SPEECH TECHNOLOGY IN CLINICAL ENVIRONMENTS:

J. van Doorn[†], *S. McLeod*[†], *E. Baker*[†], *A. Purcell*[†], *W. Thorpe*[‡]

[†] School of Communication Sciences and Disorders, The University of Sydney
[‡] National Voice Centre, The University of Sydney

ABSTRACT

Traditionally, perceptual judgement of speech disorders by clinicians has been a cornerstone of speech language pathology. Increasingly, it is being argued that acoustic speech analysis should supplement aural perception in the clinic. For successful clinical application of speech technology, experts in acoustic analysis generally agree that a working knowledge of acoustic phonetics, digital signal processing and the literature on the acoustic characteristics of speech disorders are required. However, it is not necessarily compellingly obvious to clinicians. This paper examines the issues by examining how the various components fit into the clinical picture. It examines when and how speech technology can be used by clinicians, arguing that clinicians need to be able to do much more than just operate the system. The paper concludes that successful integration of speech technology into clinical environments provides an opportunity to for technologists and clinicians to work together to produce effective speech technology for clinical applications.

1. INTRODUCTION

Acoustic speech analysis is now an easily accessible tool that opens up exciting possibilities in clinical speech measurement. Affordable, commercially available speech analysis systems that run on personal computers are now available for applications in speech clinics. (eg *Kay CSL* and *IBM SpeechViewer* systems).

Speech technology is already a well-established research tool for investigating speech disorders. In reviewing the field Farmer (1) refers to "an exhausting but not exhaustive" list of reported findings in the literature. This research is providing an essential component for successful clinical application of speech technology ie an expanding body of knowledge about the acoustic characteristics of many speech disorders. It has also generated a group of people whose expertise in analysing speech has given them a full appreciation of the requirements for accurate, useful information to be derived from a speech signal. Among these people there is general consensus about the information that underpins successful clinical application of acoustic analysis of speech disorders (1), (2), (3):

- A detailed understanding of the acoustic speech signal, and its relationship to speech production and perception.
- An understanding of requirements for recording and storing the speech signal in a form suitable for analysis using digital techniques.
- An appreciation of the process of digital signal processing required to generate acoustic speech features
- Up to date knowledge of the body of literature relevant to specific speech disorders

Although there is little argument among experts in acoustic analysis that a working knowledge of the above areas is required for successful use of acoustic analysis, it is not necessarily compellingly obvious to clinicians or students of speech language pathology. This paper examines the issues "in context" for clinical applications, looking at where the various components fit into the clinical picture.

2. WHEN TO USE SPEECH TECHNOLOGY

2.2 Supplement to Clinician's Aural Perception

Traditionally, perceptual judgement of speech disorders by clinicians is a cornerstone of speech language pathology. Perceptual judgments still provide a "golden standard" for professional judgments about speech disorders. Clinicians are trained to become expert listeners in phonetic transcription of speech disorders, and in judgements of speech qualities such as nasality and voice quality. Clinicians need to decide whether there is any advantage in using acoustic measures in place of or in conjunction with traditional perceptual clinical judgments for a particular client.

This decision is made easier if clinicians understand the capabilities and limitations of human speech perception, in particular that the perceptual process is subject to effects such as categorical perception and contextual effects. This can result in disordered speech being incorrectly and unreliably transcribed (4), (5) or to subtle acoustic differences going undetected by listening (6), (7), (8), (9). Similarly, acoustic analysis may be useful where first language experience influences our ability to perceive and produce speech sounds in another language (10). Being aware of situations where the perceptual process, even of experienced listeners, is subject to the influence of higher processing effects, will alert clinicians to the potential situations where acoustic analysis may provide valuable insight that could not be achieved by listening alone.

In other cases perceptual judgments of speech qualities such as voice quality and nasality are susceptible to problems that influence their reliability, even with expert listeners (11), (12). Objective measures have the potential to enhance the reliability

of clinical judgements. However, our current knowledge of the relationship between acoustic measures and perceptual judgments for nasality and voice quality is far from complete. It is important that the clinician is aware of the current state of knowledge for acoustic features of particular disorders, and hence whether there is potentially a useful clinical outcome from using speech technology.

2.2 Visual Feedback

There is a major application of speech technology in providing visual feedback to clients. Clinicians can make use of potential motivational and instructional aspects of visual displays of acoustic features (either as spectrographic information or as a graphic display eg in a game) for particular clients. Programs such as the *IBM SpeechViewer* and the *CSL-Pitch* module of the *Kay CSL* have well-developed sets of games with graphical interfaces that are specifically designed for use in speech pathology clinics.

3. HOW TO USE SPEECH TECHNOLOGY

3.2 Choice of Speech Features

Deciding on the appropriate speech analysis requires an appreciation of the range of available acoustic measures. This in turn entails understanding the nature of the acoustic speech signal, and knowing established measures such as fundamental frequency and formant frequency that are used to describe it. Then it is important to discriminate those acoustic features that are known to be cues for perceptual judgments, from those that are used to quantify certain characteristics of speech, even when their relationship to speech perception is uncertain. When using speech technology to investigate speech disorders, selection of speech features depends also on the available knowledge of the acoustic characteristics of the disorder in question.

Where speech technology is used to generate graphic displays to provide visual feedback for the client, the clinician should understand what acoustic parameters are being extracted automatically from the speech signal, and how they relate to the graphic display.

3.2 Successful Extraction of Acoustic Features

Technical aspects of speech analysis are fundamentally of little interest to most clinicians. To the clinician, using speech technology often means little more than "knowing what buttons to press". Familiarity with the actual operation of the equipment is in fact an essential clinical skill. Ironically, it is probably conceptually the least important, and yet it is the most visible for the client, so that the credibility of the clinician and the technology is under threat while the clinician is operating the equipment. It must be remembered that mastering the buttons is only a part of mastering the use of technology in clinical settings.

The first challenge is to convince speech language pathologists that they require sufficient knowledge to understand the capabilities and limitations of digital signal processing for their professional credibility when they are using speech technology (13). It is probably best demonstrated by illustrating the consequences of inappropriate analysis on the clinician's ability to interpret the information. The next challenge is to present concepts such as the rudiments of spectral analysis of speech and the associated digital signal processing techniques (eg Fast Fourier Transforms, Linear Predictive Coding), and the fundamentals of algorithms used in automatic feature extraction (eg voicing parameters), with minimal reliance on maths, physics and engineering (2).

Using speech technology in the clinic begins with acquiring a suitable speech sample. Speech analysis can be adversely affected by noise in the speech signal (14). Clinicians need to be conscious of this influence, and be familiar with recording techniques that will optimise signal quality. The importance of choosing as quiet a clinical environment as possible, using unidirectional microphones close to the client's mouth, and following instructions regarding the use of specific microphones should be explained (15).

Digitisation of the acquired speech signal is an integral part of modern speech analysis. The fundamentals of digitising a speech signal, in particular the effects that digitising rate can have on the accuracy of analysis, need to be understood. For instance, the clinician should be capable of deciding that if spectral information about fricatives (where there spectral energy extends to around 12 kHz), sampling rate needs to be at least 24kHz to obtain accurate spectral information (2).

In modern speech analysis systems, the complex mathematical processing becomes invisible to the user, especially when "default" settings are used on the system to improve their user friendliness. Manufacturers of these systems go to great lengths to point out in the user manuals that default settings are not necessarily suitable for all clients under all conditions. The acoustic-phonetic effects of speaker and context, along with the exact requirements for particular feature extraction demand some level of customising. Inappropriate use of the technology will lead to meaningless outcomes and incorrect interpretation, and consequent disenchantment with technology as a viable clinical tool.

4. INTERPRETING THE OUTCOMES

Accurate interpretation of the information provided by speech analysis is critical to the successful integration of speech technology into the clinic. It relies on an understanding of the acoustics of speech production and perception, and on the body of knowledge about speech disorders. It should be acknowledged it may not currently be possible to interpret certain acoustic information in the light of our present knowledge about the acoustic characteristics of disordered speech.

Results of speech analysis in clinical settings are most often presented in one of three ways:

- direct displays of speech features eg wide and narrowband spectrograms, power spectra,

fundamental frequency contours, intensity contours.

- graphic displays to represent certain speech features such as an expanding balloon to represent increasing speech intensity (*IBM SpeechViewer*).
- tables of numerical data such as the calculated voicing parameters from Kay MDVP program, or the statistics page showing average nasalance values on the Kay Nasometer.

There are issues of interpretation that are specific to each of these types of data presentation.

4.1 Direct Representation of Spectral Features

Knowledge of acoustic phonetics will provide a familiarity with visual representations of speech, in particular the characteristics of spectrographic representations of the acoustic speech signal. However, it should be remembered that clinical speech analysis rarely occurs in ideal noise-free environments that are known to researchers in acoustic phonetics. The clinician will have to contend with acoustic interference from the environment. Clinicians need to be able to visually separate signal from noise, and familiarity with the acoustic characteristics will assist with that.

Professional credibility dictates that clinicians should be able to recognise and explain anomalous information by being aware of sources of error inherent in the speech signal itself (16), (17) and in the analysis process (16), (2). For instance, Lindblom (16) points out the difficulties associated with analysing speech that has a high fundamental frequency, an issue of particular importance to speech pathologists working with young children. It is important for clinicians to recognise the characteristics of the most commonly occurring problems that occur when analysing disordered speech.

Even when acoustic information accurately represents the speech signal, it is essential that clinicians critically appraise the findings in light of perceptual judgement and previous clinical experience. When measurements obtained from speech technology contradict aural perception, clinicians need to understand the potential sources of that contradiction. There are limitations to the human perception of speech system, just as there are limitations to information obtained from speech technology. Clinicians will only be able to properly assess the problem if they have a good grasp of the theoretical aspects of speech perception and speech analysis.

4.2 Representation of Spectral Features via Computer Graphics and Games

Representing speech features as computer graphics is an appealing concept for clinical applications. However, using graphics means that the entire process of feature extraction becomes invisible. The behaviour of graphics on computer screens in games and other visual feedback exercises is best understood (particularly when there is unexpected behaviour) in light of the principles of the underlying algorithms and analysis processes that generate the graphical displays. They provide the clinician with the necessary background to assess the accuracy of the information on the screen, thereby reducing frustration brought about by unreasonable expectations of the capabilities of the system. Most importantly, an understanding of underlying processes for visual feedback equips the clinician with sufficient knowledge to be able to explain to the client what speech features are affecting the visual display.

There is a need for systematic clinical evaluations of visual feedback systems that become available for clinical use. Published results for rigorous evaluation studies will assist clinicians in judging the potential clinical usefulness of particular systems.

To illustrate some of the difficulties that can occur when graphics are used as visual feedback, we will use some examples from two of our own studies. For instance, we found that the "Chart" mode of *Kay's Sonamatch* program has difficulty with automatic detection of the first two vowel formants for some speakers (18), detecting and plotting (F0, F1) instead of (F1, F2) on the visual display .The problems with automatic formant detection are still being actively pursued (19), and future versions of *Sonamatch* could well benefit from that research. In the meantime, it will be important for clinicians to be aware of the potential difficulty with *Sonamatch*.

Another study assessed the potential benefit of using visual feedback of nasalance to assist with nasality problems of deaf speech (20). During that study, McFarlane discovered an apparent discrepancy in the feedback obtained from averaged nasalance time traces and unaveraged traces. The discrepancy could be explained by the moving average calculation used for the averaged trace, but it caused initial confusion in the interpretation of the client's ability to keep nasalance values below a threshold value.

4.3 Numerical Information

Numerical information is most often interpreted in terms of available normative data, or cutoff values for a predetermined boundary between normal and disordered speech characteristics. Judging numerical information in this way requires knowledge of the origins of the normative data, and being able to assess whether or not it is suitable for a particular client. For instance, many speech parameters depend on factors such as age, gender, dialect, and phonetic context. Clinicians should ensure that normative data is appropriate for particular clients. Even then, interpretation of cutoff boundaries between the extremes of the normal range and the clinical population is contentious. For instance, there is debate whether cutoff measures for nasalance measures should be determined from normative data, or from sensitivity and specificity measures on clinical populations. In either case there appears to be an overlap of nasalance measures near the normal/abnormal cutoff boundary (21).

Another difficulty with derivation of speech parameters, particularly voice parameters, arises when algorithms are unsuitable for disordered speech, or become unreliable in the presence of a clinical environment (14).

5. CONCLUSION

The future of speech technology in clinical environments presents challenges for technologists and clinicians alike. For instance, the speech technology community will need to continue to assess the limitations of current techniques and algorithms, and refine them so that machine-generated errors in analysis are largely eliminated. The noisy environments found in clinics will be largely unavoidable, and so there is a need to develop analysis systems that are robust against noise. Speech technologists can draw on clinicians' experience, working with them to improve designs for graphical interfaces.

On the other hand, there are challenges for the speech pathology community on top of the need to have fundamental understanding of speech analysis techniques. These challenges include providing a more extensive knowledge base that demonstrates the acoustic characteristics of disordered speech, and their relationship to traditional perceptual descriptions. Assessing disordered speech often requires knowledge of normal speech characteristics, and there is still a large scope for development of normative data. Finally the speech pathologists have a professional responsibility to conduct systematic assessment of speech technology in clinics, and work with technologists to provide efficient, accurate, reliable and friendly systems.

6. REFERENCES

1. Farmer, A. "Spectrography" in Ball, M. and Code, C. (Eds) *Instrumental Clinical Phonetics*, Whurr Publishers, London, pp22-63, 1997.
2. Kent, R. and Read, C., *The acoustic analysis of speech*, Singular Publishing Group, San Diego, 1992.
3. Baken, R. and Daniloff, R. (Eds), *Readings in clinical spectrography of speech*, Singular Publishing Group, San Diego & Kay Elemetrics, Pine Brook NJ, 1991.
4. Buckingham H. and Yule, G. "Phonemic false evaluation: theoretical and clinical aspects", *Clinical Linguistics and Phonetics* 1, 113-125, 1987.
5. Shriberg, L. and Lof, G. "Reliability studies in broad and narrow phonetic transcription", *Clinical Linguistics and Phonetics* 5, 225-279, 1991.
6. McLeod, S., van Doorn, J. and Reed, V. "Homonyms and cluster reduction in the normal development of children's speech", *Proceedings of the 6th International Conference on Speech Science and Technology*, Adelaide, 331-336, 1996.
7. Scobbie, J., Hardcastle, W., Fletcher, P. and Gibbon, F. "Consonant clusters in disordered L1 acquisition. A longitudinal acoustic study", *Proceeedings of the XIIIth International Congress of Phonetic Sciences,* V2, 706-709, 1995.
8. Forrest, K., Weismer, G., Elbert, M. and Dinnsen, D. "Spectral analysis of target-appropriate /t/ and /k/ produced by phonologically disordered and normally articulating children", *Clinical Linguistics and Phonetics* 8, 267-281, 1994.
9. Gerken, L., & McGregor, K. "An overview of prosody and its role in normal and disordered child language", *American Journal of Speech-Language Pathology* 7, 38-48, 1998.
10. Flege, J. "Speech learning in a second language" In Ferguson, C., Menn, L. and Stoel-Gammon, C. (Eds) *Phonological development. Models, research, implications.* York Press, Timonium, 1992.
11. Fletcher, S. ""Nasalance" vs. listener judgements of nasality", *Cleft Palate Journal* 13, 31-44, 1976.
12. Kreiman, J., Gerratt, B. and Precoda, K. "Listener experience and perception of voice quality", *Journal of Speech and Hearing Research* 33, 103-115, 1990.
13. Crump, J. "High technology instrumentation and the speech-language pathologist", *Journal for Computer users in Speech and Hearing* 7, 83-87, 1991.
14. Ingrisano, D., Perry, C. and Jepson, K. "Environmental noise: A threat to automatic voice analysis", *American Journal of Speech-Language Pathology* 7, 91, 1998.
15. Tatham, M. and Morton, K. "Recording and displaying speech" in Ball, M. and Code, C. (Eds) *Instrumental Clinical Phonetics*, Whurr Publishers, London, pp1-21, 1997.
16. Lindblom, B. "Accuracy and limitations of Sonagraph measurements", *Proceedings of the Fourth International Congress of Phonetic Sciences*, 188-202, 1962.
17. Peterson, G. and LeHiste, L. "Duration of syllabic nuclei in English", *Journal of the Acoustical Society of America* 32, 693-703, 1960.
18. van Doorn, J., Shakeshaft, J., Winkworth, A., Hand, L. and Joshi, S. "Models of Australian English vowels for commercial visual feedback systems", *Proceedings of the ESCA workshop on Speech Technology in Language Learning (StiLL98),* Stockholm, 53-56, 1998.
19. Alvarez, A., Martinez, R., Nieto, V., Rodellar and Gomez, P. "Continuous formant-tracking applied to visual representations of the speech and speech recognition", *Proceedings of the 5th European Conference on Speech Communication and Technology,* Rhodes, Vol2, 653-656, 1997.
20. McFarlane, E. "*Visual feedback of nasalance with hypernasal hearing impaired speakers*", Unpublished Honours Thesis, University of Sydney, 1992.
21. van Doorn, J. and Purcell, A. "Nasalance levels in the speech of normal Australian children" *Cleft Palate-Craniofacial Journal,* 35, 287-292, 1998.