ReCALL **16** (1): 103–113. © 2004 Cambridge University Press
DOI: 10.1017/S0958344004000813 Printed in the United Kingdom

Computer generated feedback on vowel production by learners of English as a second language

DAVID BRETT
Facoltà di Lingue e Letterature Straniere, Università degli Studi di Sassari, Via Roma 151, 07100 Sassari, Italy
(*email: bellabrett@tiscali.it*)

Abstract

The acquisition of new vowel sounds is one of the most challenging tasks for the adult learner of a second language. It is also an aspect which requires a great deal of time and individual attention. Learners of English are particularly hampered by the irregular orthographic representation of vowel sounds. This paper reports on research in progress into the development of an application to provide learners with real time feedback on their vowel production. This process includes the calculation and plotting of formants on a graph using the PRAAT programme. The patterns which result are highly reminiscent of the traditional vowel chart, which in turn reflects articulation, i.e. articulatory changes are reflected on the graph. This application can provide feedback on both pure vowels and diphthongs. Some of the difficulties and obstacles which as yet prevent it from being a tool which can be used by the learner autonomously will also be outlined and discussed. Foremost amongst which are: the question of speaker normalisation (how to compensate for the difference in values between male, female and pre-pubertal speakers) which may be resolved by way of statistical elaboration of the formant data; and the methods used to isolate the vowel sound from the rest of the signal.

1 Introduction

The acquisition of new vowel sounds is one of the most challenging tasks for the adult learner of a second language. It is also an aspect which requires a great deal of time for the learner and individual attention from teachers. While the first requirement depends very much on the learner's motivation, the second is very often beyond his or her control, and in most learning contexts the amount of time that can be dedicated to the practising of vowel sounds with individual students is normally extremely small, if not entirely non-existent. Therefore it seems natural to resort to the use of computers, to Computer Assisted Pronunciation Teaching (CAPT) to help learners improve this aspect of their pronunciation. Yet, while much work has been done in the field of perception (see Ingram & Park (1997) and Escudero (2002)), that of evaluation of speech production remains to be fully explored.

https://doi.org/10.1017/S0958344004000813 Published online by Cambridge University Press

Although this study arises from needs which have emerged in the context of learning English as a second language at tertiary level in Italy (i.e. aged 18 upwards), it is quite probable that many readers working in different contexts will identify readily with the following. Native speakers of Italian have quite a difficult task to face when confronted with the vowel repertory of English (Standard British English is the variety being referred to here and henceforth) as it consists of a number of pure vowels that is almost double that of Standard Italian, twelve (Roach, 1991: 14–19), as opposed to seven (Canepàri, 1992: 54), respectively, and there is an analogous situation in terms of diphthongs. To put it simply, Standard Italian, like a number of other widely spoken languages such as Arabic and Japanese (International Phonetics Association, 1999), lacks central vowels. The native speaker of Italian must hence learn to distinguish among a large number of sounds which are to all intents and purposes exotic. It is obviously not sufficient to perceive these differences: the sounds must also be produced effectively.

Yet the complexity of the matter does not end here: it is a well-documented fact that learners of English are severely hampered by the irregular orthographic representation of vowel sounds. For an Italian speaker, who has learnt to read and write using a highly predictable phonographic system this is all very hard to comprehend. The end result is that the majority of Italian speakers of English, even at tertiary level, have rather vague ideas about the nature of the English vowel repertory. Great difficulty is found, for example if the learner is asked to group words together according to their vowel sounds. While individuals who learn the language for basic uses, e.g. travel, essential communication at work etc., cannot be reasonably expected to display a complete grasp of the phonological system, those in need of greater linguistic skill, e.g. those who use the language extensively in the workplace, non-native speakers who intend to teach the language, university students majoring in the language, etc., may not be able to do without firm knowledge of the sound system.

2 The analysis of formants

The present study examines a system for providing feedback on learners' vowel production based on the analysis of formant data. For those unfamiliar with the topic a brief explanation may be necessary: analysis of the acoustic qualities of vowels shows peaks at certain frequencies. The frequency at which these peaks appear differs from one

Table 1. *The mean formant values of pure vowels of a female BBC broadcaster. Data from Deterding (1997). The first two formants are normally considered sufficient for the distinction of vowels in the repertory of Standard British English. The third formant becomes of relevance when distinguishing between front spread and rounded vowels, that can be found in languages such as French and German*

Vowel sound	Mean F1	Mean F2	Mean F3
heed	304	2664	3248
had	1067	1690	2791
who'd	333	1529	2657

https://doi.org/10.1017/S0958344004000813 Published online by Cambridge University Press

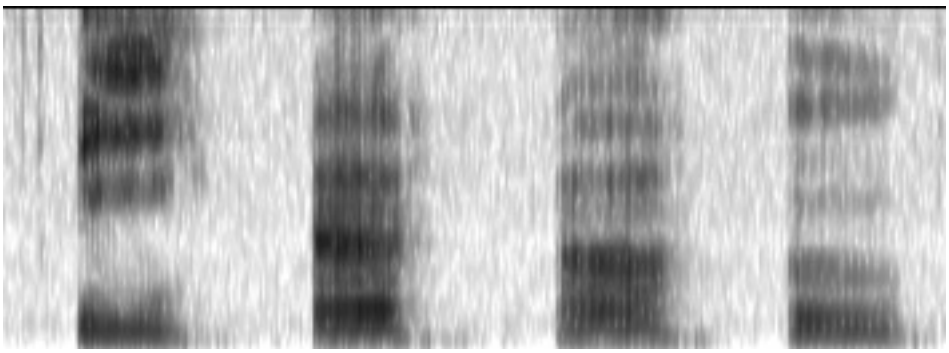


Fig. 1. Spectrogram of four English vowels produced by a male native speaker: the *beat*, *bat*, *Bart*, and *boot* sounds; formants visible as dark bands; x-axis= time, y-axis=Hz (0-5000).

vowel sound to another (Figure 1).

Deterding (1997) in a landmark study measured the formant values of pure vowels produced by five male and five female BBC broadcasters. As an exemplification the results for one female speaker are displayed in Table 1.

Furthermore, the values when plotted inversely (i.e. as minus values) on a graph with F2 and F1 on the x and y-axes, respectively, bear a distinct resemblance to the traditional vowel chart (Figure 2), which in turn is directly connected to articulation. In other words, visual information, extracted from a vowel sound produced, bears a direct (albeit inverse) relationship to the articulatory position the speaker adopted when producing the sound, and vice versa.

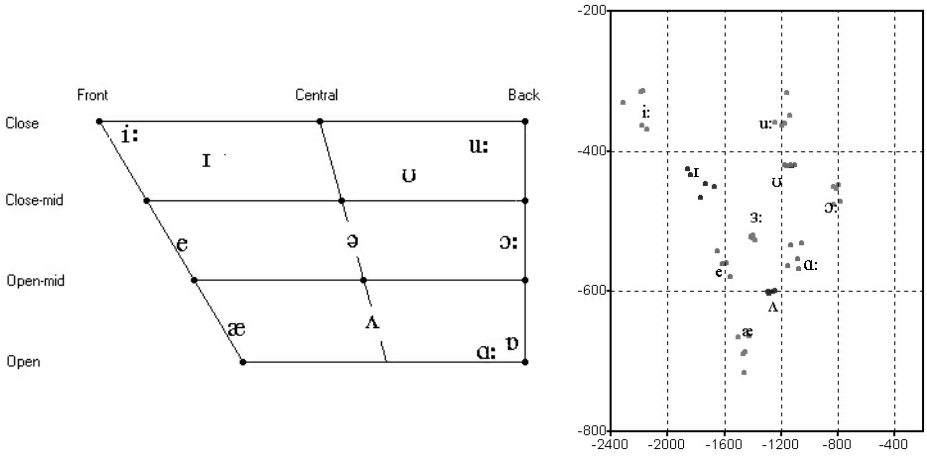


Fig.2. Comparison between the vowel chart (left) and a graph (right) plotting formant values of a male native speaker of Standard British English (x-axis = -F2 (Hz), y-axis = -F1 (Hz)).

https://doi.org/10.1017/S0958344004000813 Published online by Cambridge University Press

2.1 Programs for formant analysis

This paper describes a project that endeavours to provide reliable, clear and useful feedback on vowel production to learners of English as a second language. While there is a discrete number of software packages available for the analysis of speech, the vast majority are not suitable for two main reasons:

1. Several freeware programs for the elaboration of spectrograms, e.g. Speech Analyzer (Summer Institute of Linguistics) and WASP (Department of Phonetics, University College London) do not allow in-depth study and treatment of formant values. While Speech Analyzer does provide a facility for the plotting of various combinations of formant values, it is not particularly reliable and furthermore does not allow formant values to be exported.
2. Two programs have applications specifically developed for the elaboration of formant values, SonaMatch (Kay Elemetrics) and Dr. Speech (Tiger Software) – these are of excellent quality but their cost is an obstacle to most language centres. This is particularly true of SonaMatch. For an in-depth description of the application of this programme to CAPT see Molholt (1998).

2.2 The PRAAT program

This study has been carried out using PRAAT, a program that can be downloaded free of charge from www.praat.org. This program has a huge number of possibilities for the analysis of speech, and has the further advantage of being easily modified for specific research purposes – results can also be exported to Excel-compatible spreadsheets. Use was made of PRAAT's inbuilt Linear Predictive Coding (LPC) formant analysis function and scripts were written to make a large number of steps automatic in order to make the use of the program as straightforward as possible.

The following is a brief description of the working of one of the scripts for the analysis of pure vowels:

1. The learner chooses the type of analysis to be conducted, e.g. MALE PURE VOWEL SYMBOL PLOT, and a menu with radio buttons appears where the symbol of the sound to be practised is selected.
2. On pressing the OK button the learner hears a recording of a native speaker producing the chosen sound and immediately afterwards has 1 second to reproduce the sound.
3. The signal is then analysed in terms of intensity to isolate the vowel sound from the rest of the signal, i.e. the time slots for which intensity is below a certain figure will be discarded. Once extracted the vowel sound is analysed at intervals of 0.01 second. and the formants at each point identified. The mean of the values for each formant is then calculated.
4. The values are then plotted on a graph, in this case plots are indicated by the phonemic symbols of the chosen sound; use may also be made of coloured dots, that are more practical when a large number of sounds is analysed.

The whole process lasts about 3–4 seconds, giving swift feedback on the learner's pro-

https://doi.org/10.1017/S0958344004000813 Published online by Cambridge University Press

duction. However, several problems remain to be resolved:

1. Since the isolation of the vowel sound is carried out on the basis of intensity, great attention must be paid to avoid contamination of the sample by external noises, either from the environment or the hardware used (this is especially important if headsets are used).
2. The problem outlined in (1) also influences the choice of whether to analyse a vowel produced in context or as an isolated sound. A vowel sound is normally the segment with the greatest intensity in a syllable, however some phonemes, e.g. voiceless fricatives, can often exceed the vowel in intensity and hence will be extracted along with the vowel, clearly falsifying the reading. Furthermore, vowels in context, even if classified as pure vowels rarely show constant values as the influence of preceding and succeeding articulations is great. For this reason, if sounds in context are to be analysed it is advisable that minimal pairs be used, in so much as is possible.
3. The feedback after 3–4 seconds referred to above, while swift, is certainly not real-time. Feedback would clearly be more effective if this lapse of time were reduced considerably and updated continuously. For discussion of how this may be obtained see Section 3, which deals with further development.
4. The values that are plotted in this way are relative – one cannot, unfortunately,

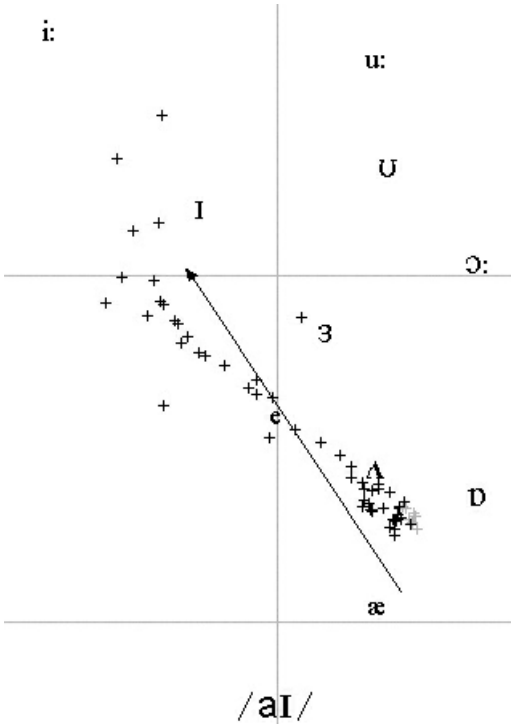


Fig.3. A 'rainbow' plot showing formant values of a male native speaker of Standard British English producing the 'my' diphthong (x-axis = -F2 (Hz), y-axis = -F1 (Hz)).

https://doi.org/10.1017/S0958344004000813 Published online by Cambridge University Press

place a symbol on a chart and say 'aim for this', simply because the acoustic characteristics of sounds produced by different individuals, even of the same sex, differ greatly. Therefore, the values that are plotted must be interpreted in a relative fashion, and a series of readings is necessary for the intelligibility of the pattern. Typically 4–5 readings of vowels at extreme ends of the chart, i.e. the beat, bat, Bart, and boot sounds, are required to provide a context for the reading of other sounds.

In the case of diphthongs, slightly different treatment is required since it is necessary to illustrate how articulation changes over time. Consequently, there are three variables to be displayed: F1, F2 and Time. The problem of how to display the alteration of values is quite easily resolved: one simply plots the individual values for each time interval instead of calculating an average. The rendering of the third variable, time, can be carried out by colour coding the time intervals, by changing hue every 0.05 seconds. One easily understandable and attractive colour scheme is that of the rainbow: magenta, red, etc – the term 'rainbow plotting' has been chosen for this method (Figure 3).

Apart from the complications noted above with pure vowels there is a further difficulty here: if we plot all the values measured then we may be dealing with up to 50 plots on the graph and it is quite clear that these must be erased before any more are plotted in order to prevent the graph from becoming completely unintelligible. Hence, direct comparison cannot be made between sounds, and the problem of how to provide parameters to allow interpretation of the data emerges. In this case the difficulty may be overcome by preparing a personalised graph for each user plotting symbols indicating the main 'cardinal points' of vowel production, against which the flux of values for diphthongs can be compared.

2.3 Models for correct pronunciation

Even though the graphs are to be interpreted in a relative way, the problem of what models we are to provide as 'correct' pronunciation emerges and is unavoidable. It must be stated very clearly what accent is being used as a model to avoid scenarios in which native speakers with accents that differ from that of the model could appear to have 'incorrect' pronunciation. For the purposes of this study the vowel repertory of Standard British English was used, the reasons for which go beyond sociolinguistic considerations. To put it simply, Standard British English is easier to analyse than many others as it does not have rhotic vowels. Rhoticism involves curling up the tip of the tongue (Laver, 1994: 270) and the resulting formant values bear more of a resemblance to diphthongs than to pure vowels (Figure 4). As a result the plotting of single dots or symbols representing averages of values involves a distinct risk of over-simplifying, if not falsifying the characteristics of the sound.

2.4 Normalisation

One important issue to be addressed is that of how to compensate for the widely differing acoustic qualities of vowel sounds produced by individuals with vocal tracts of

https://doi.org/10.1017/S0958344004000813 Published online by Cambridge University Press

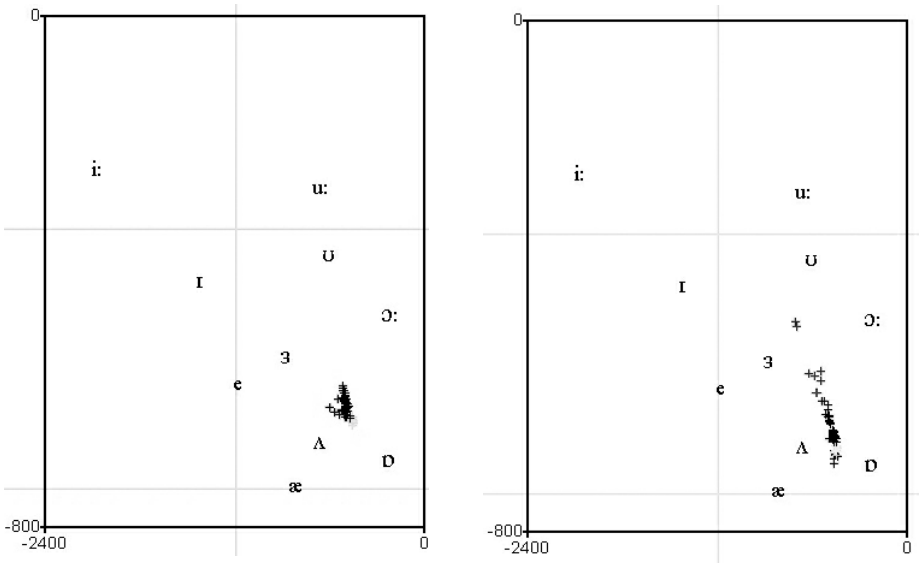


Fig. 4. Plots of formant values for the 'cart' vowel without (left) and with (right) rhoticism. In the latter case far greater flux can be seen in the values (x-axis = -F2 (Hz), y-axis = -F1 (Hz)).

different dimensions, in practical terms: children, women and men; there may even be substantial differences between individuals within these categories. The human brain is an extremely efficient processor of this type of material and the way in which it elaborates the raw acoustic data of an /æ/ produced by a child, an adult female and an adult male, for example, and succeeds in recognising it as the same phoneme has yet to be fully understood. This process, called normalisation, has raised great interest in recent years as speech recognition has started to become a realistic alternative to the keyboard. Adank (1999) evaluated a large number of different methods of vowel normalisation and concluded that the z-score was one of the most effective. A PRAAT script has been written to carry out this statistical analysis of formant data, the execution of which does not undermine the swift delivery of feedback, and the results would seem to confirm Adank's conclusions (Figure 5). However, while the use of z-scores does seem to compensate for speaker specific differences to a satisfactory degree, its application does complicate somewhat the task at hand. In short, the z-score highlights differences between values, and treats them as relative rather than absolute data. Therefore, if ten samples of the same vowel sound (e.g. /i:/) are analysed the result will be a series of plots that are spread all over the graph, as their relatively small differences are amplified. On adding more samples of a different vowel sound with formant values that differ notably (e.g. /æ/) the first ten plots will suddenly merge as the relative differences within the group become minimal in comparison to the new readings. All this means that the learner cannot start practising immediately. In order to avail of the advantages of this elaboration, a series of readings must be taken first: typically ten readings for vowels at the extreme ends of the chart, the *heed*, *had, hard* and *who'd* sounds. These values can then be exported and loaded each time the individual learner starts the exercise, so the initiation process, which lasts only a few minutes, need not be repeated.

https://doi.org/10.1017/S0958344004000813 Published online by Cambridge University Press

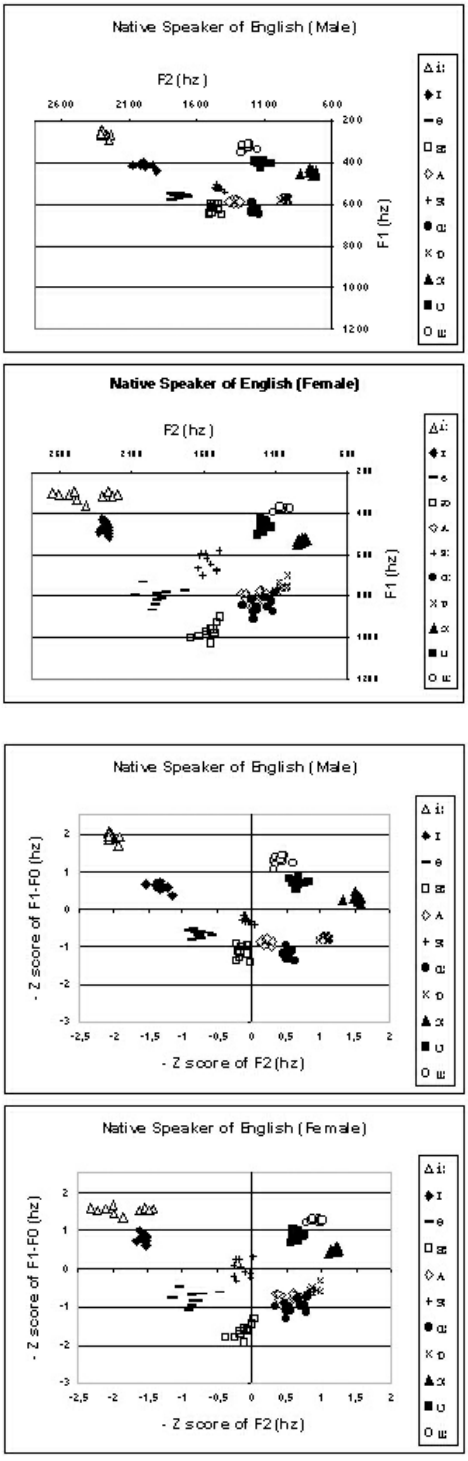


Fig. 5. Graphs showing the difference between raw formant data (top) and that which have been elaborated with the z-score method (bottom). In the latter case it can be seen that direct comparison can be made between the results for the male and female subjects. From Brett (in press).

https://doi.org/10.1017/S0958344004000813 Published online by Cambridge University Press

Practice on problematic sounds, central vowels in the case of Italian speakers of English, can then be started. Care must be taken however not to add too many new values, which would cause a distortion of the relative model prepared for each individual, i.e. after about ten readings the initial model should be reloaded.

3 Further development

In order to address the problem of rendering the feedback more user friendly, the author is currently exploring the potential of using the extremely versatile multimedia tool Macromedia Flash MX to create a graphically-improved interface. This program has already been used by the author to create applications for the practice of vowel perception and phonemic transcription skills (Figure 6).

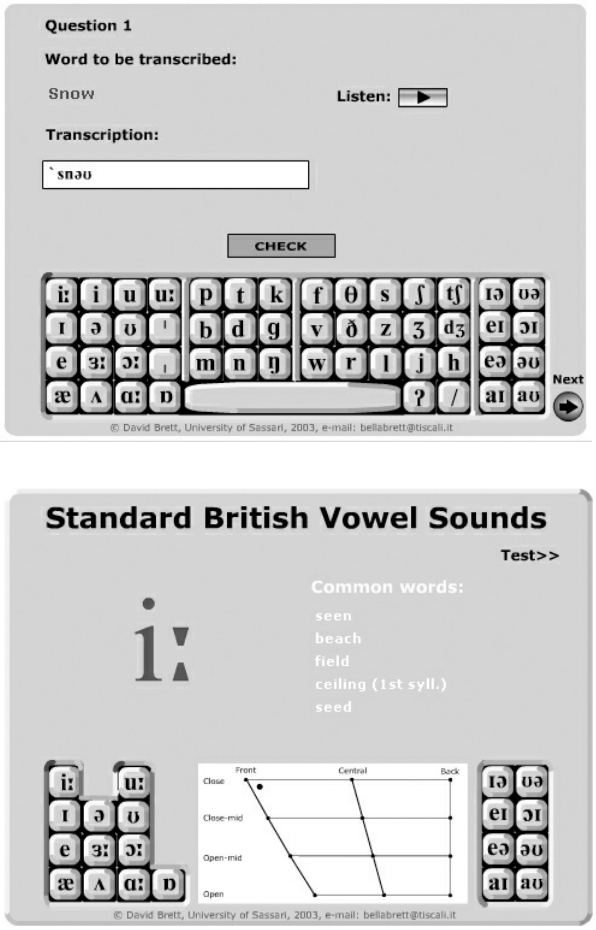


Figure 6. Two applications created by the author using Macromedia Flash MX: (top) Standard British English Vowel Sounds and (bottom) Transcription Exercises. This material can be viewed online at the author's website: http://web.tiscali.it/davidbrett.

https://doi.org/10.1017/S0958344004000813 Published online by Cambridge University Press

The essential difficulty lies in programming the Flash interface to intercommunicate efficiently with PRAAT; the prospects are, however, promising. Once the data is loaded by Flash, using ActionScript's LoadVars command, the modalities of visualisation are numerous. The following is a summary of the main points of interest:

1. Sliding bars can determine the scale at which values are to be plotted, providing ideal settings for male and female speakers with the possibility of fine adjustment. This may resolve in part the problem of reproducing normalisation.
2. Phonetic symbols can be moved using the drag'n'drop technique, therefore plots can be annotated with ease.
3. Flash can handle the calculations necessary for the 3D rendering of objects, hence the third formant may also be plotted.
4. Finally, the opacity of plots can vary from 100% to 0%, useful for the rendering of decay in the plotting of real-time calculations, i.e. the plot referring to the most recent value is the most opaque, the rest become progressively more transparent.

These possibilities make Flash an ideal program for the creation of an improved interface, moreover, the LoadVars command can be repeated continuously, permitting the interface to constantly update with the data provided by PRAAT, hence allowing the possibility of providing real-time feedback.

4 Conclusion

The applications for the PRAAT program described in this article represent a step forward in the field of CAPT, even when compared with the two specifically developed programs which have been referred to above (section 2.1), because:

1. the scripts are open source and can be modified, developed and tailored to the specific needs of the training situation;
2. PRAAT is free, whereas the other two are expensive and do not even offer a trial version. High initial costs do not encourage workers in the field to try out and evaluate new methods and technology.

Nevertheless, it cannot be denied that this project, while promising, has far to go before an individual learner can use it without qualified help to gain easily useful, clear feedback on his or her vowel production.

One of the main drawbacks concerns the nature of the PRAAT program itself: it has been developed as a research tool, and not a didactic tool. Therefore, the possibilities of representing feedback graphically are basically limited to graphs that can be read only by the user with at least basic knowledge of articulatory phonetics. The SPECO Project (Vicsi, *et al*, 1999) provides an excellent example of software which gives attractive feedback that is easily understandable for the uninitiated (in this case children); however it focuses on different aspects of speech production from those discussed in this paper.

This discussion may be rounded off by concluding that using PRAAT to provide feedback on vowel production with its current interface is not a viable option: this conclusion should not detract from the program itself, as it was not designed for this task. For

https://doi.org/10.1017/S0958344004000813 Published online by Cambridge University Press

the specific use to which it is to be put in this case, a specific interface must be created, and the program Macromedia Flash MX would seem to be the most promising program to use for its manufacture.

References

Adank, P. (1999) Acoustic vowel normalisation: dealing with different sources of variation, XIVth *International Conference of Phonetic Sciences* (ICPhS), San Francisco, 1593–1596.

Brett, D. (2003) Improved vowel production with the PRAAT programme, *Proceedings of Annual Meeting of the TESOL Pronunciation and Computer Special Interest* Group, London, June 2002.

Brett, D. (in press) The PRAAT Programme in the L2 Pronunciation Class: Formant Plotting and the Use of Z Scores, *Annali della Facoltà di Lingue e Letterature dell'Università degli Studi di Sassari*, 2002.

Canepàri, L. (1999) *Manuale di Pronuncia Italiana*. Bologna: Zanichelli

Deterding, D. (1997) The formants of monophthong vowels in standard southern British English pronunciation, *Journal of the International Phonetic Association* (**27**): 47–55.

Escudero, P. (2002) The perception of English vowel contrasts: acoustic cue reliance in the development of new contrasts. In: Leather, J. and James, A. (eds.), *Proceedings of the 4th Symposium on the Acquisition of Second Language Speech*, New Sounds 2000, University of Klagenfurt, 122–131.

International Phonetics Association (1999) *Handbook of the International Phonetics Association*. Cambridge: Cambridge University Press.

Ingram, C. L. and Park, S. G. (1997) Cross-language vowel perception and production by Japanese and Korean learners of English. *Journal of Phonetics* **25**: 343–370.

Laver, J. (1994) *Principles of Phonetics*. Cambridge: Cambridge University Press.

Molholt, G. (1998) *Accent Reduction Via Acoustic Analysis*. New Jersey: Kay Elemetrics Corp.

Roach, P. (1991) *English Phonetics and Phonology*. Cambridge: Cambridge University Press.

Vicsi, K., Roach, P., Oster, A., Kacic, Z. and Barczikay, P. (1999) SPECO: a multimedia multilingual teaching and training system for speech handicapped children. *Proceedings of the XIV International Congress of Phonetic Sciences* (ICPhS), San Francisco, 1361–1364.

https://doi.org/10.1017/S0958344004000813 Published online by Cambridge University Press