THE INTERACTIVE AUDITORY DEMONSTRATIONS PROJECT

Martin Cooke, Helen Parker, Guy J. Brown and Stuart N. Wrigley

Department of Computer Science, University of Sheffield
Regent Court 211 Portobello Street, Sheffield S1 4DP, UK

Email: {m.cooke,h.parker, g.brown,s.wrigley}@dcs.shef.ac.uk
http://www.dcs.shef.ac.uk/~martin

ABSTRACT

Topics in speech and hearing are well-suited to demonstrations using media other than the printed word. Currently, educators rely largely on passive formats such as the CD collections for general auditory psychophysics [6], auditory scene analysis [2] and cochlear damage [10]. Progress in programming tools and cheap, multimedia hardware now presents the potential to go much further. The *interactive auditory demonstrations* project aims to provide the user with an environment in which to explore the many phenomena and processes associated with speech and hearing. This promotes a much richer space of parameter manipulation than is possible via passive media. Further, the ability to initiate actions, repeat procedures and benefit from practically any kind of multimodal feedback enables a much wider range of learning possibilities. This paper focusses on the interface issues which are revealed by interactive exploration of the domain. A demonstration of linear prediction is presented to illustrate these issues.

1. INTRODUCTION

The interactive auditory demonstrations project aims to provide an exploratory environment to support speech and hearing education. Since late 1997, more than 25 MATLAB demonstrations have been developed and made available via the URL above. Figure 1 shows two example demonstrations while figure 2 lists some of the auditory phenomena and speech processing concepts produced to date. These represent a small fraction of phenomena and processes suited to this approach.

Previous papers [5][18] have described the motivation for this enterprise and the rationale for key implementation decisions (e.g. why MATLAB rather than Java). The purpose of the current paper is to examine the styles of interaction we believe are desirable for this domain. A newly-developed demonstration - linear predictive analysis of speech - illustrates these interface styles.

Our aim in discussing interaction is to go beyond generic issues such as consistency and robustness, although these are, of course, important and not straightforward criteria to satisfy. Similarly, we do not discuss speech visualisation techniques [4]. Instead, our focus is on the issues involved in producing tools which allow *direct manipulation* of relevant features in the domains of speech and hearing.

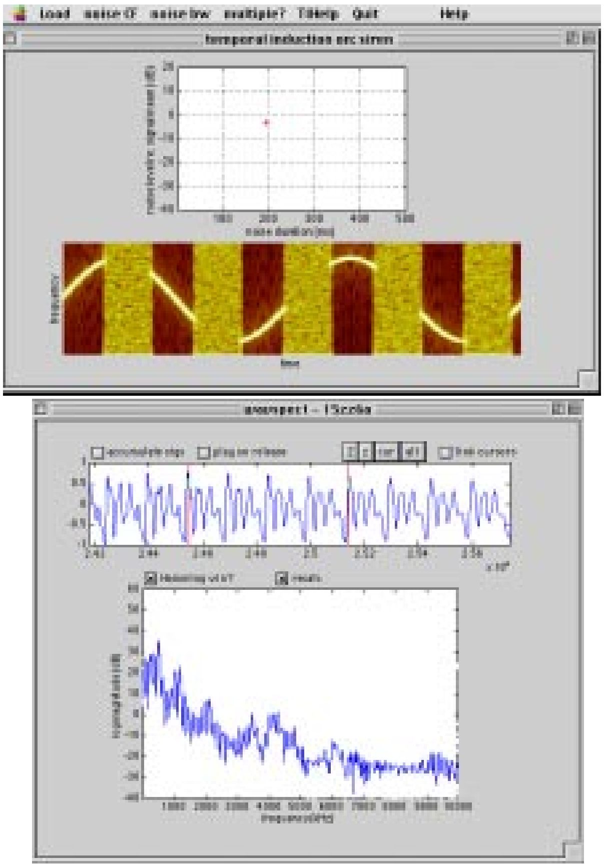

Figure 1. Top: Demonstration of auditory temporal induction [17]. Users click to hear the signal interrupted by noise bursts with duration and level specified by the grid location. Menus allow selection of signal type (speech, music, siren,...), noise properties and interruption schedule. Bottom: Demonstration of the link between the time and frequency domains. Users investigate the effect of window size, type and placement on the spectrum, which is continuously updated.

auditory phenomena	
audiometer	hearing level assessment
bmld	binaural masking level difference
bm	basilar membrane animation
detuning	mistuned harmonics
distortion	various distortions of speech
intmel	interleaved melody identification
streamer	auditory stream segregation
sws	cocktail party sine-wave speech
ti	temporal induction
vowelExplorer	double-vowel perception
speech processing	
auto	autocorrelation
ceplift	cepstral liftering
epd	endpoint detection
lpcspect	linear predictive analysis
polezero	pole-zero diagrams
timedom	time-domain speech processing
wavspect	short-term spectral analysis
vowelSeg	double-vowel segregation model

Figure 2. A selection of current demonstrations.

2. TOWARDS DIRECT MANIPULATION OF AUDITORY PHENOMENA AND SPEECH PROCESSING ALGORITHMS

Direct manipulation (Shneiderman, [15]) as a style of interaction is characterised by

- continuous representation of the world of action;
- immediate feedback of the results of user actions;
- rapid execution/reversibility of user-initiated actions;
- pointing, selecting and dragging replace execution through commands.

Shneiderman claims that direct manipulation supports novices through shortened learning times, expert users through speed of action and intermittent users, by enabling operational concepts to be retained. User anxiety is reduced by reversibility. Confidence and mastery follows from the fact that the user initiates actions and can predict responses. In short, the user is encouraged to ask ‘what if’ questions in the domain. Direct manipulation is the dominant style in most desktop environments, although the ideals stated above are not always achieved [12].

Direct manipulation appears well-suited to the domain of speech and hearing. It offers several advantages over a passive format such as audio CD demonstration.

1. It enables efficient exploration of the parameter space underlying a given phenomenon or process. For instance, the temporal induction demonstration provides access to a continuous representation of the two key factors in this effect: noise intensity and (to a lesser extent) duration. A CD is limited to a small number of (admittedly well-chosen) exemplars.
2. It allows users to draw inferences about relationships in the data. This feature is particularly suited to unfamiliar domains which are linked by seemingly complex transformations. In the time-frequency demonstration, key concepts such as time-frequency tradeoff and window placement/type are apparent in the relationship between user-initiated cause and immediate effect. The “relation between action and meaning is natural, straightforward and obvious” [7].
3. There are opportunities for both *articulatory directness*, which is concerned with the naturalness of action afforded by the interface to the user (e.g. cursors on a waveform invite movement) and *semantic directness*, in which a user recognises the meaning of objects represented at the interface (such as a static 2D graphic representation of waveforms or spectra) [7]. A further example of semantic directness which is particularly relevant in auditory psychophysics comes from allowing users to interact with a grid which is deliberately based on a familiar plot (e.g. hearing threshold, [11],p.51).
4. Tools can reinforce understanding via complementary multimodal ‘views’ of the data. In the temporal induction demonstration, the primary modality is sound, but visual feedback of the spectrogram helps to reinforce the critical notion of sufficient levels of occluding noise.
5. Further reinforcement can be obtained across demonstrations. For instance, important notions such as the effect of tapered windows and preemphasis on the spectrum can be conveyed wherever appropriate.
6. Displays persist and auditory stimulation is repeatable on demand with minimum effort. The pace of interaction and learning is under the control of the user.
7. For advanced study, the ability to link psychophysical stimuli and auditory representations is invaluable.

These potential benefits ought to lead to a deeper understanding of the domain. However, there are costs associated with direct manipulation and exploration.

1. It may not be possible to achieve response rates which are sufficient to guarantee immediacy and continuity. Lack of immediacy can lead to temporary asynchronous display of multiple representations. This would be a particular problem for tasks involving audio-visual synchrony. We explore the critical issue of timing below.
2. A danger inherent in opening up a wide parameter space is that the user gets bogged down in exploration and misses the critical regions. Devices which enhance important features and suppress less important information are required. One approach is to provide *detail-on-demand* [8] rather than have an overwhelming amount of constantly changing information to distract the user.

Timing requirements of direct manipulation

Psychologists consider three time scales - 0.1s, 1s and 10s - to be of importance in human information processing. The finest time scale is required for fusion of visual stimuli [3] (a somewhat smaller interval is required for auditory fusion). Responses to actions which are delayed by 0.1s appear to exhibit cause and effect. One second is the time required for a minimal dialogue interaction [13]. Time scales of 10s and above cover the period required to accomplish simple tasks. Interface engineers recognise that response times of computer systems need to be tuned to these human time constants [16].

Ideally, we would like computation and display to be complete within a time frame of the order of 0.1s. This rate of system response is achievable with the time-frequency demo, for instance. As the user moves the cursor, the spectrum is updated smoothly. This scale of immediacy allows the spectral response to influence the waveform selection process, which allows insights into pitch-synchronous analysis. Other demonstrations require significantly more background signal processing, leading to reduced opportunities for manipulation. However, it is important to identify the scope for immediacy at an early stage and design the interface accordingly.

3. LINEAR PREDICTION OF SPEECH

The issues described in section 2 were applied to the design of a tool for linear prediction of speech. This topic links a number of domains (time, frequency, z-plane) and involves a non-trivial amount of computation. The primary concepts to be supported include: the effect of LPC order on spectral smoothing, pole placement and the error signal; whitening of the error spectrum; reduction of residual error with increasing order; contrasting performance on voiced and unvoiced speech and the voicing information retained in the residual. Secondary concepts to be reinforced include the effect of window size, type and placement, the role of preemphasis, the relation between a signal and its Fourier spectrum, and the z-plane. The interrelatedness of these concepts makes them well-suited to the exploratory approach. Figure 3 depicts the evolution of the tool and outlines the manner in which direct manipulation issues were addressed.

4. DISCUSSION

We believe that applying principles of direct manipulation in an exploratory, multi-modal environment delivers deeper insights into speech and hearing than is possible through passive demonstration tools (although we stress that these demos are not intended to replace traditional teaching methods - an altogether more difficult task). We have yet to subject this claim to quantitative evaluation.

A wide variety of evaluation techniques are available for qualitative and quantitative analysis of interaction, ranging from direct to indirect observation and measurement of user/learner behaviour, to direct and indirect reporting by users on the experience of using our demonstrations. Pertinent components of a usability analysis are learnability, efficiency, flexibility and user attitude [14].

An analysis of users' patterns of data exploration with the demonstrations can be derived straightforwardly by recording user actions in a time-stamped log file. This will yield crude comparative indicators of depth of exploration and speed and ease of learning across users, learning exercises/ tasks and demonstrations.

Usability evaluation should also record users' subjective feelings of satisfaction with the demonstration environment. We hope to establish whether the present level of user control, engagement and the immediacy of response is sufficient to deliver the desired learning outcomes.

5. CONCLUSIONS

Speech and hearing provide fertile ground for interactive demonstrations of the form we argue for in this paper. Only a fraction of this territory has been explored to date (but see [1], [9]). Pitch perception, binaural processing and basic auditory psychophysics, for example, are ideally suited. We look forward to developing further demonstrations and welcome collaboration in this venture.

Acknowledgment: Some of the development was funded by the ELSNET LE Training Showcase, 98/02.

REFERENCES

[1] Bloothooft, G., van Dommelen, W., Espain, C., Hazan, V., Huckvale, M. & Wigforss, E. (1998) *The landscape of future education in speech communication sciences*, Utrecht Institute of Linguistics Publications ISBN 90-5434-069-X.

[2] Bregman, A.S. & Ahad, P., (1995) *Demonstrations of auditory scene analysis: the perceptual organisation of sound*, CD, MIT Press, Cambridge, Mass.

[3] Card, S. K., Moran, T. P. & Newell, A., (1983) *The psychology of human-computer interaction*, LEA.

[4] Cooke, M.P., Beet, S.W. & Crawford, M.D. (1993) *Visual representations of speech signals*, Wiley.

[5] Cooke, M.P. & Brown, G.J. (1999) 'Interactive explorations in speech and hearing', *J. Acoust. Soc. Japan*, in press.

[6] Houtsma, A.J.M., Rossing, T.D. & Wagenaars, W.M. (1987) *Auditory Demonstrations Compact Disc*. Acoustical Society of America.

[7] Hutchins E L, Hollan J D & Norman D A (1986) 'Direct manipulation interfaces', eds Norman & Draper, 87-124.

[8] Kreitzberg, C. B. (1991) 'Details on demand: hypertext models for coping with information overload' in Dillon M (ed), *Interfaces for information retrieval and on-line systems*, Greenwood Press, 169-176.

[9] MATISSE (1999) *Methods and Tool Innovations in Speech Science Education*, London, April.

[10] Moore, B.C.J. (1995) *Perceptual consequences of cochlear damage*, Oxford (accompanying CD available).

[11] Moore, B.C.J. (1997) *An Introduction to the Psychology of Hearing, 4th edition*, Academic Press.

[12] Mullet, K. & Sano, D. (1995) *Designing visual interfaces*, SunSoft Press, Prentice Hall

[13] Newell, A. (1990) *Unified theories of cognition*, Harvard University Press,

[14] Shackel, B. (1990) 'Human factors and usability' in Preece J & Keller L, eds., *Human-computer interaction: selected readings*, Prentice Hall.

[15] Shneiderman, B. (1983) 'Direct manipulation: a step beyond programming languages', *IEEE Computer, 16 (8)*, 57-69.

[16] Shneiderman (1998) 'Response time and display rate', in *Designing the user interfac*e, Addison Wesley.

[17] Warren, R.M. (1970) 'Perceptual restoration of missing speech sounds', *Science*, 167, 392-393.

[18] Wrigley, S.N., Cooke, M.P. & Brown, G.J. (1999) 'Interactive learning in speech and hearin*g', Proc. MATISSE Workshop*.

1. A **paper prototype** defined the content and opportunities for direct exploration.

2. Initial development consisted of reuse of the time-frequency demonstration and assessment of timing constraints. Unlike the time-frequency tool, continuous update of LPC-derived representations such as smoothed spectra could not be achieved. Instead, recalculation on cursor release was implemented, resulting in a few seconds wait for derived representations.

3. The next phase involved **specialisation of the interface** to convey multiple derived representations (waveform segment, error signal, DFT and LPC-smoothed spectra, error spectrum, pole locations).

4. Simplification without loss of functionality resulted from a single plot for time domain signals and a further plot for frequency domain. Colour distinguishes raw signals from LPC representations. Experienced designers depend heavily on trial and error to determine which elements are truly essential [12] p41.

5. The **final design** incorporates

- the addition of 'background' exploratory features such as click to play signals (including the residual), click to reveal pole frequency ('detail on demand')
- a tidying and grouping of similar interface elements
- the use of font size to distinguish important information (display titles) from the less important (duration of current selection, mean square error of residual)
- removal of uninformative axis information

The final interface is sparse, free of unnecessary decoration and confusing icons, yet includes all the desired content.

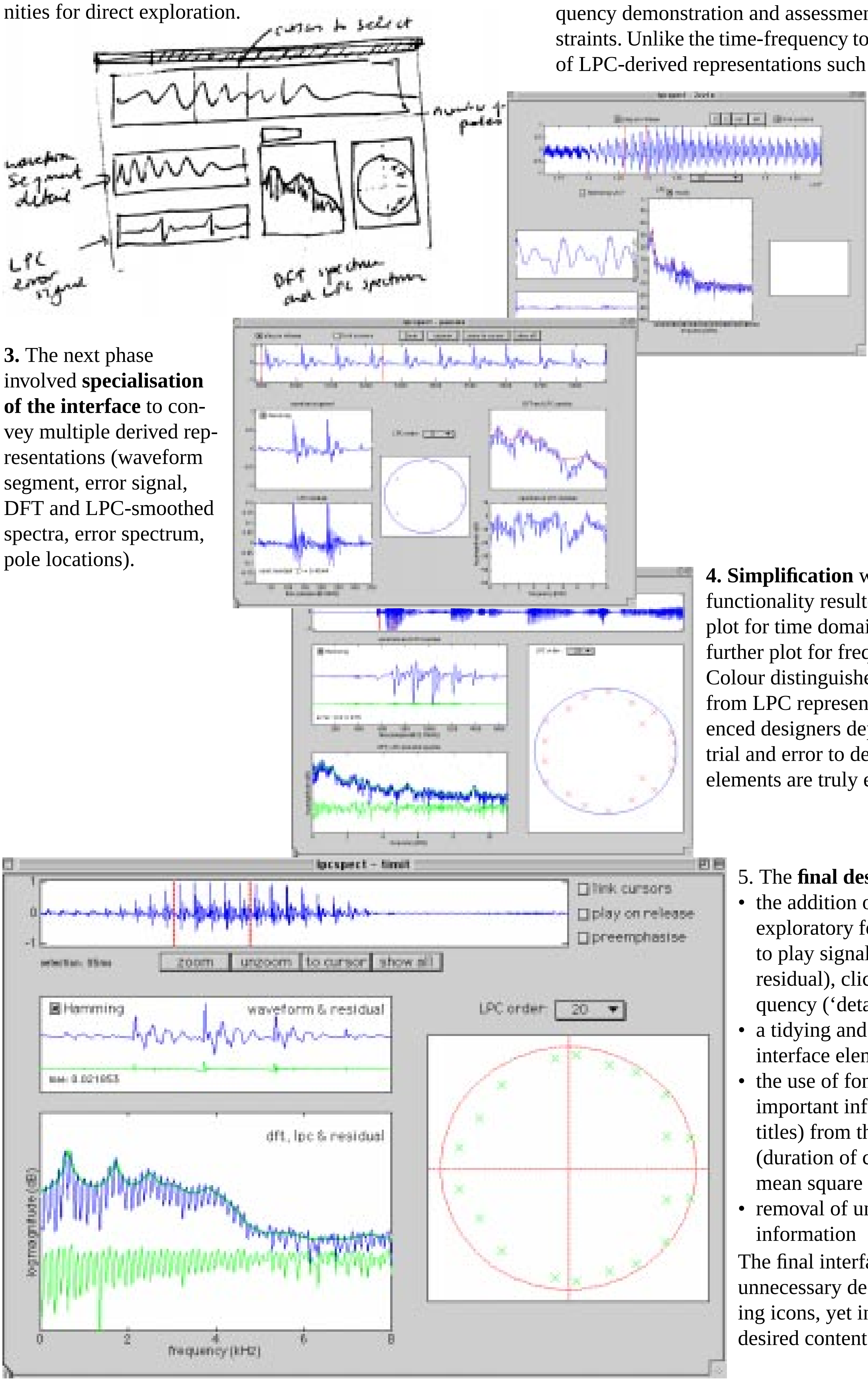

Figure 3. Evolution of a tool for exploring linear predictive analysis.