SEEING IS BELIEVING: BEYOND A STATIC 2D-VIEW OF FORMANT SPACE FOR SPEECH RESEARCH AND EDUCATION

Michael Barlow and Frantz Clermont
{spike,frantz}@cs.adfa.edu.au
School of Computer Science
University College, University of NSW

ABSTRACT – The paper describes an online resource developed for the dual purpose of education and research in speech science and technology. The resource consists of three-dimensional (3D) interactive worlds, which are currently based on formant databases and can be used not only to demonstrate familiar phenomena but also to gain new insights from within less constrained spaces. The resource's construction and availability are described, together with examples of some of the phenomena manifest in spoken vowels. Results of employing the resource in an undergraduate course in speech processing suggest that the 3D interactive approach not only is more appealing and natural than two-dimensional (2D), numeric approaches to teaching the same material, but it also enhances learning.

INTRODUCTION

Education of new speech researchers in the fundamentals of speech science and technology is a requisite of on-going, quality speech research. Regardless of whether the researcher will focus on basic issues in speech science or work in the more applied area of speech technology, a firm understanding of the principles that underpin our field is necessary. Recently, this need for quality material in the education of a researcher has been explicitly recognised with the publication of tools and material for speech education (Loizou, 1999; Cooke, 1999).

One desirable aspect of that dual endeavour (education and research) is the availability of computer tools for observing data without the restrictions imposed by the more traditional 2D-space. This issue of Scientific Visualisation arises in many disciplines, where visual representations are expected to yield further or better insights into the rich (in both dimensionality and scale) data resulting from experimentation, simulation, and analysis. In the area of speech research, this is often exemplified by the quest for a deeper meaning of the acoustic parameters employed, and for interpreting the latter in relation to speaker, phone, dialect or other sources of variability.

Until quite recently, visualisation in 3D has required specialised and expensive hardware and software, putting it out of the reach of the normal computer user. However, the last 2 years of developments in processing power of the PC and its peripherals (such as graphics cards), maturing standards for 3D data descriptions, and publicly available software, combined with the ever-increasing penetration of the internet, have shifted visualisation to a point where it is now a tool available to any user. These technical advancements, coupled with the constant move towards making education material available through the World Wide Web (an example of distance education), today imply that it is practical to incorporate scientific visualisation of real data as a component of a university-based course. Our contributions to the development and the preliminary assessment of such tools form the basis of the present paper.

First, we describe the construction of interactive, online, 3D worlds (interactive 3D spaces) through which the user may move and interact with the objects occupying that space. Over thirty 3D-worlds were constructed in which the user is able to visualise and interact with vowel and diphthong formant data, illustrations of which are offered in Figs 1 and 2.

The second and the third part of this paper describe a preliminary evaluation of the 3D VRML worlds constructed thus far, in which users are free to move through and interact with the data, viewing it from any perspective or distance. In particular, we illustrate the potency of unconstrained navigation in 3D-space by giving some fresh insights into the third-formant dependence of idiolectal differences in Australian English and of the non-linear movements of certain diphthongs. The pedagogical usefulness of our 3D formant worlds is also discussed in the context of an undergraduate course in speech processing which we teach at our School of Computer Science.

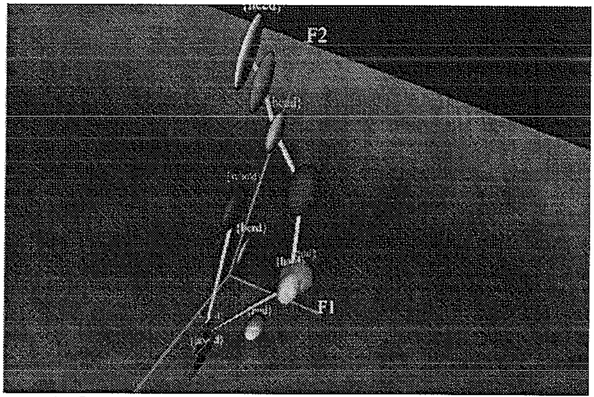


Figure 1. Snapshot of 3D VRML World of the steady-state vowels of Australian English. Each spheroid (labelled and colour-coded) represents a vowel centred at its mean {F1, F2, F3}-values with variance in those dimensions represented by the size and shape of the spheroid. The vowels are also inter-connected to yield the "classic" triangular-shaped polygon. Users can freely move through and interact with such worlds.

FORMANT DATABASES

The three lowest formant-frequencies (F1, F2 and F3) of spoken vowels have been selected to construct our current worlds, bearing in mind that they not only afford small dimensionality but they also lend themselves to meaningful interpretations in the acoustic, the articulatory and the perceptual domain. In this sense, therefore, the vowel-formant space embodies powerful properties that are useful both for educating the neophyte and for assisting the experienced researcher in the quest for a better understanding of speech phenomena. Three databases of such parameters have been adopted, which convey different perspectives on major determinants of speech variability.

First, we acquired (Watrous, 1991) a copy of the time-honoured database resulting from Peterson and Barney's classic study of the American English vowels spoken by 15 Children, 28 Women and 33 Men. Second, we used a 36-male speaker database restored (Clermont, 1996) from Bernard's (1967) pioneering study of the three varieties (Broad, General and Cultivated) of Australian English vowels. For the purpose of illustrating certain dynamic aspects of vocalic sounds, a 4-male database of diphthong formants (Clermont, 1991) was also incorporated into our choice of worlds.

VISUALISATION TOOL

The 3D-worlds were implemented in VRML (Virtual Reality Modelling Language) - a standard for 3D (ISO/IEC, 1997) on the World Wide Web - for which browser plug-ins are freely available. VRML is a 'scene description language' and supports the paradigm of the user occupying, moving through, manipulating, and interacting with the 3D world constructed.

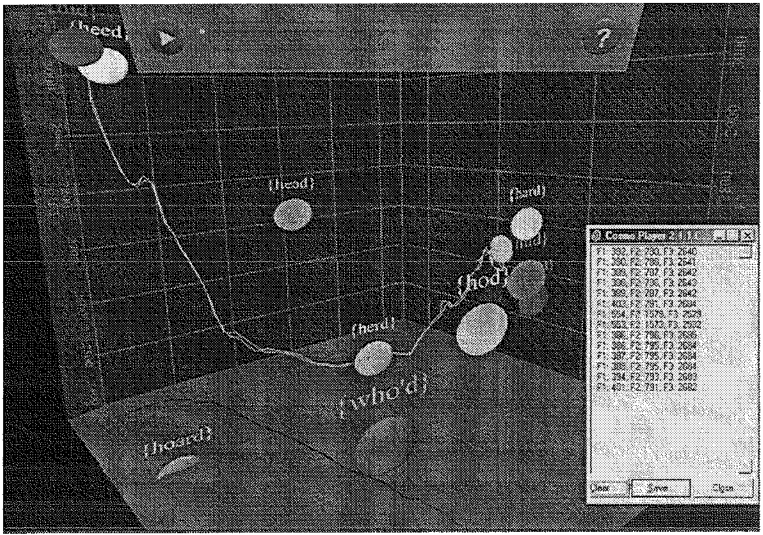
VRML programs, known as worlds, consist of specifying a scene graph – a hierarchy of objects each with their own geometry (spatial structure), appearance (colour, material etc.) and associated behaviour. The viewer's motion through and interaction with the world is supported directly by the browser, thereby freeing the programmer to describe the objects and behaviour of the world alone.

Amongst the features supported by VRML are interpolators (used to alter colour, position and shape for animations), sensors (touch and proximity, used to allow programmed behaviour when the user interacts with objects), sound sources (used to spatially encode sound), and a script object for programming in Java or Javascript (used to perform calculations and more complex behaviours). All these features were employed to enhance the dynamic aspects of the worlds created as models of the formant data.

VRML worlds were generated by a combination of direct programming in VRML and a suite of programs in MATLAB and Java,' which processed the raw formant data and generated output descriptions of the data in VRML. Software engineering principles such as modularity and code reuse were pursued not only through the development of MATLAB and Java code but a core set of VRML prototypes: a feature somewhat analogous to functions.

3D WORLDS & SPEECH PHENOMENA

Because VRML is an open standard for which browser plug-ins are freely available, it is entirely possible to take the highly desirable step of making the constructed worlds available to a large audience. The worlds described in this paper may all be found at the URL: http://www.cs.adfa.edu.au/~spike/Research/VisualSpeech/. The individual worlds exist as links from the above page, with each link grouped under a broad category (e.g., "Dialect Differences") and having a short description of what may be found by loading the world. All worlds share a set of common features such as orienting axes, pre-programmed viewpoints (traditional F1-F2 view, 'amongst the data' etc.) and a programmed tour, which flies the viewer around the data when a cube is clicked on.



```
F1: 392, F2: 790, F3: 2640
F1: 390, F2: 788, F3: 2641
F1: 389, F2: 787, F3: 2642
F1: 390, F2: 786, F3: 2643
F1: 389, F2: 787, F3: 2642
F1: 403, F2: 791, F3: 2634
F1: 554, F2: 1379, F3: 2529
F1: 563, F2: 1573, F3: 2552
F1: 386, F2: 798, F3: [illegible]
F1: 386, F2: 795, F3: 2684
F1: 387, F2: 795, F3: 2684
F1: 388, F2: 795, F3: 2684
F1: 394, F2: 792, F3: 2683
F1: 401, F2: 791, F3: 2682
```

Figure 2: Snapshot of formant space for an adult-male speaker of Australian English, which shows how the non-linear transition through the diphthong /oi/ (as in "boy") traverses the neighbourhood of certain vowels. The panel situated at top of the world permits control of the diphthong's animation, while the pop-up window at right gives a listing of {F1,F2,F3} measurements initiated by "touching" one of the vowel spheroids.

The great strength of the VRML worlds lies in the opportunity afforded by the user to view, to move through, and interact with, the data from any perspective. Clearly, static 2D views and textual descriptions cannot do the process justice. However, two brief illustrations may serve to show the potential offered by this approach, while the following section provides some concrete data via analysis of the tool's usage in a an undergraduate course.

Several of the constructed worlds employ animation and dramatically illustrate issues of linearity and non-linearity in formant space. For example, the worlds constructed for the diphthongs of a single speaker (see Figure 2) show the steady-state monophthongs for a speaker as well as the paths followed by that speaker's realisation of a diphthong. That the diphthong's dynamics manifests itself as a non-linear transition through a specific monophthong sequence (Clermont, 1991; 1993) is easily observed from a range of viewpoints inaccessible in 2D. Similarly, other worlds based on the Bernard database animate transitions from the Broad, through the General, to the Cultivated continuum of spoken vowels in Australian English. The animations show that complex, non-linear changes occur on an individual vowel basis when moving from one end of the continuum to the other. Some of the consequences of age and gender differences can also be studied in the world constructed from the Peterson & Barney database of American English vowels. Indeed, the transitions animated between the mean monophthong positions for the three groups provide a powerful illustration of the well-known inverse relationship between vocal-tract length and mean formant values (Fant, 1960; Stevens, 1971).

Another important theme exemplified in a number of the worlds currently available is that of inter-speaker differences as a source of variability in the speech signal. For instance, the inherently complex variability arising from multi-speaker realisations of the same monophthong and the regions of accentuated overlap can be easily apprehended, together with a better appreciation for the detrimental confusions in automatic speech recognition.

USE IN AN UNDERGRADUATE COURSE

We have thus far argued for a new approach to speech-parameter visualisation that extends beyond traditional, static 2-D representations and, in this regard, our 3-D VRML worlds have above been purported to afford spatially unconstrained and interactive observations. We have also implied the claim that these worlds should be pedagogically effective in the sense of enhancing learning, or at least in Chickering et. al.'s (Chickering & Gamson 1987, Chickering & Ehrmann's 1997) sense of "encouraging active learning". Whether this claim can be substantiated in practice is a relevant question, which we have attempted to address by inviting into our 3-D worlds 17 undergraduate students of Computer Science, and by interpreting the students' observations and performance in carrying out three consecutive tasks.

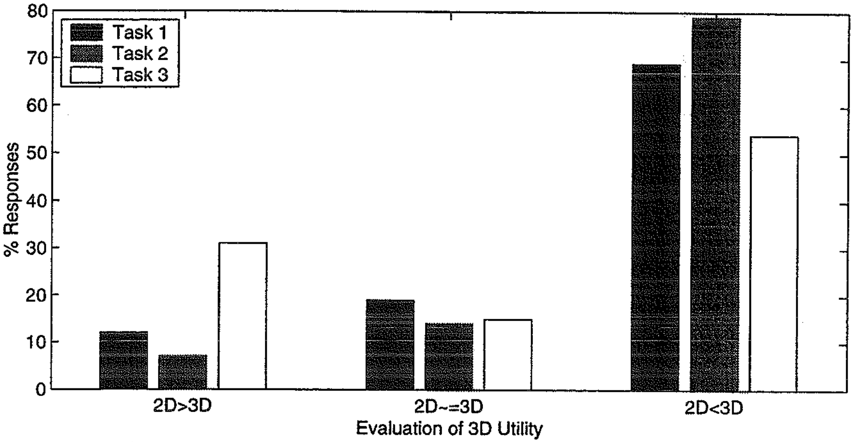


Figure 3: Student responses to the use of the 3D worlds in the three assignment tasks.

Task 1 consisted of re-examining, in an animated 3-D world, the same Peterson & Barney data, which had been employed in a previous assignment for superimposing vowel quadrilaterals in the F1-F2 plane and studying therein between-gender differences. For Task 1, students were then asked to re-assess gender differences in the extended F1-F2-F3 space of American English vowels. By contrast, Tasks 2 and 3 implicated data that had not been previously seen by the students. For Task 2, they were asked to identify major differences manifest in a quadratic surface, which was shown shifting gradually from the Broad to the General and then to the Cultivated varieties of spoken vowels in Australian English. For Task 3, the students were for the first time exposed to dynamic behaviours of Australian English /ai/ and /oi/ as illustrated in Figure 2, and thus asked to identify two major differences between these back-to-front diphthongs. Note that, in addition to their written answers, the students were requested to submit for each task an accounting summary of the number and the elapsed time of their viewing sessions, together with a preference vote (2D>3D, 2D~=3D, 2D<3D; summarised in Figure 3) based on each task accomplished. The 3 tasks were equally weighted, and the accounting summary contributed 20% towards the mark given for each task.

The results thus collected first indicate that 2 viewing sessions were needed by nearly all students in order to accomplish any of the 3 tasks. The first session was reported by students to have consisted of mainly familiarising themselves with the 3-D world and its navigating tools; the second session was then devoted to answering the questions proper. It is therefore not surprising that the elapsed times recorded by nearly all students show that the second viewing sessions were often shorter than the first ones, although the data are also consistent in highlighting that Task 3 in particular took longer to complete than the other tasks.

It is instructive to observe (see Figure 3) that Task 3 also attracted the lowest preference (54%) for a 3-D representation, by contrast with 69% for Task 1 and 79% for Task 2. It thus appears that previous 2D familiarisation of certain phenomena may be necessary and that *ab initio* exposure to a 3-D representation may be counter-productive. It should however be acknowledged that the highest 3-D preference of 79% was expressed for the second task, which admittedly involved continued 3-D exposure from, but was not necessarily easier than the first task. Insofar as preference votes are interpretable as global indicators of visual appeal and significance in the learning process, then it seems justified to speculate that our 3-D worlds do have a significant role to play.

Finally, we should be asking ourselves whether the profiles of marks recorded for the three tasks reflect any sensitivity to the use of 3-D representations. While the sample size is admittedly small in terms of the number of students and while the degree of difficulty may be perceived differently from task to task, the profiles exhibit certain features that appear to be systematic and are therefore worth interpreting in view of our earlier claim. For example, Task 1 yielded the most even distribution of marks and not necessarily the highest marks, while Task 2 yielded a somewhat skewed distribution towards the maximum mark and Task 3 clearly yielded a distribution skewed towards the maximum. This dichotomy between Task 1 and Tasks 2 and 3 could suggest the following interpretations, namely: (i) that double familiarity (2-D and 3-D) of the phenomena at hand as for Task 1 tends to cause greater homogeneity of correct answers; (ii) that the introduction of data describing a phenomenon previously unseen may require continued exposure to our 3-D worlds in order to achieve consistency or exactness in concept learning.

CONCLUDING DISCUSSION

The paper has detailed the construction and availability of a number of interactive 3D worlds, freely available upon the World Wide Web. These worlds are constructed from well known formant databases and illustrate a number of fundamental phenomena. The worlds have been used successfully in an undergraduate course on speech processing.

While all worlds described in the paper were constructed on the basis of formant data, there is no actual limitation on the data visualised. Linked from the same URL as the formant data can be found a number of 3D visualisations of the Vocal Tract, constructed from area function data (Barlow, Clermont, & Mokhtari, 2000). Similarly, work is currently in progress to visualise cepstral coefficients and other such parameters directly employed in recognition tasks. In these higher dimensional spaces a combination of principle component analysis, selection of a subset of the parameters, and encoding dimensions above three through texture, colour, and shape is being employed. Further, while most of the data visualised (diphthongs being the exception) was inherently static in nature, it is strongly felt that the true strengths of the approach will lie in visualising time-varying data.

One limitation of the current approach is that the process of 3D world generation is semi-automatic at best, with the world-builder often needing to know VRML syntax and semantics in order to achieve the more dynamic effects. This leaves most users with only the set of pre-generated worlds. In order to alleviate this problem a Java application and suite of related classes is being developed to facilitate the automatic generation of the VRML worlds for any set of data. This has several advantages as it is not only platform independent (a key feature of Java) but it places the visualisation of data directly in the hands of the user: students and researchers will be free to generate and explore 3D worlds for any data they possess. Indeed, due to its generality the application can be used to visualise any multi-dimensional data, not merely that from the speech domain.

A further area of development currently being explored is the display of the VRML worlds in an immersive virtual environment known as the WEDGE (Gardner et. al. 1999). It is hoped that the immersive aspect will enhance the learning experience and potentially offer new insights in an environment in which the user may truly interact with the data in 3D.

One of the inherent limitations of VRML is the lack of computation ability and the corresponding difficulty in providing greater dynamic control of existing worlds. Planned future work will explore the Java 3D API, an emerging standard for 3D that includes all the benefits of Java, including running as an applet inside a web page.

It is also planned to pursue our pedagogical evaluation with a view towards strengthening the preliminary evidence presented in this paper, which seems to point to interactive, 3-D representation of speech parameters as a powerful learning tool.

REFERENCES

Barlow M., Clermont F., & Mokhtari P., (2000), "From Acoustics of Speech to a 3D Vocal-Tract: Toward a Plausible Model with Real-Time Constraints", Proc. Eighth Aust. Int. Conf. On Speech Science and Technology, in proceedings, Canberra.

Bernard J.R.L. (1967), "Some measurements of some sounds of Australian English", PhD Thesis, Sydney University.

Chickering A.W. & Gamson Z.F., (1987) "Seven principles for good practice in undergraduate education", AAHE Bulletin, 39(7), 3-7.

Chickering A.W., and Ehrmann (1997), "Implementing the seven principles: Technology as Lever", http://www.aahe.org/ehrmann.htm

Clermont F., (1991), "Formant-contour models of diphthongs: A study in acoustic phonetics and computer modelling of speech", PhD Thesis, Australian National University.

Clermont F., (1993), "Spectro-temporal description of diphthongs in F1-F2-F3 space", Speech Communication 13: 377-390.

Clermont F., (1996), "Multi-Speaker formant data on the Australian English vowels: A tribute to J.R.L. Bernard's (1967) pioneering research", Proc. Sixth Aust. Int. Conf. on Speech Science and Technology: 145-150.

Cooke M., & Brown G.J. (1999), "Interactive explorations in speech and hearing", J. Acoust. Soc. Japan (E) 20(2): 89-97.

Fant G., (1960), *Acoustic Theory of Speech Production* (Mouton, The Hague, The Netherlands).

Gardner, H.J., Boswell, R.W., and Whitehouse, D. (1999). "The WEDGE Emmersive Projection Theatre", *Proc. 4th International SimTecT Conf.*: 383-385.

ISO/IEC (1997) "VRML 97", International Specification ISO/IEC IS 14772-1, www.vrml.org.

Loizou P.C., (1999) "COLEA: A MATLAB Software Tool for Speech Analysis", http://www.utdallas.edu/~loizou/speech/colea.htm.

Peterson G.E., & Barney H., (1952), "Control methods used in a study of the vowels", J. Acoust. Soc. Am. 24: 175-184.

Stevens K.N., (1971), "Sources of inter- and intra-speaker variability in the acoustic properties of speech sounds", Proc. Seventh Int. Congr. Phonetic Sciences: 206-232.

Watrous R.L., (1991), "Current status of Peterson-Barney vowel formant data", J. Acoust. Soc. Am. 89: 2459-2460.