ABSTRACT

This paper describes an algorithm which allows singing to be analysed in real time using a PC and then re-synthesised by the computer using whistled notes. The singing can also be transcribed as a series of notes on a musical stave using a MIDI file as interface. Pitch amplitude and spectral change parameters are derived from the input waveform. A sequence of musical notes is derived from a set of parameters using a set of rules. The system is designed as an entertaining, yet educational tool for children, and will be embodied in an interactive multi-media system. In its electronic form the paper has attached files demonstrating the results of the re-synthesis algorithm.

1. INTRODUCTION

One of the goals of interactive multi-media is to increase the range of modalities with which the user can interact with the computer. The use of speech recognition for computer control and the transcription has been a goal of the computer industry for many years. The use of computers for musical applications has been less widely researched. However the power of the present generation of personal computers allows the sophisticated analysis of the speech waveform. This paper describes the development of pitch and spectral estimation algorithms which allow singing to be analysed in real time using a PC and then re-synthesised by the computer using whistled notes. The singing can also be transcribed as a series of notes on a musical stave using a MIDI file as interface. A particular application of the system is as an entertaining, yet educational tool for children. Other applications include the teaching of singing and as an aid to musical composition.

Three parameters must be estimated for the re-synthesis function. They must each be performed in real time at the chosen frame rate of 20ms:

1. Estimate of pitch period.
2. Estimate of the degree of voicing, periodicity, of the speech.
3. Estimate of speech energy.
4. Estimate of the extent of change in spectral profile of the speech.

In section two the pitch algorithm and the periodicity derived from it are outlined. Section three presents the spectral change algorithm. Section four and five describe the re-synthesis technique and the transcription algorithm while section six illustrates some experimental results.

2. THE PITCH ESTIMATION ALGORITHM

This application demands a pitch estimation algorithm which is simultaneously reliable, accurate and yet computationally simple so that it can be implemented in real time using the host processor of a typical multi-media PC. The AMDF algorithm with certain refinements was chosen for this task since other algorithms based upon spectral analysis or autocorrelation [1] all appeared to be much more computationally intensive.

The AMDF algorithm [2] operates by measuring the periodicities of the speech waveform over different test pitch periods. Two adjacent sequences of speech samples are drawn from the speech, each sequence containing N samples. The distance between the two sequences is then measured, typically as a city block distance. The distance is normalised to the total energy in the two sequences and the result is defined as the periodicity at period N, at time index n, $P_n(N)$. The periodicity values range from 0 to 1 with 0 corresponding to a perfectly periodic segment of waveform. A search over a range of period values N yields the pitch period estimated as the value of N associated with lowest periodicity value. The estimate of the speech energy is produced as a by-product of the algorithm.
AMDF algorithm uses the city block distance between the
Discrimination of periodicity function. The standard
refinements must be made:

In practice this algorithm is neither sufficiently fast nor
sufficiently accurate for the target application and several
refinements must be made:

Discrimination of periodicity function. The standard
AMDF algorithm uses the city block distance between the
two segments of waveform. The resulting periodicity
function exhibits a rather shallow minimum offering insufficient sharpness, and hence pitch estimate resolution.
The sharpness of the minimum is increased significantly by
using the Euclidean distance and redefining the periodicity
function as:

\[
P_n(N) = \frac{\sum_{i=1}^{N} |s_{n-i} - s_{N+n-i}|}{\sum_{i=1}^{N} |s_{n-i}| + |s_{N+n-i}|} \quad \ldots 1
\]

\[
P_n(N) = 2 \frac{\sum_{i=1}^{N} (s_{n-i} - s_{N+n-i})^2}{\sum_{i=1}^{N} s_{n-i}^2 + s_{N+n-i}^2} \quad \ldots 2
\]

Pitch halving and doubling. The AMDF algorithm is prone
to pitch halving and doubling errors because the periodicity
function is itself periodic in the true pitch period. These
errors are catastrophic in the singing application because they lead to sudden octave shifts in the re-synthesised tune.
A simple approach to dealing with this problem is to
identify candidate pitch period harmonic values and choose
the candidate which has the lowest periodicity and which is
also consistent with pitch values estimated in immediately preceding frames. The rationale being that it is not possible
for a singer to make a full octave note change within the
duration of a few frames.

Pitch tracking. In spite of the simplicity of the periodicity
function used as the basis for pitch estimation, the
computational burden can be high because the function
must be evaluated for each of the possible pitch period
values. This may amount to several hundred separate
evaluations, making the algorithm unsuitable for real time
implementation. A satisfactory solution has been found to
use the AMDF algorithm within a pitch tracking framework
which enables the pitch period search to be restricted to a
small number of values around the estimate of pitch period
for the previous frame. This strategy works well during
periods of highly voiced and high energy speech, but fails
after the speech has been unvoiced or has low energy. A
simple yet effective solution is to control the range of the
pitch period search using the previous frame’s best
periodicity value. Typically it is found that a search range
of five sample periods around the previous pitch period for
periodicity value near zero, to a hundred sample periods
around the previous pitch period for periodicity values near
unity, provides a good compromise between search speed
and reliability of pitch estimate.

3. SPECTRAL CHANGE ESTIMATION

The perception of division of a tune into distinct notes is
frequently achieved by the singer changing the sound which
is being sung even though the pitch or energy may not change significantly. An example of this is in the “ee-ii-ee-
ii-o” phrase of the song ‘Old MacDonald had a Farm’. This
effect cannot be directly achieved in the re-synthesised tune
because it is expressed using only one sound such as a
whistle. An effective solution is to make the variation in
amplitude of the re-synthesised tune depend not only on the
variation in amplitude of the original singing, but also on
the degree of voicing and observed changes in spectral
profile of the original sung sound.

The method requires that a single scalar be evaluated
whose magnitude lies in the range 0 to 1 which represents
the extent of spectral change between successive frames of
the original singing. The method used in the target
application is to evaluate the modulus of the scalar dot
product \( S \) of the vector descriptions of the spectrum of the
current and previous frames, \( \mathbf{F}_{n-1} \) and \( \mathbf{F}_n \) respectively.

\[
S = \frac{\mathbf{F}_{n-1} \mathbf{F}_n^*}{|\mathbf{F}_{n-1}| |\mathbf{F}_n|} \quad \ldots 3
\]

In practice, evaluation of the full resolution power
spectrum of the sung speech on a frame by frame basis
using the FFT is both unnecessary and computationally
impractical given the required real time operation of the
system. A satisfactory approach offering sufficient speed
and resolution is simply to analyse the sung speech using a
bank of four second order filters whose centre frequencies
are chosen carefully in the range 1000Hz to 3000Hz, so that
change of sound generates a significant change in the four
dimensional spectral vectors.

4. RE-SYNTHESIS OF SINGING

The re-synthesised singing may be performed using an
arbitrary sound such as a sinusoidal whistle, or any other
continuously sounded musical instrument. The pitch of the
synthesised singing is controlled directly by a median
smoothed sequence of the pitch estimate values whilst the
amplitude is controlled both by the energy of the original
singing, changes in its spectral profile, and the degree of
voicing in the speech. Amplitude control using the voicing
parameter is required to ensure that unvoiced sounds of
significant energy are not used to generate a note which
would have an inappropriate pitch. The spectral change
parameter is used to deal with singing in which pitch does
not change from note to note in the note sequence, and in
which the sequence of notes is not punctuated by low
energy intervals.

The exact method of using each of the speech parameters to
control the amplitude and pitch of the re-synthesised
singing has been found by experiment. As an example, the
expression for the time domain synthesised waveform for
whistled sinusoidal notes is shown in equation 4a, 4b and
4c. Equation 4a evaluates the appropriate increment in
phase of the sinewave between sample instants using the
current pitch period estimate $\tau_p(n)$. 4b evaluates the
current total phase and 4c scales the sinewave by the
current energy $E_n$, voicing $V_n$ and spectral change term
$S_n$. The voicing and spectral change terms are defined to
be in the range 0 to 1.

$$\Delta \phi(n) = 2\pi f_0 / \tau_p(n) \quad \ldots 4a$$
$$\phi(n) = \sum_{i=-\infty}^{\infty} \Delta \phi(i) \quad \ldots 4b$$
$$s(n) = E_n V_n S_n \cos(\phi(n)) \quad \ldots 4c$$

Each of the control parameters used in the equations above
are specified on a sample by sample basis even though they
are only estimated at the frame rate. The sample rate
values are simply obtained by interpolation between
successive frame estimate values.

5. NOTE ENDPOINT DETECTION FOR MIDI ORCHESTRATOR

One of the educational objectives of the system is to allow
the notes of sung speech to be displayed on a musical stave
so that the user can see immediately what they have sung
and modify their singing as necessary. The widely used
MIDI Orchestrator software is designed to perform just this
function but requires the music to be expressed in a MIDI
file format in which the pitch, loudness and timing of each
note are appropriately encoded. The problem is therefore
to find an algorithm which is able to convert the parameters
extracted from the sung speech into note start and stop
timings along with appropriate amplitude and pitch values.
To maintain the interactive and real time operation of the
system, this process of note endpoint detection must be
performed in a single pass without the delay which would
be incurred by waiting for a global view of the entire
sequence of sung speech parameters. The approach
adopted is to use a series of rules based upon the values of
sung speech energy, periodicity, spectral change and pitch
change to decide when a note has started and when it stops.
The rules need be rather complex to deal with the
characteristics of real sung speech but some simplified
examples are shown below to illustrate the process:

A note starts if: There is no note currently active AND the
speech energy is greater than a threshold AND the periodicity is greater than a threshold.

A note stops if: A note is currently active AND the speech energy is less than a threshold
OR
the speech periodicity is less than a threshold
OR
the speech pitch period has changed by more than one
semitone.

6. EXPERIMENTAL RESULTS

The operation of the pitch estimation and note detection
algorithms are illustrated by the set of contours shown
below which were generated by a woman singing “jingle
bells - jingle bells - jingle all the way”. Fig. 1a, 1b, 1c, and
1d show the raw pitch frequency estimate, periodicity
value, energy, and spectral change values for the sung
phrase. The pitch frequency is expressed as a musical
semitone value. It can be seen that there are high values of
pitch periodicity corresponding to each of the notes which
are sung. The energy follows a similar pattern, but has
broader peaks, which cover parts of each note which are not
highly voiced. The usefulness of the spectral change
contour is shown by the presence of distinct peaks near the
end of the sung phrase “all-the-way” where neither the
periodicity or energy contours show distinct note start and
end points. The information from each of these contours is
combined using the note end-point detection rules and
yields the musical note contour shown in Fig. 1e. Finally,
the timing of the notes is shown in Fig. If in which the
pitch contour has been punctuated with zero values in
intervals between the notes detected by the algorithm so
that the duration of each note is evident.

The sound files attached to this paper demonstrate the
operation of the system. A0259S01.WAV is a recording of
the original sung phrase, A0259S02.WAV is a re-synthesis
of the phrase using the contours in figures 1a to 1d, and
A0259S01.WAV is the re-synthesis of the phrase using the
Midi Orchestrator strings, based upon the note interval and
pitch contour in Fig. 1f. [sound A0259S01.WAV, A0259S02.WAV, A0259S03.WAV]

REFERENCES

Figure 1a Pitch Contour

Figure 1b Periodicity Contour

Figure 1c Energy Contour

Figure 1d Spectral Change Contour

Figure 1e Note Contour

Figure 1f Note Intervals and Pitch