feature Identificat- ion (AFI)
Automatic Melt Process (AMP)
Mongre l Tone
Representat- ion Decoder (RDC)

Figure 2 – The diagrammatic version of the operation of the morphing algorithm in the Audio Morphing Package

Spectrally, the harmonic component frequencies need to be identified - assuming we are dealing with harmonic spectra. To locate these frequencies, the spectrum is scanned for peaks at every temporal location in the time- frequency representation. The peaks must be checked at every temporal location, as the harmonics will vary in amplitude and frequency location temporally. This harmonic search algorithm, defined by Hope and Furlong (Hope and Furlong 1998) has been found to work successfully and efficiently for FFT lengths greater than 128 and this conclusion is authenticated by the graphs in Figure 3.

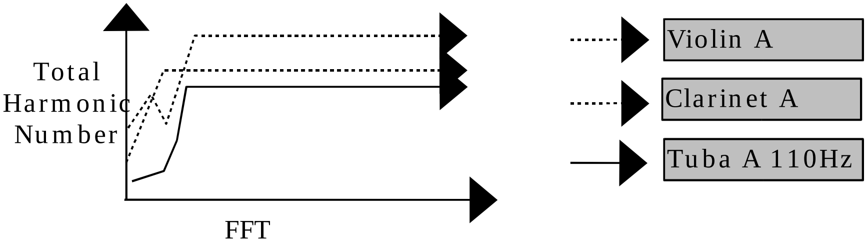

Figure 3 – Operational test results regarding the AFIA as a harmonic identifier, for varied FFT lengths on a number of tones

After locating the spectral features from the signal, the next thing that must be done before the AFIA can be considered complete, is to locate the temporal features that are being considered perceptually significant, and hence the only ones being sought in the AFI. These are the *Onset*, the *Decay*, and hence the *Steady State* which is sandwiched between them, as seen in Figure 1. This process is again outlined by Hope and Furlong (Hope and Furlong 1998).

3.2 Blending between two tones.

The final stage in the creation of a mongrel tone is to take the resulting matched features from the AFIA and *blend* or *melt* them into one another. The question arises as to which blend algorithm is perceptually optimal. However, this is not the primary concern in the creation of this AFIA. In Mongrel, the use of a linear interpolation for the blend is used. Whether the chosen joint time-frequency representation and interpolation procedure is acceptable or not in manufacturing a hybrid sound will be concluded using listener results. The most significant part of the linear interpolation is the method by which the matched features are warped into each other, be it linear or logarithmic or whatever. The whole process begins at the point of the user’s choice of content. The user specifies a ratio of tone A to tone B in the input of the interface to the Mongrel Software. This ratio is specifically for the *Automatic-Melt-Process* (AMP). This uniquely defines how far along the linear interpolation the AMP proceeds in operation.

Firstly, the spectral features, are correctly blended to their intermediate location. The new location is based on the ratio given of tone A to tone B. Finally, the temporal features are blended and time-warped on a linear scale to match each other at the correct intermediate specification. Once this new *Intermediate* joint time-frequency representation is realised it is forwarded to the *representation decoder* (RDC) or *TF-distribution Inverter*, which generates the inverse transform, resulting in a re-synthesised sonic signal which is the newly generated *Mongrel Tone*. The Mongrel package has been shown to operate efficiently and accurately in the generation of new mongrel tones. The quality of these mongrel tones is discussed and analysed through results and conclusions in section 4.

4 Results and conclusions

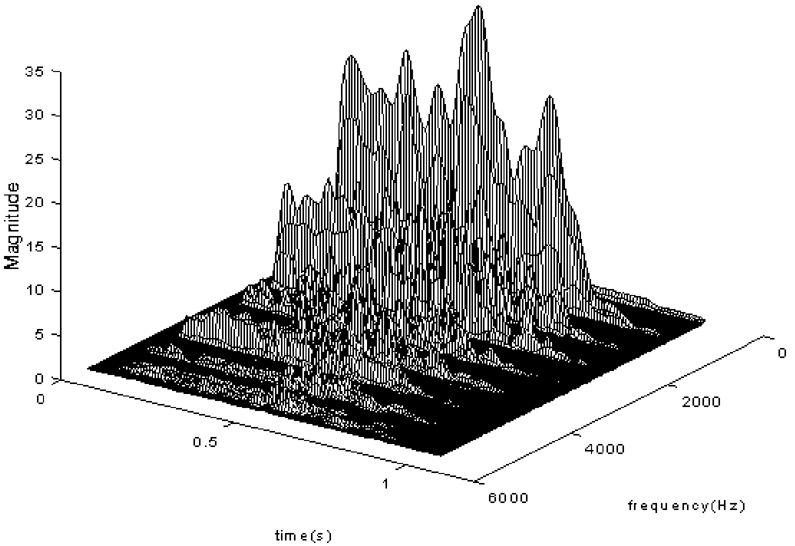

Figure 4 – A Morphed Tone: containing 50% Clarinet (pitch 660Hz) and 50% Violin (pitch 660Hz), duration 1 sec, sampled at 11025Hz.

Above in Figure 4, we can see an example of a mongrel tone which contains 50% of each of the two original tones. The morph was created using the Mongrel package. It must be remembered however, that what can be seen as alterations are not always the most perceptually identifiable changes, so to ensure that the morph is in fact approaching a compromise between the two parent tones, a sample test on listeners .is more appropriate. To carry out the tests, a number of morphs were achieved using several different orchestral instrument samples. Among those used were oboe, clarinet, violin, saxophone, piano, French horn, trombone, tuba and in the case of a composition - *Requiem Introit* - a child's voice and an organ were also used. Three different sets of listener tests were done using the resulting morphed samples.

Firstly, listeners were invited to listen to 32 morphed sounds and suggest the name of the two sounds morphed together and the percentage of each instrument's timbre in the new sound. Secondly, listeners were asked to listen to twelve sounds, this time with the names of the instruments given. What was required from the listener was to suggest, as accurately as possible, the percentage of each instrument in the pairing. Thirdly, listeners were let listen to 4 morphed sounds, each containing three instruments. Two of the instruments from each triple had first been morphed together evenly, and then morphed a certain amount with the third sound. Listeners were asked to try and identify the three sounds, and the groupings.

The results of these tests lead to several conclusions. Out of all the instruments used in the testing, the violin and piano were the most widely recognised. This was due to the strength of some of their characteristics, for example the onset of the piano tones was stronger than the strength of any other onset in the sample group. Even when the instrument was nearly totally suppressed with its morphing partner, it was still ever so slightly identifiable due to the strong imbalance in onset. A good example of this would be the morph of piano and oboe, where the oboe at 90% of the maximum representation in the paring, could still not suppress the onset of the piano, due in no small way to its own feeble onset.

Listeners had difficulty identifying the instrument when the saxophone and trombone were morphed together. When the non-musically trained listeners are isolated, it is noted that they had a lower recognition level for most of the instruments when they were morphed together, this holding true for the 'easier' instruments such as piano and violin as well. The most interesting morph from the orchestral instruments, has to be the morph of clarinet and oboe, which at a certain level generates a timbre that is almost inseparable to that produced by a saxophone.

Future work on automatic feature identification should involve further correlations of listener test results with the methods employed in the morphing process. This will continue to ensure that ‘perceptively’ optimal methods are being sought and hopefully result in significant advances in the quest for the optimal hybrid sound.

5 References

Holloway B., Tellman E., Haken L. (1995). Timbre Morphing of Sounds with Unequal Numbers of Features. *Journal of the Audio Engineering Society,* 43 (09), 678-689 (September 1995).

Slaney M., Covell M., Lassiter B. (1996). Automatic Audio Morphing. *Proc. ICASSP*. (IEEE.Atlanta, GA, May 7-10, 1996).

Walker B., Fitz K. (1996). *Lemur*. (University of Illinois, Urbana, IL 61801, USA).

Stumpf C. (1926). *Die Spracblante*. Berlin and New York: Springer-Verlang.

Furlong D.J., Hope C.J. (1997). Time-Frequency Distributions for Timbre Morphing: The Wigner Distribution versus the STFT. *Proc. SBCMIV* (Brasilia, Brasil, August 1997).

Furlong C.J., Hope C.J. (1998). Endemic Problems in Timbre Morphing Processes: Causes and Cures. *Proc. Irish Signals and Systems Conference 98*, 10-16 (DIT, Dublin, Ireland, June 1998).