Proceedings of the Institute of Acoustics

VOICE PARAMETER ESTIMATION FOR SOUND PROCESSING CONTROL

I.S. Gibson, D.M.Howard

Parallel and Signal Processing Applications Group,
Department of Electronics, University of York, Heslington, York YO1 5DD. UK.
(isg100@unix.york.ac.uk)

1. ABSTRACT

Computers have made it possible to analyse and estimate parameters from the voice using digital storage and signal processing techniques. The processing time to manipulate sounds digitally is decreasing rapidly as technology advances. This increases the potential of the voice (with its wide range of expression) to act as a controller for computer music systems. This paper describes a system which has been implemented to control synthesized music through the use of voice in conjunction with a graphical user interface. The voice is used to build amplitude and pitch envelopes to control a synthesized sound and filters are available, whose envelopes are controlled by mapping certain parameters obtained from a Linear Predictive Coding model of the vocal tract. A graphical interface enables parameters relating to panning, reverberation, and overall amplitude of the sound to be entered.

2. BACKGROUND

There are a variety of methods and techniques for producing and representing electronic computer music. Music itself, whilst being an art form, contains elements which can be expressed mathematically (such as rhythm and harmony), as well as containing more abstract elements such as timbre. Computer music requires the composer to specify all of these parameters. Once specified, the composer can control and refine the resulting music through a process of experimentation [Dannenberg, 1993].

The Musical Instrument Digital Interface (MIDI) has been adopted by most manufacturers of computer-based music instrument hardware as the standard protocol for communication between devices. However, MIDI was primarily developed for use with musical keyboards. There are a number of disadvantages to this approach of producing music. They include difficulty in the production of microtonal music and of fine timbral control in live performance.

Music Computer Languages offer the ability to control synthesized sounds completely. However, these languages are usually sequential in nature, and are not always intuitive to use. Often the programmer must be the performer, the composer and the instrument designer; tasks which would be tackled by different people in a traditional composition environment.

When designing a new system for producing electronic music in real-time the following issues should be addressed:

- an unambiguous and clear user interface,
- item an intuitive user interface, and
- simultaneous effective control of several musical parameters.

The MIDI keyboard provides only a limited number of control parameters in real-time. Typically, these are as follows:

- pitch,
- velocity,
- after-touch,
- pitch bend, and
- modulation.

It would be useful to explore other methods of capturing the data needed to control synthesized sounds. The voice is able to change in pitch, amplitude and timbre. Everyone who can speak has some control over these parameters, and professional singers have a great deal of control over them. Using the voice is a skill which most people make use of every day. Therefore a system using voice input would be instantly accessible to many without the need to learn new skills, given an intuitive user interface. It could even be used to encourage people to practice and improve their voice skills.

This paper is concerned with the creative mapping of voice parameters on to synthesized sound parameters. The aim is to achieve a system which will be intuitive and natural in its use. The main problem is the processing of parameters in real-time but there are a number of problems which must be addressed. Fundamental frequency estimaion routines must wait for at leasta complete waveform cycle to be captured before processing a result. Speech waveforms are spectrally complex. While this means that the voice is a rich source of time-varying parameters, the processing power required to estimate them is significant.

3. ANALYSIS AND SYNTHESIS SYSTEMS

Three systems have been developed for purposes of voice analysis and sound synthesis.They were implemented on an Indigo (SGI) Iris 4D workstation and written in C.

3.1 TIMBRE ANALYSIS SYSTEM

This section describes a system built for the purpose of timbre analysis. It was designed with the following objectives:

- to allow recording and playback of standard AIFF format audio files,

- to produce a spectrogram from a soundfile and
- to analyse the timbre of a sound.

A soundfile is selected or recorded, and subsequently processed by the system. Amplitude is normalised, and an FFT is applied. The resulting spectrogram has 512 bands. A set of attributes based upon those suggested by [Grey, 1976] are displayed graphically on the spectrogram and also numerically. They are as follows:

- the bandwidth of the sample (A),
- the difference (in milliseconds) between the highest and lowest
- frequency onset, (B) and
- the amount of low amplitude, high frequency energy in the attack segment (C).

The sample bandwidth is calculated by taking the difference between the highest and lowest frequencies present whose strengths are above a user-defined threshold level. As the threshold is increased so only the strongest frequency levels are displayed.

In order to establish the amount of low amplitude, high frequency energy in the attack segment the user must define the following parameters:

- the attack portion of the sample,
- the area of high frequency in the sample and
- the threshold of the low frequency level used for analysis.

Six consonant sounds were processed using this system. Each consonant had an /i/ vowel following it. The length of the attack period for each consonant was set at 80 milliseconds. High frequency, low amplitude energy was observed in the region 6003 Hz to 10.4 khz. The results are shown in table 1.

Consonant	A	B	C
/t/	3196	112.6	6.8
/d/	11817	56.3	3.0
/k/	13226	8.0	7.2
/g/	3729	176.9	0
/p/	5530	193.0	0
/b/	5369	193.0	0

Table 1. Results showing sample bandwidth, difference in highest and lowest frequency onset, and low amplitude high frequency energy.

Using this method of timbral classification (A,B,C above) the timbral differences between each consonant are clear . It is anticipated that this system will be expanded and used to further the research of mapping voice tc .ynthesized sounds. This would probably involve mapping timbral characteristics of voice (estimated using the criterion above) to produce a similar timbral characteristic in the synthesized sound.

3.2 LPC ANALYSIS SYSTEM

This system allows voice samples to be analysed, either in real time or non-realtime using linear predictive coding (LPC). LPC works on the principle that a sample may be calculated from a series of delayed samples. This results in each sample *s(n)* being predicted from a series of past samples *s(n-1) to s(n-k)* each of which are multiplied by a corresponding filter coefficient. LPC uses one of the following algorithms:

- covariance,
- autocorrelation and
- lattice.

The covariance method can provide the most accurate results but the filter coefficients can become unstable. It is best used in cases where the waveform is pitch synchronous. The autocorrelation method and lattice method have been adopted for this research. The former can be used to calculate reflection coefficients from filter coefficients whereas the latter returns reflection coefficients directly.

The advantages of using LPC are numerous. It bridges the gap between the time and frequnecy domain. It is equally valuable for both storage of speech and synthesis of speech. The prosodic information is separated from the segmental information. It can deal more rationally with aperiodic waveforms because it works in the time domain.

Vowel sounds have been analysed with this system, each of them producing different and distinct patterns of filter co-efficients. Two examples are shown in figures one and two..

VOICE PARAMETER ESTIMATION FOR SOUND PROCESSING CONTROL

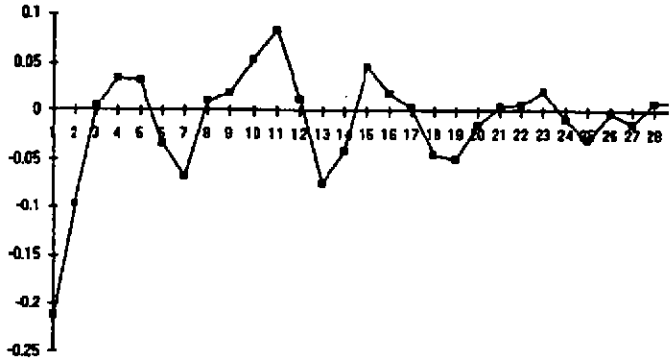

Figure 1. LPC filter co-efficients for /I/

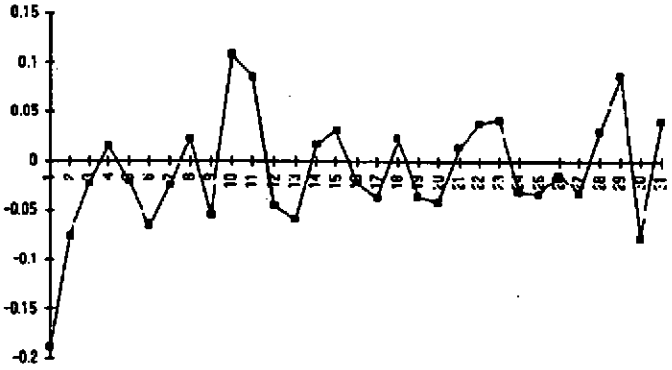

Figure 2. LPC filter co-efficients for /e/

A desirable quality of any music producing system would be the ability to morph between two synthesized sounds as the performer changes between vowel sounds. As an initial investigation the system could be trained to recognise two distinct vowel sounds and to detect changes from one to the other.

VOICE PARAMETER ESTIMATION FOR SOUND PROCESSING CONTROL

3.3 ANALYSIS/SYNTHESIS SYSTEM

This system allows manipulation of sound parameters using graphics and voice control. Sound generation is handled for the present the programming language CSound [Vercoe 1986].

The system generates source code for the language. Sound icons are created using voice input. Voice is used to control the following sound envelopes:

- fundamental frequency,
- high-pass filter,
- low-pass filter and
- amplitude.

Fundamental frequency estimation is achieved with the use of the electrolaryngograph [Abberton et al, 1989]. This device relies on the ability of human tissue to conduct high frequency electricity. It measures the changes in impedance across the vocal folds. This is achieved with the use of 2 electrodes placed across the surface of the neck. The fundamental frequency is obtained by measuring the rate at which the vocal folds vibrate. A particular advantage of the electrolaryngograph is that it does not suffer from the effects of external noise. Also, it is unaffected by interference from the vocal tract.

Software used with the Electrolaryngograph measures the closed quotient (CQ) time [Davies et al, 1986]. There are closed phase (CP) and open phase (OP) portions of the waveform. OP begins at a point roughly corresponding to 3/7 ths of the cycle's peak to peak amplitude.

Some singers find wearing the electrolaryngograph electrodes slightly uncomfortable. However, the resulting signal obtained is fairly reliable and accurate providing the electrodes are positioned to give a maximum amplitude electrolaryngograph output waveform (Lx). If they are placed at the exact level of the vocal folds then the signal to noise ratio is optimised. Some resistance is encountered from the skin which may be minimised by ensuring that the electrodes are fixed firmly in place.

Amplitude estimation is achieved by measuring peak-to-peak values of waveforms sampled through the audio port. These values are used to produce amplitude envelopes in Csound.

Filter information estimation is achieved with the use of a linear predictive coding model of the vocal tract. The open back vowel sound 'ae' (as in 'bad') is used to open the filter. This is achieved by mapping a co-efficient to the filter cut-off frequency.

Score and Orchestra files are produced for use with CSound (figure 7). A standard AIFF digital audio sound file is generated and played directly through the SGI Workstation speaker (or headphone) port.

```
instr 1
kpanenv linseg 0.5, 0.3000, 0.5053, 0.3000, 0.5105, 0.3000, 0.5000,
0.3000, 0.5000, 0.3000, 0.5000, 0.3000, 0.5105, 0.3000, 0.4842, 0.3000
kampenv expseg 5000, 0.3000, 2530.0000, 0.3000, 2660.0000,
0.3000, 2556.0000, 0.3000, 2530.0000, 0.3000, 2530.0000, 0.3000,
2530.0000, 0.3000, 2530.0000, 0.3000
krevenv linseg 0, 0.3000, 0.0000, 0.3000, 0.0000, 0.3000, 0.0000,
0.3000, 0.0000, 0.3000, 0.0000, 0.3000, 0.0000, 0.3000, 0.0000, 0.3000,
kpitenv linseg 74,0.2,74,0.2,74,0.2,84,0.2,84,0.2,95,0.2,95,0.2,114,0.2,
114,0.2,133,0.2,133,0.2,158,0.2,158,0.2,160,0.2
klocampenv linseg 0,0.2,10000,0.2,7543,0.2,6415,0.2,5141,0.2,5068,0.2,
4047,0.2,3139,0.2,2548,0.2,1821,0.2,1235,0.2,688,0.2,868,0.2,72
klpenv linseg 24000, 3.0, 24000
khpenv linseg 0, 3.0, 0
a1 oscil kampenv+klocampenv,(p4+(p4*(kpitenv/50))),1
a1 reson a1, p4, klpenv, 1
a1 areson a1, p4, khpenv, 1
aout1 = (a1*(1-kpanenv))
aout2 = (a1*(kpanenv))
aout3   reverb aout1, 5
aout4   reverb aout2, 5
aout5 = aout1+(aout3*krevenv)
aout6 = aout2+(aout4*krevenv)
aout5 balance aout5, aout1
aout6 balance aout6, aout2
outs aout5,aout6
endin
```

Figure 7. CSound orchestra file.

4. SUMMARY

The on-going research outlined in this paper is directed at making available voice controlled music synthesis systems for singers of all abilities. Voice parameter estimation is possible in

real-time but more difficult to accomplish in conjunction with real-time sample generation. Further research will be directed involving the use of MIDI in order to free processor time for voice analysis.

5. ACKNOWLEDGEMENTS

This work is funded by the EPSRC grant reference 93314618.

6. REFERENCES

[Abberton et al, 1989] ABBERTON E., HOWARD D., & FOURCIN A.(1989). Laryngograph Assessment of Normal Voice: A Tutorial. *Clinical Linguistics and Phonetics*, 3, (3), 281-296.

[Clark, 1992] CLARK J. & YALLOP C. (1992). *Phonetics and Phonology.* Massachusetts, USA: Blackwell Publishers.

[Dannenberg, 1993] DANNENBERG R.B. (1993). Music Representation Issues, Techniques, and Systems. *Computer Music Journal*, 17, (3), 20-30.

[Davies et al, 1986] DAVIS P., LINDSEY G.A., FULTTER M. & FOURCIN A.J. (1989). Variation in Glottal Open and Closed Phase for Speakers of English. *Proceedings of the Institute of Acoustics*, 8, 539 - 546.

[Grey ,1976] GREY J. (1976). Multidimensional Perceptual Scaling of Musical Timbres. *Journal of Acoustical Society of America*, 61, (5), 1270-1276.

[Howard and Lindsey, 1987] HOWARD D.M. & LINDSEY G.A. (1987). New Laryngograms of the Singing Voice. *Proceedings of the 11th International Congress of Phonetic Science (USSR:Tallin)*, 36, (3), 406-407.

[Vercoe 1986] VERCOE, B.L. (1986). *The Csound Manual.* Cambridge Massachussets: Experimental Music Studio, Media Laboratory, MIT, Cambridge.