# Technology For Real-Time Visual Feedback In Singing Lessons

David M. Howard

*Research Studies in Music Education* 2005 24: 40
DOI: 10.1177/1321103X050240010401

# Technology For Real-Time Visual Feedback In Singing Lessons

David M. Howard

## Abstract

Computer technology has advanced to the point where audio analysis techniques, previously the hallowed preserve of speech science and engineering research laboratories, are now available to anyone who makes use of a modern PC machine in the home or office. Much of this analysis can be carried out in real-time. Thus it is currently perfectly feasible to make available real-time visual feedback in singing lessons, and a number of systems are now available. There is a considerable body of (well-established) literature describing quantifiable changes that occur when the singing voice is trained in a context within which visual feedback can be offered. This paper discusses voice production and those elements that can be quantified using today's multimedia PC machines, the use of real-time visual feedback in singing lessons and the importance of some understanding of the principles behind typical analysis algorithms employed in such systems. It is shown that real-time displays are useful in singing lessons. Displays should be clear and unambiguous and users must be aware of what is expected of them. User-controlled parameters should be altered with caution with an informed expectation of what the output might look like and the impact of any errors that are likely to occur.

The advent of the personal computer has stimulated rapid advances in both hardware and software to the point where today's machines have processing and storage capability which would only have been found on expensive large laboratory machines a quarter of a century ago. Recent developments that have made the acquisition, processing and storage of large amounts of audio and more latterly video material commonplace, primarily for entertainment purposes, mean that machines found in the office and at home are capable of carrying out significantly complex audio processing in real-time. Such audio analysis techniques were once the domain of engineering and specialist speech science laboratories only. Whilst real-time visual displays have been used for developing particular vocal skills in the past, for example in the speech therapy clinic, they have typically involved the use of complex and expensive external hardware interfaces to acquire and process the input data. Modern PC machines designed for audio and video entertainment incorporate all facilities required for the acquisition, processing and storage of audio data. Further, for all but the most demanding applications, external hardware is no longer a necessary requirement, although it might be desirable in some circumstances.

Awareness of the capabilities of modern computers has also grown to the point where many singers and singing teachers are beginning to seek out the possibilities of quantitative support for their training activities, for progress monitoring, to provide feedback during training and to deepen their understanding of the physiology and acoustics of singing. For many highly cost-effective opportunities exist for making progress in such endeavours, due to the quite extensive range of freeware available for download via the internet. There are, however, potential pitfalls in the area of audio analysis which could result in data analysis that is specious. The origin of these pitfalls lies within the research techniques themselves that are employed for carrying out the audio analysis. Most make use of quite complicated signal processing techniques which incorporate

controls with which their operation can be optimised for the particular input data being used. Setting such parameters to best effect requires a degree of knowledge of output expectation (KOE), often with respect to the algorithm being employed. Whilst familiarity with the operation of algorithms is rooted in science and engineering, and therefore typically quite a distance away from the previous experience of either singing teacher or student, KOE can be gained with experience and an intelligent approach to using the technology. This paper provides examples of how analysis parameters can alter the observed output, as well as pointers to how increased KOE can be gained.

Formal singing teaching is a process where expertise is handed down from teacher to student, generation by generation. As such, it is both an oral and aural tradition. It relies for its success on the aural ability of the teacher to analyse qualitatively the student's sung output, the oral ability of the teacher to describe what is needed to alter the student's sung output appropriately to enable improvement to be achieved, the student's aural ability to understand the teacher's instructions and the student's oral ability to put the instructions into practice. This is a process that is usually supported through the use of imagery (Moorcroft, 2002), where 'psychological hooks' or concepts designed to promote the use of postural gestures that are deemed to be appropriate for the production of a sung output are employed by the teacher when feedback is provided. Examples of such hooks from the author's own experience of having singing lessons include: "begin speaking as though your voice was flowing in a light blue colour, and as the words", or "sing as if smelling a rose", or "sing through an imaginary hole in the top of your head" or "focus the sound beam of energy further forwards/backwards in the mouth". Such psychological hooks have been used and passed down over many teacher-pupil generations, and they do indeed appear to work for many students in terms of enabling them to produce an appropriate vocal output.

There are potential points of weakness in this process. For a student to make progress, any explanation of how to alter behaviour must be clear, unambiguous and properly understood. The student must also have a clear idea of what constitutes improvement. One potential source of confusion is that teachers tend to make use of their own inventory of psychological hooks which are rooted in the tradition from which they emerged, and the student has to 'tune-in' to the underlying message or gesture. The analysis of the student's vocal output is qualitative, and unless some recording device is employed, progress cannot be readily reviewed in anything but an anecdotal way. Many students and teachers are starting to seek explanations that describe the physiological and scientific reality of singing, so that they can become much more in touch with the physical means to the singing end.

There is a wealth of data available now on the singing voice which is based on quantitative measurements (Sundberg, 1987; Titze, 1994) and this provides the basis for the application of a quantitative approach to monitoring changes in the vocal output of a singer. There is research centred on various quantifiable aspects of the singing and speaking voice and one outcome might be replacement and/or enhancement of the present rather diverse terminology, imagery and psychological hooks through the use of reliable, repeatable and recordable quantifiable measures of singing performance.

The availability of real-time audio recording, processing and storage on today's multimedia computers means that quantitative knowledge about the singing developmental process can now be applied during lessons. In addition, the use of real-time visual feedback provides additional advantage in the teaching process. For example, feedback has been described by Welch (1985) as having a key role in the learning process as follows. Traditionally, the teacher provides a target which the

student attempts to imitate, and the teacher provides feedback. Following the feedback, the student makes another attempt and so the process repeats. However, the provision of real-time visual feedback *during* the student's vocal response enables the effects of any modifications to be observed immediately and synchronously with the vocal response. Real-time visual feedback provision thereby enables the student to make another attempt straight away based on scrutiny of the feedback during the previous attempt, which gives an immediate indication as to where change is required.

The usefulness of real-time visual feedback in singing voice training has been demonstrated for: (a) primary school children developing note pitching skills (Howard & Welch, 1993; Welch, Howard & Rush, 1989) and (b) singers in training (Rossiter, Howard & De Costa, 1996; Thorpe, Callaghan & van Doorn, 1999; Nair, 1999; Welch et al., in press). These studies confirm that singing development can be enhanced with real-time quantitative feedback providing: (a) the information supplied is meaningful, valid and useful; (b) students can use it during their lessons with input from their teacher; (c) progress is monitored from lesson to lesson and (d) students use it to enhance their practice experience between lessons. Howard and Welch (2003) have pointed out, however, that such systems will never replace singing teachers for two key reasons.

1. It is often the case that a change can be made to a measured parameter by more than one means. For example, an increase in larynx closed quotient is associated with voice training (Howard, 1995), but it can also be increased through the use of a pressed phonation (Sundberg, 1987), which is a completely unsuitable and potentially damaging voice quality with which to sing.
2. There are aspects of vocal training where the judgment of another human is required and where a real-time display would not be appropriate such as: stagecraft, performing musically, working with accompanists, working with conductors, working with directors, communicating with the audience, gesture, posture, ornamentation, etc. The use of real-time displays would enable singing teachers to spend more lesson time for these essential and often somewhat neglected musical aspects of performance (Howard & Welch, 2003).

They further suggest that the application of such displays could also benefit all who experience vocal difficulties in everyday life, such as teachers, lecturers, public address announcers, politicians, media presenters, journalists, market traders, stockbrokers, tour guides, town criers, carers, health workers and parents.

This paper considers the basic mechanism of human voice production and data that can be quantified to track changes during singing training. It then considers technology that can be used to provide real-time visual displays in singing lessons and its usefulness, and discusses the potential implications of changing any user settings provided with the underlying algorithms in terms of the extent to which the resulting display can be altered and perhaps compromised (Howard, 2002).

## Human voice production

The human vocal instrument can be considered in terms of its acoustic function as three main elements: a power source, a sound source and sound modifiers. These can be related physiologically to the action of: the lungs, the vocal folds and the vocal tract respectively. Figure 1 illustrates these, alongside an equivalent mechanical model indicating their acoustic function. The double-ended

arrows signify those parts of the instrument that can be moved during voice production.

Air is drawn into the lungs by enlarging their space through muscular action, thereby lowering the lung pressure with respect to atmospheric pressure (equivalent to pulling the piston in Figure 1 downwards). Air is expelled by muscular contraction which shrinks the lungs, producing a lung pressure that is higher than atmospheric and causing air to flow outwards (equivalent to pushing the piston upwards). Singers learn to control their breathing to enable them to sing long notes and musical phrases in a controlled manner, often termed 'breath support'. In the western bel canto tradition, for example, particular emphasis is placed on keeping the upper chest, shoulders and neck relaxed with relatively low-lying larynx and shoulders.

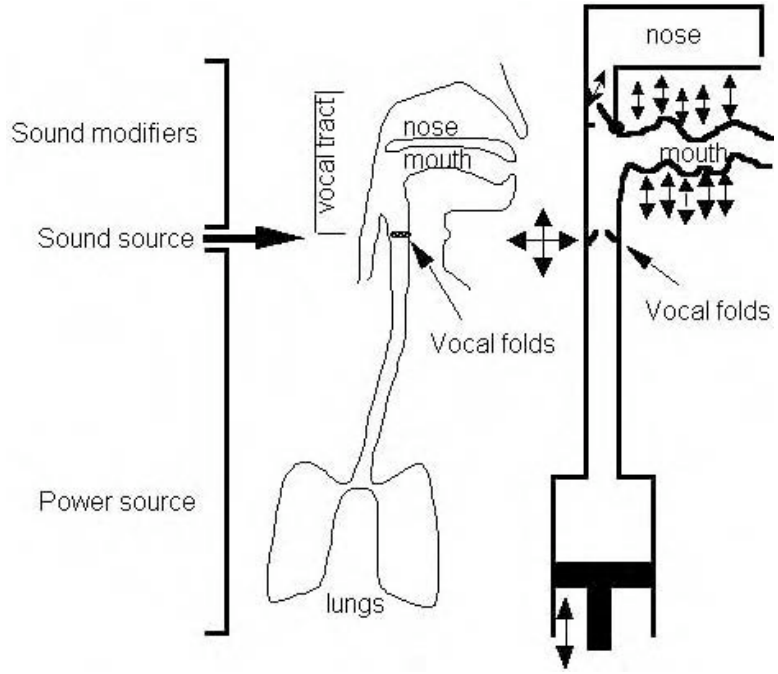


Figure 1
*The three key elements of the human vocal instrument. (Adapted from Howard & Angus, 1998.)*

The sound source during sung notes is the acoustic result of the vibration of the vocal folds within the larynx. The pitch of a note is related directly to the number of times per second that the vocal folds vibrate, known as the 'fundamental frequency' or '$F_0$'. The note A above middle C, which is used as the orchestral tuning reference, has an $F_0$ of 440 Hz. When a soprano sings this note, her vocal folds therefore vibrate 440 times a second. If she were to sing the A an octave higher, her vocal folds would vibrate at double this $F_0$, at 880 times a second, or 880 Hz. A professional singer will have a pitch range of around three octaves, whereas during conversational speech, the pitch range is around one octave. The note produced when the vocal folds vibrate is controlled in part by the tension and the thickness of the vocal folds themselves, a large proportion of which are muscle. Low notes are produced when the vocal folds are relaxed and thick. Stretching the vocal folds raises the pitch of the note up to a point where the folds can be stretched no further. Notes higher than this are produced by reducing the mass of the vibrating vocal folds through muscular action that holds a significant proportion of the vocal fold tissue

rigid, and reusing the same stretching mechanism to increase the pitch higher still. These 'voice breaks' or 'register breaks' are heard as tone colour variation as singers vary $F_0$ over a wide range (Sundberg, 1987; Titze, 1994). Key skills that must be developed by a budding professional singer are the development of a pitch range of around three octaves and essentially inaudible register changes.

The sound modifiers (see Figure 1), which comprise the mouth and nose airways between the larynx at one end and the lips and nostrils at the other, are known as the 'vocal tract'. The acoustic characteristics of the vocal tract change as the volume is altered, mainly by moving the tongue, jaw and lips, and this enables the different sounds that make up language to be produced. These acoustic characteristics manifest themselves primarily as a series of resonant peaks in the acoustic frequency response, known as 'formants', and the centre frequencies of the formants change as the articulators are moved (Sundberg, 1987; Howard, 1999). The first four formants are numbered from the lowest frequency resonant peak upwards as $F_1$, $F_2$, $F_3$ and $F_4$ (see left-hand illustration in Figure 2).

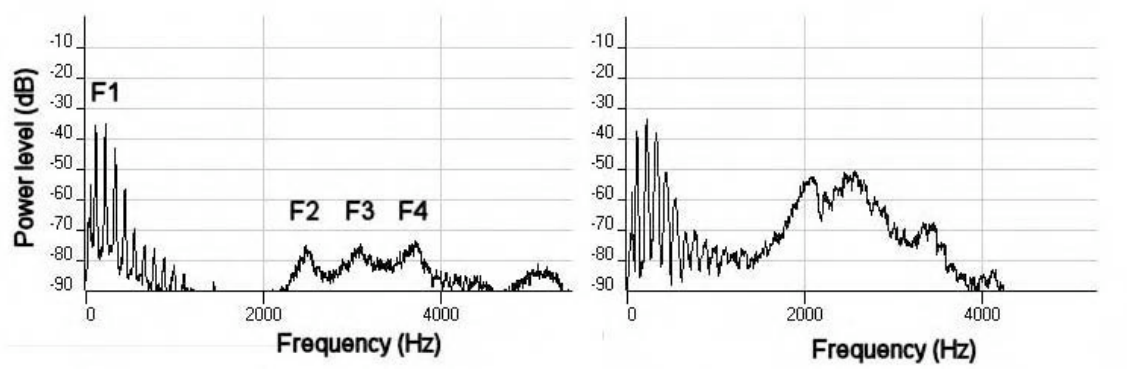


Figure 2
*Power spectra for the vowel /i:/ sung by an adult male in a non-projected style (left) and projected (western classical) style (right) at the same pitch*

The trained opera singer can be heard by an audience when singing on a large stage, accompanied by large orchestral forces without using any amplification device. This is achieved in practice by lowering the larynx and enlarging the pharynx; often achieved in the singing studio through the use of psychological hooks such as "singing on the point of the yawn", or "singing as if an orange is stuck in the throat". The acoustic result is a clustering together of the third, fourth and fifth formants into a composite formant peak referred to as the singer's 'formant cluster' (Sundberg, 1974; 1987). This is illustrated in the right-hand plot of Figure 2, where the singer's formant peak can be seen in the projected (western classical) production in contrast to the non-projected production (compare the left- and right-hand sides of the figure respectively).

## Acoustic analyses in vocal education

Analyses that are potentially useful in the context of vocal education are those that provide information to complement what the teacher is trying to achieve. In a project to investigate the application of technology in the singing studio, Howard et al. (2004) consulted a Liaison Panel of singing teachers, voice specialists, a linguist and a psychologist, to establish those areas where singing teaching might benefit from real-time visual displays. The following is a summary of essential

aspects which the Liaison Panel felt occurred frequently as pedagogical issues, and which were perceived as having the potential to benefit most through the use of a real-time visual display in the studio.

1. Voice quality (e.g. bright, mellow, breathy);
2. Consonants;
3. Vowel quality and length;
4. Pharyngeal widening;
5. Legato/staccato singing;
6. Registers;
7. Resonance;
8. Pitch accuracy;
9. Larynx position;
10. Tongue positioning for vowels;
11. Jaw position for pitch;
12. Head/neck alignment;
13. General posture;
14. Breathing.

The main techniques available from the speech science laboratory that have the potential to run in real-time on a standard multimedia PC machine and to provide some support for issues such as these include fundamental frequency measurement, spectral measurement to provide an estimation of formant frequencies and other resonant properties and spectrographic measurement to enable small- and large-scale timing differences to be evaluated. It is also possible to provide an estimate of the current area of the oral tract which has the potential to inform those identified issues that relate to tract shape, such as numbers 4 and 9-13 in the above list. Each is briefly summarised below.

$F_0$ estimation has been the subject of research since the 1920s and a large number of techniques have been developed. In practice, the choice of an $F_0$ estimation technique should be made with direct reference to the particular demands of, and acceptable errors that can be tolerated with, the intended application (Hess, 1983). For a real-time display of $F_0$ to be useful in the singing studio, it is especially important that some output is provided for beginners, even for notes that are tentative, weak in energy and breathy. Many $F_0$ estimation techniques would provide no output for such notes due to the way that they differentiate between clearly sung notes and background noise. A peak-picking technique, originally designed for use with cochlear implants (Howard, 1989), has been found to be appropriate in such situations, since it provides a compromise by not only encouraging those with tentative non-confident productions, but also by having good overall accuracy in its $F_0$ estimation.

Spectrum and spectrographic measurement are commonly applied techniques for acoustic analysis of speech and singing, enabling an examination of the frequency components present in a sound, either instantaneously (spectrum) or at a particular time (spectrograph) (e.g. Koenig, Dunn & Lacy, 1946; Potter, Kopp & Green, 1947; Fry, 1979; Borden & Harris, 1980; Baken, 1987; Baken & Daniloff, 1991; Rosen & Howell, 1991; Kent & Read, 1992). The spectrum is a measurement of the energy present in different frequencies at a given instant in time, and a spectrogram

consists of time and frequency plotted horizontally and vertically respectively, with increasing energy being indicated as a greater darkness of marking on a grey or colour scale. Knowledge of the nature of the frequency components present in the sung output enables the frequencies of the formants and the singer's formant to be estimated, as well as discussions of resonance strategies to be informed.

## Real-time visual displays

Modern multimedia personal computers are ubiquitous, and a considerable body of software is now available for the analysis of parameters relating to voice production; indeed, some excellent examples are available for download from the internet as PC freeware (WWW-1). Some of these operate in real-time and, therefore, could be employed during voice education and training sessions. Displays that are commonly found in such software include:

- the input microphone waveform against time;
- $F_0$ against time;
- a spectrogram;
- a power spectrum.

Software that is available specifically designed for use in singing lessons includes *SingandSee* (Thorpe, Callaghan & van Doorn, 1999; WWW-2), *VoceVista* (Nair, 1999; WWW-3) and *WinSingad* (Howard et al., 2004; WWW-4). All three provide real-time displays of the input microphone waveform against time, $F_0$ against time, a spectrogram and a spectrum, and all can write and read the input data to and from a file. The screen layout is completely different in each program, and each allows the display content to be altered by the user to suit the situation in which it is being used. *SingandSee* provides a variety of representations of $F_0$, including a graphical plot against time, a note on a musical stave, the note name and the note position on a piano keyboard. In addition, it provides an input level meter, as well as a fast (instantaneous) and slow (integrated) version of the spectrum. *VoceVista* offers a selection of screens which are different combinations of the displays, and it additionally offers a split screen to enable comparisons to be carried out between two separate sung inputs. This feature might, for example, be used to provide a student with a target display from the teacher. *WinSingad* offers other displays and a different screen design philosophy, and it is explored in more detail below.

The development of *WinSingad* follows directly from experience gained with the design of real-time visual feedback systems for both the development of pitching skills with children (Howard & Welch, 1989; 1993) and the ALBERT system (Rossiter & Howard, 1994). Its development in terms of how the information is presented (rather than which displays are implemented) was informed by a Liaison Panel brought together in the context of supporting the work of professional singing teachers in their singing studios as part of a project funded by the UK Arts and Humanities Research Board (AHRB) (Howard et al., 2004). These experiences have indicated that the choice of acoustic analyses and the design of any software environment for application in vocal training should be informed by consideration of at least the following features.

- The algorithms should be reliable and fit-for-purpose.
- The algorithms should need little or no user adjustment of controls.
- The output(s) should be plotted in real-time.[i]
- Displays should be clear and unambiguous.
- Help pages should give clear instructions for operating the software.

- Help pages should briefly and non-scientifically overview the display data.
- The overall screen display should be uncluttered.
- Use of the user controls should be intuitive.
- The link between the display and the taught activity should be clear.

*WinSingad* is a Windows application which has been developed using Microsoft Visual Studio C++. Each display is presented as a separate panel within the main *WinSingad* window, and the user has control of: whether a panel is visible or hidden; the vertical order in which the panels appear relative to each other; the colours used for the plot, background, axes and text and the thickness of the line used for the plot. Any user-controllable parameters relating to any processing algorithm are associated with the appropriate panel(s). The *WinSingad* system offers the following real-time displays which can be viewed on the screen singly or in any combination each in its own panel:

- Input acoustic pressure waveform against time;
- Fundamental frequency against time;
- Short-term spectrum;
- Narrow band spectrogram;
- Spectral ratio against time;
- Oral tract area;
- Mean/mid oral tract area against time;
- Real-time web camera window.

The spectral ratio analyses relating to the vocal tract area and the web camera are unique to *WinSingad* and an example screen showing just these outputs is shown in Figure 3. The spectral ratio is designed to provide a simple display relating to the energy in the singer's formant region. It is calculated as the ratio between the energy in the singer's formant frequency band (2.0 kHz to 4.0 kHz) and the complete spectrum (Rossiter & Howard, 1994), which constrains its value (essential for real-time plotting) to remain between 0 and 1. Rossiter and Howard (1994) have demonstrated that a spectral ratio calculated by this method increased for singers in training, and Rossiter, Howard and De Costa (1996) found that this ratio was also raised on average for all subjects taking a short period of vocal tuition to develop their speaking voices.

Analysis of the oral tract area of the vocal tract is possible through a technique known as linear prediction (Markel & Gray, 1976), which is widely used for speech analysis and coding. Markel and Gray discuss a technique which enables the output from a linear prediction analysis to be transformed into the cross-sectional areas of a series of equal length cylindrical tubes which acoustically account for the sound being analysed. The lowest panel in the *WinSingad* screen in Figure 3 shows an example oral tract display for the vowel in 'bells' from *There were bells* from 'The Music Man' (Willson, 1983) sung by a soprano. Around twenty tubes are required to model a 17.5 cm adult oral tract. Although the oral tract is not cylindrical, the approximation has been found to be useful in the context of a real-time visual display for voice training (Rossiter, Howard & Downes, 1995).

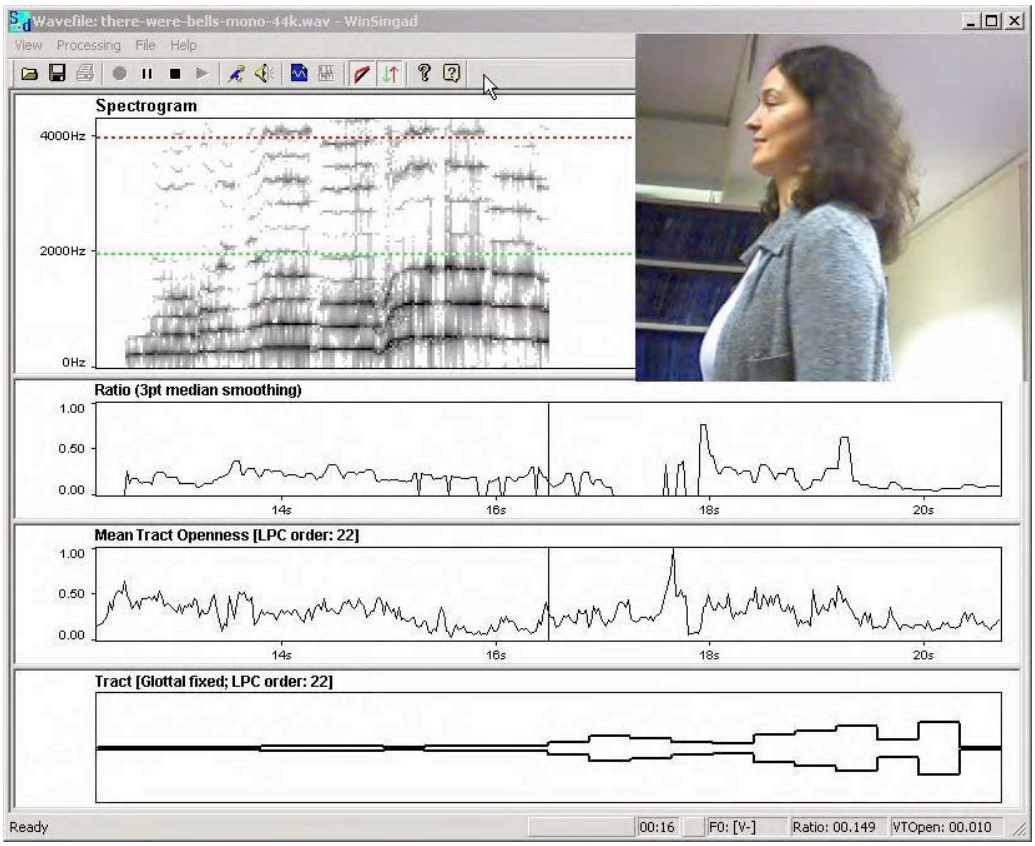


Figure 3
WinSingad *spectrogram, spectral ratio, mean oral tract openness and oral tract plots for a section of* There were bells *from 'The Music Man' (Willson, 1983) sung by a soprano*

## User-adjustable parameters

A key consideration when making use of any analysis system is whether the results obtained are both accurate from a technical point of view, and appropriate in terms of their usefulness for the task in hand. These considerations are generally fully explored for real-time visual displays during their development, since an analysis technique is selected for inclusion based on its demonstrated usefulness during experimental trials. There is, however, one unknown for the designer when a display is in use, namely if any controls for the analysis algorithm itself are provided to be adjusted by the user. Alteration of any such controls can have a potentially major effect on the nature of the display itself, and such changes do not relate to the input itself; rather they are an artefact of the operation of the algorithm.

In order to illustrate this in the most general way, examples are offered for $F_0$ and spectrographic displays commonly found in real-time visual feedback software for vocal development.

### *User-adjustable parameters and $F_0$ analysis algorithms*

All $F_0$ analysis algorithms are likely to have user-controllable parameters, particularly the level of the input signal. This might be user-controlled via the master volume sliders provided in Windows, an external control on a microphone amplifier or as a control provided within the analysis software itself. The $F_0$ algorithm used in

*WinSingad* is based on peak-picking (Howard, 1989) and it has some user controls whose effect is briefly described.

Figure 4 illustrates the effect on the $F_0$ display of adjusting the level, or 'gain', of the input signal for an arpeggio on the vowel in 'baa' sung by a tenor. The default value is 0.25, and the figure shows how variation of the input gain can have a substantial effect on the $F_0$ output, especially towards high and low extreme gain settings. Comparison with the $F_0$ output for an input gain setting of 0.25 suggests that too low a gain can result in part of the output being either missing altogether or broken up, whilst too high an input gain setting can result in $F_0$ values being indicated for background noise where there is no sung input. The effect for this example can be seen with respect to the changing levels of the sung input waveform.

In practice, a default setting will typically always be used. However, there might be situations where adjustment of the input gain is appropriate. For example, suppose a comparison was being made of the $F_0$ output during a phrase which included sections sung quietly (*pp*) and loudly (*ff*). An appropriate input gain setting for the *pp* section might be too high for the *ff* section, with an output perhaps similar to that illustrated for a gain of 0.7 in Figure 4, or vice versa where the result might be similar to that shown for a gain of 0.04 in the figure. The potential of this effect can be seen by comparing the presence or absence and accuracy of the $F_0$ output with the changing levels of the sung input waveform shown in the figure even for just this single arpeggio, where no perceived dynamic change was intended.

In order to improve the visual overall look of a measured parameter such as $F_0$ on a visual display in terms of irregular variation such as the additional spikes in the $F_0$ outputs for input gain settings of 0.04 and 0.1 in Figure 3, a technique known as 'output smoothing' can be employed. The purpose of output smoothing is to smooth out irregularities in a plot by calculating some form of average of the data themselves, and it is very common to make use of a median measurement. The median is the middle value of an odd number (3, 5, 7, etc.) of data values, and it is calculated as follows. Following the calculation of the output data samples, the median (middle) value is taken of each output data sample and either the next and the previous sample (3-point median), or the next two and the previous two (5-point median), or the next three and the previous three (7-point median) and so on. The result of median output smoothing is that extreme high and low values are lost since they will never be the median value, and large irregularities are removed. The key advantage of taking the median rather than for example, the mean value, is that all values in a median smoothed result are actual values that have been measured, whereas a mean value usually will not be equal to any measured value.

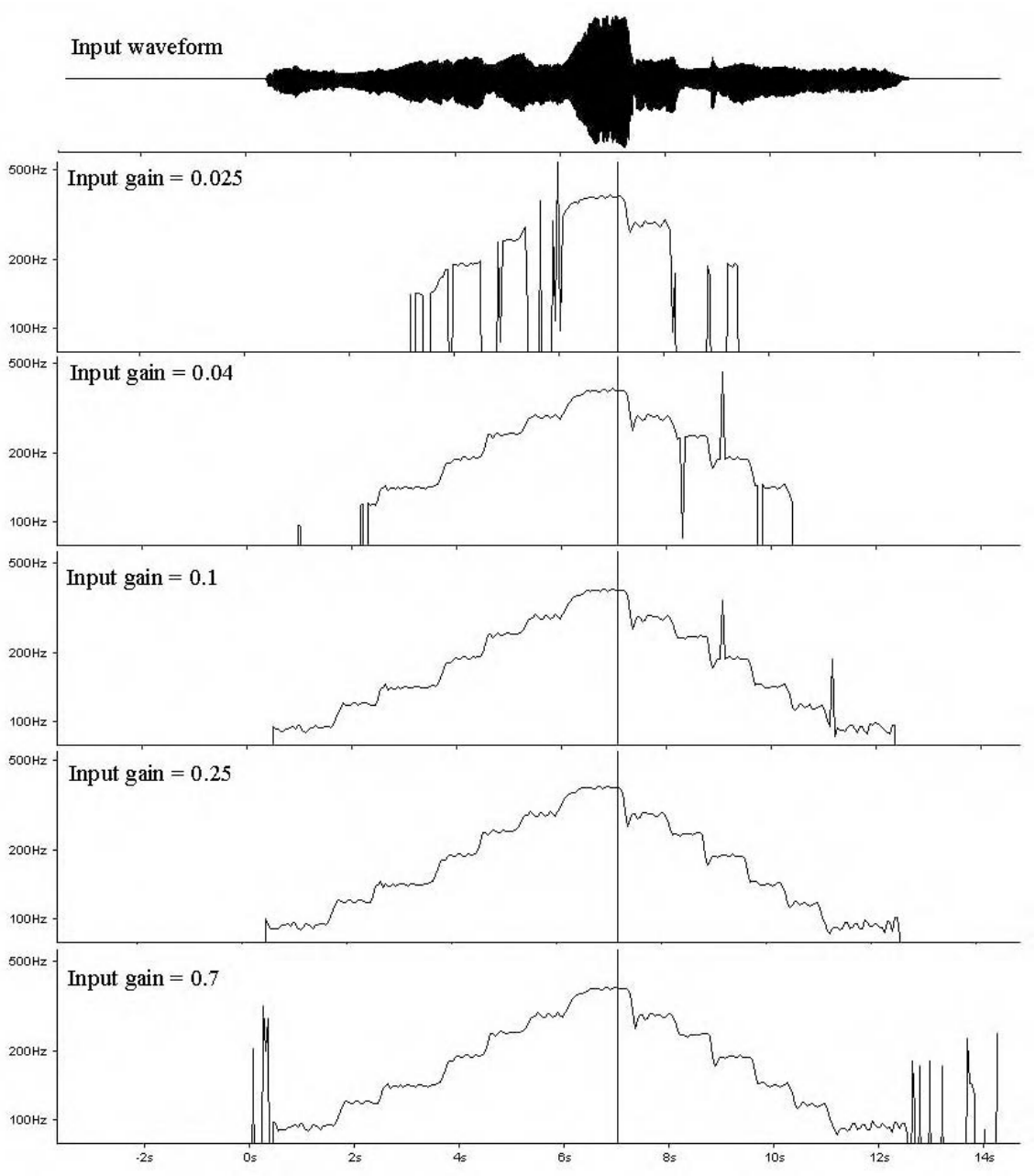


Figure 4
WinSingad *fundamental frequency plots with input gain settings of 0.025, 0.04, 0.1, 0.25 and 0.7 for a two-octave arpeggio on the vowel in 'baa' sung by a tenor. No output smoothing is employed.*

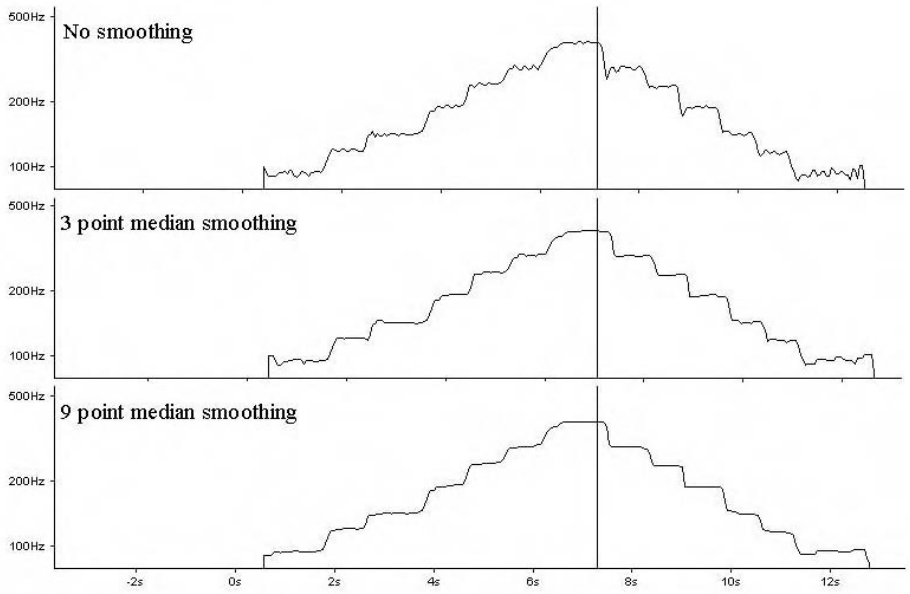


Figure 5

WinSingad *fundamental frequency plots with 0-, 3- and 9-point smoothing for a two-octave arpeggio on the vowel in 'baa' sung by a tenor. Input gain is 0.25 throughout.*

Figure 5 shows the effect of 3- and 9-point median smoothing on the $F_0$ contour for an arpeggio on the vowel in 'baa' sung by a tenor (the same signal analysed in Figure 4), where the input gain setting is 0.25. As the number of points used for the smoothing is raised, it can be seen that the output becomes smoother, but is it appropriate here? The main irregularity in the plot of the raw $F_0$ data (no smoothing) for this sung example is vibrato (visible note by note on the 'no smoothing' plot), and smoothing it out may not be what is desired.

The use of output smoothing can, however, be beneficial when there are clearly erroneous output values. By way of example, the effect of smoothing an output that is compromised by an inappropriate setting of a processing parameter will be demonstrated. The peak-picker used in *WinSingad* is sensitive to which way up the input waveform is (Howard, 1989), or its input 'polarity'. Figure 6 shows plots for the same data analysed in Figure 5, but here the input waveform has been inverted by means of a *WinSingad* toolbar button or dialog interaction. It is clear that the change in the $F_0$ output is enormous, and that the no smoothing result is highly misleading when compared to those in Figure 5. The effect of output smoothing is very clear here, and it demonstrates that the higher the number of points used, the greater the smoothing effect. The figure also demonstrates that smoothing *cannot* correct errors, something that is often commonly misunderstood. It is clear from the figure that there are still three notes even in the $F_0$ output with 9-point median smoothing that jump to values that are far too high.

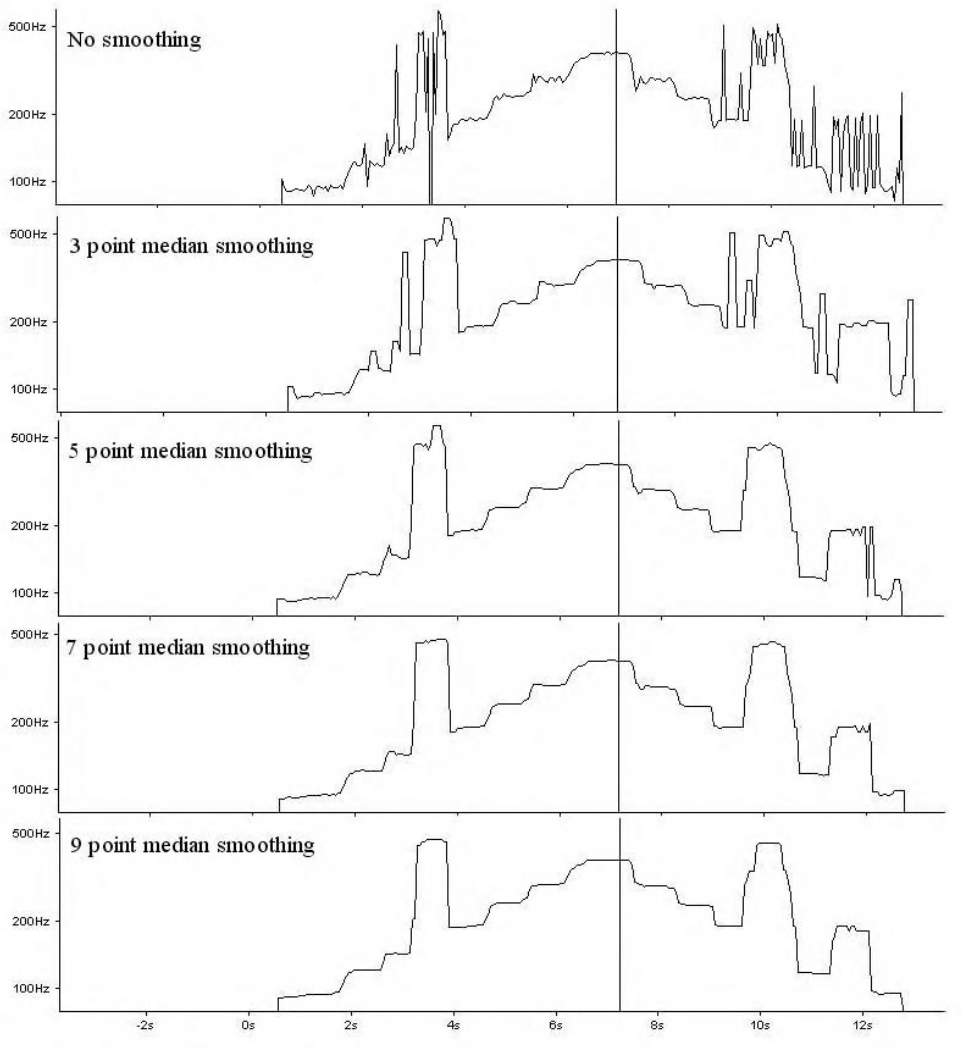


Figure 6
WinSingad *fundamental frequency plots for an inverted input waveform with 0-, 3-, 5-, 7- and 9-point smoothing for a two-octave arpeggio on the vowel in 'baa' sung by a tenor. Input gain is 0.25 throughout.*

***User-adjustable parameters and spectrograms***

A spectrogram provides a basic and essential analysis to investigate the acoustic nature of sound as a plot of the energy of the frequency components in the input signal against time. Traditionally, time and frequency are plotted on the horizontal and vertical axes respectively, with increasing energy being indicated as a greater darkness of marking on a grey scale. Modern computers have enabled the use of a colour rather than a grey scale, since they are well equipped for colour plotting, and many consider the colour scale to be more visually appealing. Using a colour scale for a spectrogram can, however, be misleading. The visually abrupt change between two adjacent colours, such as red to orange or blue to yellow, is very strongly suggestive of a distinct and discrete boundary between the energy levels that are represented by these colours. Any conclusion along these lines would be inappropriate since such a colour change represents simply a step in energy change, say from 65.0 dB to 65.1 dB.

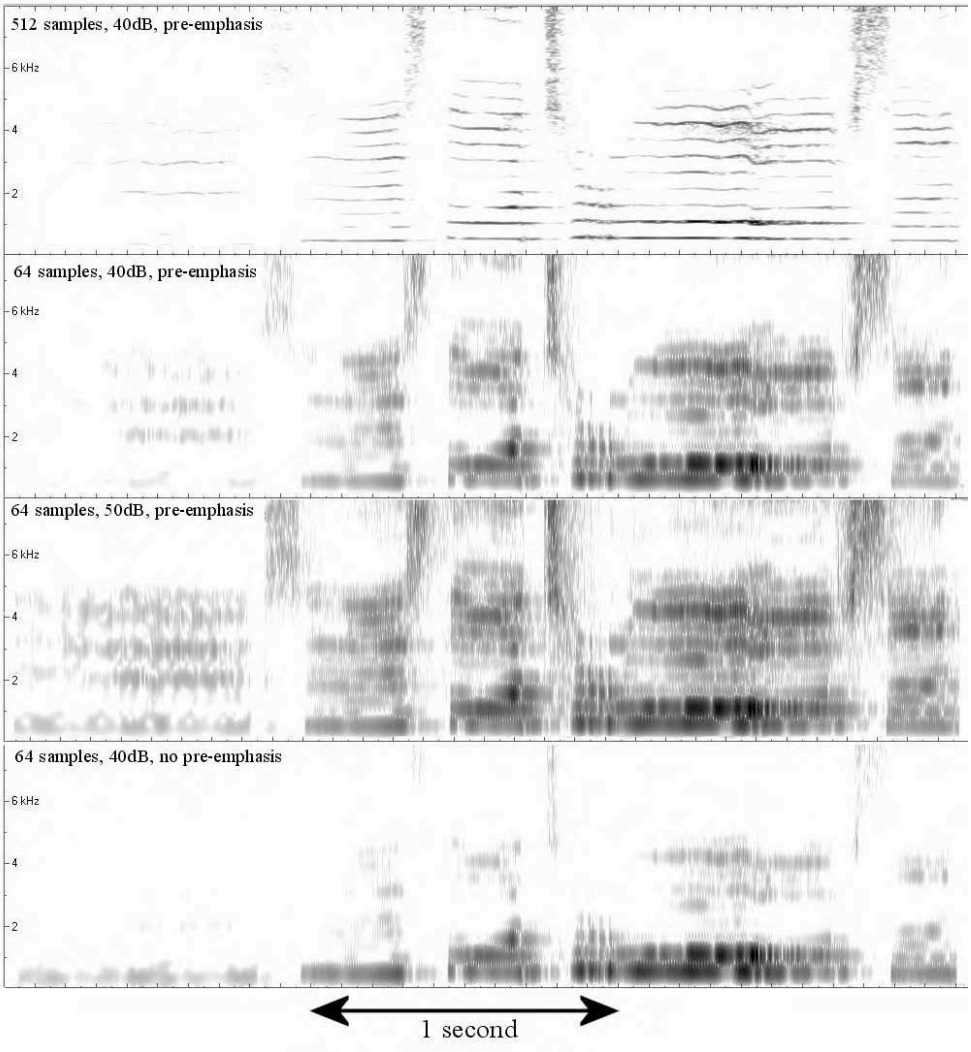


Figure 7
*Spectrograms to show the effect of the number of samples analysed, the use of pre-emphasis for 'this is the truth' from the Herefordshire carol* This is the truth from above *arranged by Vaughan Williams (Jacques, Willcocks & Rutter, 1972) sung by a soprano. (Key: 64 or 512 samples refer to the length of the analysis frame; 40 or 50 dB refers to the dynamic range of the spectrogram; pre-emphasis refers to whether or not a filter has been used to compensate for the average spectral slope of human speech or singing.)*

The essential technical consideration in spectrographic analysis is the accuracy available for frequency and time representation and measurement. Any increase in either the frequency or the time resolution can only be achieved at the expense of the other; increased accuracy in time resolution can only be achieved by reducing accuracy in frequency resolution and vice versa. In the context of singing analysis, this means that a spectrogram that offers a clear representation of the harmonics of the sung sound will be somewhat blurred in terms of the location of short-time events, such as note and consonant onsets. The frequency and the time resolution are controlled in spectrographic analysis by altering the number of samples being analysed in each frame.

Figure 7 shows spectrograms of 'this is the truth' from the Herefordshire carol *This is the truth from above* arranged by Vaughan Williams (Jacques, Willcocks &

Rutter, 1972) sung by a soprano. This figure enables the effect of the number of samples analysed, the dynamic range employed and the use of pre-emphasis to be illustrated. The greater the number of samples being analysed the better the frequency resolution but the poorer the time resolution and vice versa. The better frequency resolution for the spectrogram based on analysing 512 samples manifests itself in the thin essentially horizontal lines that make up its spectrogram in the figure. These horizontal lines are the individual harmonics of the input sound, and their thinness portrays the frequency accuracy of this particular spectrographic representation. As the number of samples is reduced, the nature of the spectrogram changes considerably. For 64 samples (40 dB and pre-emphasised), the underlying structure of the spectrogram is based more on vertical lines, or 'striations', each of which represents a glottal closure for sung notes. In addition, timings of events are better represented when a lower number of samples are employed for spectrographic analysis. This can be seen in this example particularly around the onsets and offsets of the consonants (the consonants can be identified in the spectrogram as having energy in the 4-8 kHz region).

Figure 7 also shows the effect of changing the dynamic range of the spectrogram itself. This refers to the energy range that is plotted between black and white. Usual practice (followed here) is to equate the highest energy in the signal to black. The effect of changing the dynamic range can be seen here for example at 40 dB and 50 dB (both for 64 points with pre-emphasis). The higher the dynamic range, the more low-level information appears on the spectrogram, but the general difference in levels in the various parts of the spectrogram becomes less obvious. An analysis of background noise will appear on the spectrogram if the dynamic range is set too high, and many important features can become lost on the spectrogram if the dynamic range is set too low. The value normally chosen is 40 dB as a suitable compromise.

The final aspect of spectrographic analysis that can change the nature of the output is whether or not pre-emphasis is used. Pre-emphasis is employed when analysing the human vocal output in order to even out the spectrum in terms of high and low frequencies so as to make best use of the dynamic range available on the spectrogram itself. The acoustic output during speech or singing tends, on average, to be lower at higher frequencies. This effect can be observed by comparing the 64 sample analyses in the figure with and without pre-emphasis, where the majority of the plot above 2 kHz is missing in the latter. Of particular note is the almost total absence of the energy during the consonants when analysed without pre-emphasis. The default setting in spectrography is to have pre-emphasis turned on, but if there is interest in knowing what the actual energy relationships are in the human vocal output, then it should be turned off (if a control is available).

## Conclusions

Today's computers that are commonly found in the home and the office are now more than capable of carrying out sophisticated acoustic analysis of the sort that was associated with advanced speech science laboratories just a decade or so ago. The interest in making use of such systems is growing as those involved in vocal pedagogy realise their potential for application in the professional voice studio. Such software is now available for the PC, both as freeware downloadable from the internet and as commercial products. It turns out that proper informed use of such software requires some knowledge of voice science and at least an understanding in overview terms of how such analyses function. This is particularly true when setting user parameters in order to obtain the best displays, since such adjustments can have

a direct impact on the features observed, sometimes through the introduction of artefacts.

The application of real-time visual feedback in singing lessons has been discussed in terms of how it can be used, what issues should be considered when setting it up, the kind of acoustic parameters that might be displayed and how adjustment of controls that are usually made available to the user have the potential to drastically change the form of the displayed output. It is suggested that the setting of such parameters to best effect requires a degree of knowledge of output expectation (KOE), which is often related directly to the nature of the algorithm being employed. Experience with such displays in the singing studio has suggested that they are found to be useful and helpful by both teachers and students alike. Understanding of basic principles, monitoring of progress and support in practice are all aspects that have been enhanced through employing real-time visual displays.

It is clear, though, that such technology will never completely replace the singing teachers, since there are many important aspects of singing where technology has relatively little to offer, including: building a repertoire, sight-singing, use of language, working with a conductor, stagecraft and performance practice. Such areas are likely to remain the sole preserve of the singing teacher, albeit with increased support from researchers.

## Acknowledgments

Part of this work was supported under an Innovation Award from the UK's Arts and Humanities Research Board (AHRB) numbered B/IA/AN8885/APN15651. The author is indebted to the teachers, students and co-researchers who were involved in this project, as well as to the reviewers for their very helpful comments.

## References

Baken, R. J. (1987). *Clinical measurement of speech and voice*. Boston: College-Hill Press.

Baken, R. J., & Daniloff, R. G. (1991). *Readings in clinical spectrography of speech*. San Diego: Singular Publishing Group.

Borden, G. J., & Harris, K. S. (1980). *Speech science primer*. Baltimore: Williams and Wilkins.

Fry, D. B. (1979). *The physics of speech*. Cambridge: Cambridge University Press.

Hess, W. (1983). *Pitch determination of speech signals: algorithms and devices*. Berlin: Springer.

Howard, D. M. (1989). Peak-picking fundamental period estimation for hearing prostheses. *Journal of the Acoustical Society of America*, *86*(3), 902-910.

Howard, D. M. (1995). Variation of electrolaryngographically derived closed quotient for trained and untrained adult female singers. *Journal of Voice*, *9*(2), 163-172.

Howard, D. M. (1999). The human singing voice. In P. Day (Ed.), *Killers in the brain* (pp. 113-134). Oxford: Oxford University Press.

Howard, D. M. (2002). The real and non-real in speech measurements. *Medical Engineering and Physics*, *24*, 493-500.

Howard, D. M., & Angus, J. A. S. (1998). Introduction to human speech production, human hearing and speech analysis. In F. A. Westall, R. D. Johnson & A. V. Lewis (Eds), *Speech technology for telecommunications* (pp. 30-72). London: Chapman and Hall.

Howard, D. M., & Welch, G. F. (1989). Microcomputer-based singing ability assessment and development. *Applied Acoustics*, *27*(2), 89-102.

Howard, D. M., & Welch, G. F. (1993). Visual displays for the assessment of vocal pitch matching development. *Applied Acoustics, 39*(3), 235-252.

Howard, D. M., & Welch, G. F. (2003). Real-time visual displays for singing development. *Journal of the Indian Musicological Society*, *34,* 7-23.

Howard, D. M., Welch, G. F., Brereton, J., Himonides, E., De Costa, M., Williams, J., & Howard, A. W. (2004). WinSingad: a real-time display for the singing studio. *Logopedics Phoniatrics Vocology*, *29*(3), 135-144.

Jacques, R., Willcocks, D., & Rutter, J. (1972). *Carols for choirs 2*. Oxford: Oxford University Press.

Kent, R. D., & Read, C. (1992). *The acoustic analysis of speech*. San Diego: Singular Publishing Group.

Koenig, W., Dunn, H. K., & Lacy, L. Y. (1946). The sound spectrograph. *Journal of the Acoustical Society of America*, *18,* 19-49.

Markel, J. D., & Gray, A. H. (1976). *Linear prediction of speech*. Berlin: Springer-Verlag.

Moorcroft, L. (2002). Embracing alternative methodologies: science and imagery in the teaching and performance of singing. In C. Stevens, D. Burnham, G. McPherson, E. Schubert & J. Renwick (Eds), *Proceedings of the 7th international conference on music perception and cognition* (pp. 561-564). Adelaide: Casual Publications.

Nair, G. (1999). *Voice tradition and technology: a state-of-the-art studio*. San Diego: Singular Publishing Group.

Potter, R. K., Kopp, G. A., & Green, H. (1947). *Visible speech*. New York: Van Nostrand.

Rosen, S., & Howell, P. (1991). *Signals and systems for speech and hearing*. London: Academic Press.

Rossiter, D., & Howard, D. M. (1994). ALBERT: a system for interactive analysis and display of voice source and acoustic parameters. *Proceedings of the Institute of Acoustics, 16*(5), 301-308.

Rossiter, D. P., Howard, D. M., & Downes, M. (1995). A real-time LPC-based vocal tract area display for voice development. *Journal of Voice*, *8*(4), 314-319.

Rossiter, D. P., Howard, D. M., & De Costa, M. (1996). Voice development under training with and without the influence of real-time visually presented biofeedback. *Journal of the Acoustical Society of America, 99*(5), 3253-3256.

Sundberg, J. (1974). Articulatory interpretation of the 'singing formant'. *Journal of the Acoustical Society of America, 55,* 838-844.

Sundberg, J. (1987). *The science of the singing voice*. Dekalb, Illinois: Northern Illinois University Press.

Thorpe, C. W., Callaghan, J., & van Doorn, J. (1999). Visual feedback of acoustic voice features for the teaching of singing. *Australian Voice*, *5*, 32-39.

Titze, I. (1994). *Principles of voice production*. Englewood, NJ: Prentice Hall.

Welch, G. F. (1985). A schema theory of how children learn to sing in tune. *Pyschology of Music*, *13*(1), 3-18.

Welch, G. F., Howard, D. M., & Rush, C. (1989). Real-time visual feedback in the development of vocal pitch accuracy in singing. *Psychology of Music*, *17,* 146-157.

Welch, G. F., Howard, D. M., Himonides, E., & Brereton, J. (accepted for publication). Real-time feedback in the singing studio: an innovatory action-research project using new voice technology. *Music Education Research.*

Willson, M. (1983). *The music man.* Wisconsin: Hal Leonard Corporation.

WWW-1: <http://www-users.york.ac.uk/~dmh8/freeware-web.htm>.

WWW-2: <http://www.singandsee.com/>.

WWW-3: <http://www.vocevista.com/>.

WWW-4: <http://www-users.york.ac.uk/~dmh8/winsingad.htm>.

[i] 'Real-time' means fast enough so that the user perceives no delay between the production of a sound and the associated plot appearing on the display.

## About the Author

*David Howard* studied at University College, London, from where he graduated with an accredited First Class Honours BSc (Eng) degree in Electrical and Electronic Engineering in 1978 and a PhD in Human Communication in 1985. His PhD was concerned with the presentation of pitch information via a cochlear implant hearing aid. He became a lecturer at University College in Experimental Phonetics from 1979. His external interests in music (singing, conducting and keyboard playing) led him to move in 1990 to the Department of Electronics at the University of York to lecture in Music Technology. There he was promoted to a senior lecturer in 1993 and to a personal chair in 1996.

David's research interests are focused around the analysis and natural synthesis of singing, speech and music and the application of real-time displays for professional voice users. When not at work, he might be found conducting his small, mainly a cappella choir from the tenor line, playing the organ, sailing or singing as a Deputy Songman in York Minster.