VISUAL FEEDBACK OF ACOUSTIC VOICE FEATURES IN VOICE TRAINING

C. William Thorpe

School of Communication Sciences & Disorders,
University of Sydney

ABSTRACT: Speech technology has long been utilised to within computer-assisted tools for voice training. Such tools, providing real-time visual feedback of specific characteristics of the voice, complement traditional training methods. However, in order for such technology to be successful, specific aspects of voice training must be considered including the nature of voice perception, motor learning processes, and what information can be reliably extracted from the voice signal. This paper examines the factors that must be considered in developing voice training technology, and provides an overview of the techniques that have been applied to extract information from the voice signal and present it visually for real-time training purposes.

INTRODUCTION

A potentially exciting application of speech technology is the provision of visual feedback in voice training. Traditionally, voice training in areas such as singing, acting, and speech therapy, is based on a model of practice coupled with teacher-provided feedback. The teacher instructs the student as to what kind of performance to practice, and provides feedback about how well his or her performance matches the desired target ideal. Typically, the feedback concerns aspects of either the acoustic quality of the vocalisation, or physiological aspects such as breathing. However, even with highly competent teachers, any feedback given by the teacher has limitations, which may reduce the learning efficiency. Firstly, the teacher must be able to perceive a particular quality of performance in order to provide comment. Some aspects of physiology are of course invisible to the observer, even by manual sensation, and even some acoustic aspects may be difficult to perceive precisely except in highly trained individuals (eg subtle differences between vowels within categorical spaces). Secondly, the teacher must communicate the feedback to the student, either in words, by touch, modelling a “correct” performance, or perhaps by means of metaphor. Thirdly, the student must be able to understand the feedback and use it to alter the subsequent performance in order to effect the desired improvement. Finally, the teacher may be in a classroom situation where attention to individual students is by necessity limited.

Providing ancillary feedback of particular aspects of a student’s performance offers the possibility to overcome some of the limitations of traditional teacher-provided feedback. Firstly, such feedback is not limited by (auditory) perceptual processes, but may highlight aspects of the performance that are not easily perceived. Secondly, the feedback can be communicated directly in a visual format that directly mirrors the performance, and furthermore, this feedback can be contemporaneous with the student’s vocalisation. This implies that it may be easier to integrate the knowledge gained into improvements in performance, compared to feedback that is delayed.

The purpose of this paper is to provide an overview of the various issues involved in providing visual feedback for the purposes of voice training. In Section 1, issues relating to learning processes themselves are discussed, followed by an overview of how the characteristics of voice perception need to be taken into account during feedback. Section 3 describes the types of acoustic analyses that are available to extract relevant information from the voice, while Section 4 describes ways in which this information can be visually presented to the student. Finally, some of the pedagogical considerations of incorporating computer technology in voice training are discussed in Section 5. The focus here (and the examples included) is on the application to singing, but the factors considered are also broadly relevant to other vocal training tasks including accent training, speech therapy, etc.

1. LEARNING AND FEEDBACK

Vocal training can be considered as a form of motor training in which it is desired to obtain extremely accurate control of all the muscle groups involved in vocalisation. The training process involves a series of trial productions, with feedback about each trial being used to improve subsequent productions. This model of learning involves both internal and external feedback mechanisms. The internal loops include direct proprioception of muscle activity, and indirect auditory perception of the sound output and kinaesthetic perception of the body's responses to vocalisation (eg respiratory pressures, vibrations, posture). The external loops generally consist of feedback by a teacher regarding the quality of the production or some aspect of how it was made (Welch, 1985). This feedback provides the student with some meaningful knowledge about how well their production satisfied some ideal goal. This "knowledge of results" (KR) must then go through a process of interpretation, so that the student's subsequent production can be modified appropriately to improve performance (Schmidt, 1975).

Considering the motor learning aspect of voice training, it is instructive to consider some of the recent research in modelling the learning of motor control tasks (e.g. . Brashers-Krug et al., 1996; Wolpert & Kawato, 1998). These studies suggest that the process by which knowledge of results (obtained by feedback from the student's performance) is transformed into improved performance proceeds by two distinct steps. In the first stage, which occurs during practice itself, the production errors highlighted by feedback are processed through an inverse model of the production mechanisms, so that appropriate corrections can be made to the motor control programs for subsequent trials. A second stage of learning occurs later (even after the period of feedback practice has ended), in which the knowledge gained during the practice is used to update a forward mode, that is able to actually plan and implement the muscle actions required for a desired performance output. This second phase is of course most important because in order to perform to the highest standard (or to produce speech rapidly enough for running communication), the student must be able to generate neuro-muscular output directly from an internal representation of the desired output, without the necessity to make minor adjustments in response to feedback as the performance proceeds – ie the motor control must be predictive, via the forward model, rather than reactive, via the inverse model (Perkell et al, 2000).

The feedback model of motor control learning has important implications for voice training, because if students are not able to easily make the link between the feedback provided and the motor control adjustments that are required to improve their performance, it may be difficult to integrate the knowledge of results into their forward model of the motor control mechanisms. In particular, feedback is often delayed, and so the student must interpret what the teacher says in light of the memory of how the performance was produced (Welch, 1985). Also, there may be problems of comprehension between the student and teacher, either due to differences in the meaning attached to terminology, or because of difficulties in verbally explaining aspects of voice production and perceptual qualities. Provision of real-time feedback of appropriate characteristics of the voice, in a visual format, sidesteps many of these difficulties by providing knowledge of results that is concurrent with the production, allowing immediate correction of any errors as the student continues to phonate, and a direct association between the student's internal motor control programs and the acoustic output.

Naturally, for feedback to be useful and contribute to learning, it must be appropriate for the learning task, and provide reliable knowledge of results. Reliability implies both consistency, so that similar productions result in similar feedback outcomes, and also accuracy, so that when the student approaches the desired production, the feedback converges towards the target. If inaccurate feedback is provided, it may worsen learning performance, although if the student's own perceptual feedback is strong enough, that can override any conflicting external feedback (Buekers et al., 1994). It is also useful if the feedback provides an indication of both the magnitude of any error, and the direction in which the student must change in order to reduce the error (eg "much too high", rather than "wrong").

2. PERCEPTION AND CONFUSION

A key issue in learning to sing or speak in a new manner is that the perception of auditory and vocal qualities is extremely non-linear, and highly variable between individuals. In particular, many vocal qualities (e.g. vowel identity and musical intervals) are perceived categorically, with boundaries determined by prior exposure and training. This means that small acoustic changes can result in either negligible or large perceptual changes depending on whether a category boundary is crossed. Indeed, if a student's perceptual boundaries are different from what is required by the training context, then it may be very difficult for the student to perceive a difference that the teacher requires. Because of the link between voice production and voice perception, one can consider that these categories exist in the student's inverse model for transforming feedback into motor control corrections.

Therefore, it is perhaps necessary to learn the appropriate perceptual boundaries before students can make use of auditory feedback in modifying their voice production.

Because visual feedback is not subject to the same auditory perception processes, it is able to bypass any perceptual difficulties the student may have, and provide a direct representation of the voice acoustic signal. Of course, it is desirable that the representation provided is relevant with respect to perceptual qualities of the voice. For instance, pitch is a perceptual quality that is closely (but not linearly) related to the fundamental frequency of the voice signal. Also, music generally conforms to a culturally-determined scale in which only particular values of pitch, and intervals between pitch values, are allowed. Beginning students must learn to recognise and produce these intervals and pitch values. Provision of visual feedback to aid in pitch training should therefore provide feedback in terms of musical intervals and pitch accuracy, according to the particular scale desired.

3. ACOUSTIC ANALYSIS

The medium of vocal production (whether singing, acting performance, or speech communication) is primarily the acoustic signal by which the voice is conveyed to listeners. Therefore, one should be able to extract any information that is relevant to voice performance from that signal. However, as implied by the complexities of human voice perception, identifying and extracting the required information is not always a straightforward task. Indeed, even human listeners can have difficulty in perceiving particular qualities of the singing voice, for instance the identification of vowels at high pitch (Scotto Di Carlo & Germain, 1985).

Information conveyed by the voice is diverse and complex, including characteristics such as pitch, vowel identity, intonation, prosody, and timbre, each of which involve a rich set of variations. Acoustically, each of these characteristics can be broadly related to particular attributes of the acoustic waveform, although there are significant interactions between them. For instance, pitch corresponds largely to the fundamental frequency of the signal, although there is some perceptual interaction with the spectral content of the sound (eg vowels with higher formants tend to sound sharper). Dynamic loudness corresponds to the acoustic power in the signal, also depending on the spectral distribution of the power. Vowel identity largely relates to the frequencies of the first two or three resonances (formants) in the voice signal, although this relationship becomes more difficult to observe (both acoustically and perceptually) as fundamental frequency increases. Consonant identity relates to both spectral and temporal patterns (eg distribution of broad-band noise for fricatives; timing of energy bursts for stops). Finally, qualities of vocal timbre correspond both to spectral distribution of energy (eg magnitude of the “singer’s formant”) and to temporal patterns in the sound – such as vibrato rate, extent, and establishment.

Traditionally, display of voice acoustics has been by means of the spectrogram, which shows both spectral and temporal patterns by means of a pseudo-three dimensional image of the acoustic energy at each point in time and frequency. There is a fundamental trade-off between time and frequency resolution, but essentially all the information in the acoustic signal is represented in the patterns displayed on the spectrogram. Several authors have attempted to catalogue the details of spectrographic displays resulting from various speech (Potter et al, 1967) and singing (Nair, 1999) vocalisations, but the surfeit of information contained in the spectrographic image can make interpretation difficult, particularly given the natural variability in vocal output. This “information overload” is particularly pertinent in the context of real-time feedback for training purposes. If interpretation of the visual information requires a high cognitive load, then the student may be distracted from the task of actually integrating the feedback provided. Also, because the visual representation is generally by means of a colour or grey-scale mapping of the sound energy, the spectrographic image can vary markedly with relatively minor changes in acoustic signal strength (for instance if the student moves relative to the microphone). Evidently, both the teacher and the student require some training in order to interpret the patterns they see on the spectrogram and so make the best use of the information available.

In order to abstract facets of the information contained in the spectral-temporal patterns comprising the voice sound, it is necessary to invoke a model of the desired characteristic. For instance, vowel identity can be modelled with respect to the resonances within the vocal tract that produce the sound. Therefore, by identifying the frequencies and bandwidths of resonances in the voice spectrum, one can extract information that relates to the vowel identity. As another example, vibrato is a sinusoidal variation (of around 5-8Hz) in the fundamental frequency of vocal fold vibration. One can therefore extract the fundamental frequency contour from a vocal recording, and fit a sinusoidal function with free parameters being frequency, extent, and onset phase.

Obviously, any model must reflect relevant aspects of either how the voice is produced (eg source-filter model) or how it is perceived (eg frequency and amplitude scaling; critical bands; etc). In particular interest to singing, models of some perceptual qualities (eg “colour”, “warmth”, etc) are still being developed, although there are some perceptual studies available that have been able to correlate various perceptual attributes with acoustic features (Ekholm et al., 1998). Their results suggest that the perceptual attribute of “resonance” is well-correlated to the acoustic power in and around the “singer’s formant” (a strong resonance between 2.5 and 3.5kHz that appears to be caused by a clustering of the third, fourth, and fifth formants (Sundberg, 1994)). The acoustic power in the frequency band between 2-4kHz, relative to the acoustic power in the band 0-2kHz also correlates well with the use of “projection” in the singing voice (Thorpe et al, 2001), and has been used to provide an indication of breaks in the voice over register changes (Bogg & Thorpe, 2000).

Other acoustic measures that appear to represent perceptual characteristics include the overall spectral slope (Bloothooft & Plomp, 1988) and the levels of individual harmonics within single critical bands (Sundberg, 1994). Ekholm et al showed that there was some relationship between the perceptual attribute of “colour” and the measurements of vibrato rate, extent, and onset delay at the start of a note.

4. VISUAL REPRESENTATION

If we take the view that spectrographic representation of the voice is overly complicated for real-time use in voice training (although it may have an important role in examining features of the voice at a more leisurely pace, particularly since it contains such a wealth of information) then we must consider how best to represent information for the student. There are several requirements for a good visual feedback display. Firstly, it must be simple to interpret, so that the student can absorb the information conveyed with little cognitive load. Secondly, it must convey the information in a manner that is relevant to the required learning task. For instance, a display that shows pitch error in terms of a meter display is good if one is interested in refining pitch accuracy on a single note (indeed, such displays are often used for electronic pitch tuners). However, it does not convey information about the magnitude of pitch intervals, which may be better provided by a moving line kind of display. Thirdly, it is helpful if the display relates to some physical aspect of how the associated voice characteristic is produced. For instance, the typical acoustic vowel chart is displayed with the directions of the formant frequency axes reversed, so that the up and down direction corresponds (approximately) to movements of the jaw, and the left to right axis corresponds to front to back movements of the tongue. This allows the student to directly transfer the feedback of vowel positions displayed on the chart into corresponding articulatory manoeuvres, without a great amount of cognitive processing.

An alternative approach to displaying the formant frequencies of vowels directly is to estimate the vocal tract shape that could have produced the observed resonances, and to display a representation of that shape (Rossiter et al., 1994). However, the usefulness of vocal tract shape as a feedback display is affected by limited proprioception of ones own vocal tract configuration. Also, the solution for vocal tract shape can suffer from non-uniqueness, making this type of display prone to reliability problems.

Other types of visual representations have included pictorial representations whose appearance is controlled by some specific acoustic feature. For instance, the IBM Speech Viewer™ has displays such as a balloon that enlarges as acoustic power increases, pictures that change colour depending on whether voiced or unvoiced sounds are produced, and action pictures where the movement is controlled by voice features such as a specific vowel quality being produced. These types of displays can be appealing to younger learners in particular (this software is designed for speech therapy training in children) where it is required to hold the interest of the student over the course of training. However, the disassociation between visual display and acoustic quality means that it is necessary to first explain what the various changes in pictorial output imply with respect to changes in voice production.

Rossiter & Howard (1992) described an interesting form of visual display in which multiple acoustic attributes (eg fundamental frequency, closed quotient, etc) are mapped into a virtual space comprising a 3-dimensional spatial field together with attributes of colour and object shape to convey the vocal characteristics as they are produced. There is however, a disassociation between the display and any “real-world” attribute of voice production or perception, implying that applying this type of display to a learning process would require some additional cognitive load.

5. COMPUTERS VOICE TRAINING – PEDAGOGICAL IMPLICATIONS

Although there seem to be good reasons for employing visual feedback as a training tool in voice training, and emerging evidence for its usefulness, actually utilising the technology in teaching practice is of course fraught with difficulty. Most teachers operate under very tight constraints of both time and content, with specified material required to be taught for courses or exams. Without clear pedagogical guidelines for how visual feedback technology should be utilised within a lesson, its introduction may simply lead to bad experiences for both student and the teacher, with valuable classroom time being wasted.

Visual feedback technology has been applied to a variety of pedagogical situations. Welch et al. (1989) investigated its use in pitch training in a classroom situation with school children. In their study, two groups of children received visual feedback, one group also being guided by reinforcement by the teacher. They found that the combination of (real-time) visual feedback and verbal feedback from the teacher was more effective than visual feedback alone, but both provided better learning outcomes than relying on a traditional teacher-only method (although in a classroom situation the amount of teacher feedback provided to each student is of course limited).

Several authors have reported on the use of spectrographic displays of the singing voice within traditional studio teaching (Nisbet, 1995; Miller & Doing, 1998; Nair, 1999). All reported some degree of success in this application, using the spectrographic analysis as a feedback tool. Nair’s book explores some of the complexity of spectrographic displays, with in-depth explanations of the typical features that occur for various vocal exercises and voice qualities. In our experience with real-time visual feedback, the successful use of spectrographic displays has depended on the experience and confidence of the teacher in interpreting the patterns that appear on the screen.

In our work, we have experimented with using real-time visual feedback in several pedagogical settings, including single-lesson “enhanced” classes (Callaghan et al, 1999) and in ongoing regular lessons. In both types of setting, teachers and students reported that the feedback had been beneficial, both to the learning experience, and to their understanding of their voices. Analyses of acoustic changes in the students’ voices however showed a wide spread of responses to the training, with only some students having measureable improvements in abilities such as pitching accuracy.

CONCLUSIONS

There appears to be an emerging convergence between the availability of computing power required to provide real-time analysis and display of voice characteristics, and the desire of voice teachers in many areas of this field to make use of more sophisticated forms of feedback. However, the development of appropriate analysis algorithms that can provide the relevant information to students and teachers still requires more work. For such technology to be successful in assisting with voice training, it must cater to the motor-learning requirements of providing real-time and reliable feedback of how well the student is approaching a performance target. The reliability of the feedback relies on the development of algorithms that take into account the characteristics of voice perception, and that provide information that is relevant in the context of the particular voice application. Quantitative assessment of feedback technology in learning environments however requires further development of methods for assessing progress in singing ability.

ACKNOWLEDGEMENTS

This paper reports aspects of an ongoing research effort in collaboration with Jan van Doorn and Jean Callaghan. Financial support was provided by the ARC (# A00104496).

REFERENCES

Bogg L & Thorpe W (2000), 'Register Change in the Countertenor Voice', *Australian Voice*;6:23-29

Brashers-Krug T, Shadmehr R, Bizzi E. (1996). Consolidation in human motor memory. *Nature*, 382 (6588), 18 July 1996, 252-255.

Buekers, M.J., Magill, R.A., & Sneyers, K.M. (1994). Resolving a conflict between sensory feedback and knowledge of results, while learning a motor skill, *Journal of Motor Behavior*, 26(1): 27-35.

Callaghan, J., Thorpe, W. & van Doorn, J. (1999). Computer-assisted visual feedback in the teaching of singing. In M. Barrett, G. McPherson & R. Smith (Eds), *Children and music: Developmental perspectives*, 105-111. Launceston: University of Tasmania.

Ekholm, E, Papagiannis, GC, & Cagnon FP. (1998). Relating objective measurement to expert evaluation of voice quality in Western Classical Singing: Critical perceptual parameters. *Journal of Voice*, 12(2), 182-96.

Miller, R. & Franco, J.C. (1991). Spectrographic analysis of the singing voice. *The NATS Journal*, 48(1): 4-5, 36.

Miller, D. & Doing, J. (1998). Male passaggio and the upper extension in the light of visual feedback. *Journal of Singing*, 54(1): 3-14.

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

Nisbet, A. (1995). Spectrographic analysis of the singing voice applied to the teaching of singing. *Australian Voice*, 1: 65-68.

Perkell JS, Guenther FH, Lane, H, Matthies ML, Perrier P, Vick J, Wilhelms-Tricarico R & Zandipour M (2000) 'A theory of speech motor control and supporting data from speakers with normal hearing and with profound hearing loss', *Journal of Phonetics* 28 :233-272.

Potter RK, Kopp GA, & Kopp HG (1966) *Visible Speech*, 2nd Ed. New York : Dover.

Rossiter, D. & Howard, D.M. (1992) 'Cyberspace visualisation of vocal development data', *Proceedings of the Institute of Acoustics*, 14(Pt 6):241-248.

Rossiter, D., Howard, D.M., & Downes, M. (1994) 'A real-time LPC-based vocal tract area display for voice development', *Journal of Voice*, 8(4):314-319.

Rossiter, D. & Howard, D.M. (1996) 'ALBERT: A real-time visual feedback computer tool for professional vocal development', *Journal of Voice*, 10(4):321-336.

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

Schmidt, R.A. (1975). A schema theory of discrete motor skill learning. *Psychological Review*, 82 (4): 225-260.

Scotto Di Carlo, N. & Germain, A. (1985). "A perceptual study of the influence of pitch on the intelligibility of sung vowels", *Phonetica* 42:188-197.

Sundberg, J. (1994). 'Perceptual aspects of singing', *Journal of Voice*, 8:106-122.

Thorpe, CW, Callaghan J, van Doorn J (1999), 'Visual feedback of acoustic voice features: New tools for the teaching of singing', *Australian Voice*;5:32-39

Thorpe CW, Cala, SJ, Chapman J, and Davis PJ, 'Breathing patterns with projection of the singing voice', *Journal of Voice* 2001;15(1):86-104

Wapnick, J. & Ekholm, E. (1997) Expert consensus in solo voice performance evaluation. *Journal of Voice*, 11(4):429-36.

Welch, G.F. (1985). A schema theory of how children learn to sing in tune. Psychology 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.

Wolpert DM, Kawato M. (1998). Multiple paired forward and inverse models for motor control. *Neural Networks*, 11 (7-8):1317-1329.