SOME EXPERIMENTS INVOLVING THE ANNOTATION OF A LARGE SPEECH AND NATURAL LANGUAGE DATABASE

P.E.Kenne, M.J.O'Kane and H.Pearcy
The University of Adelaide

ABSTRACT. We report on the effect of different interfaces on the manual adjustment of a segmentation of a speech waveform provided by an automatic segementation process.

INTRODUCTION

A major difficulty for both speech recognition systems and natural language systems is the large effort required to port such systems to a new application. Both speech and natural language (NL) systems require large amounts of training data (for example, the DARPA Air Travel (ATIS) database (Hirschman, 1992) consists of approximately 15000 utterances collected at five sites). Both data collection and annotation are labour intensive activities.

All court proceedings in Australia are recorded, and transcripts are produced for over 95% of them. The recordings together with the transcripts provide a rich source of data for speech and NL training. Until recently, the recordings of court proceedings have not been suitable as training data for speech recognition systems due to the often poor quality of the recording. Auscript (the Australian court reporting service) is installing facilities to provide recording to DAT. At present (August 1994), a small number of courts have had digital recording equipment installed.

A major difficulty in using these data to derive a speech recogniser training database is that the transcripts are not in any way time aligned with the audio data (not surprising given their source and the current application of the transcripts). A further difficulty is that due to Auscript editorial policy, the transcripts are not a faithful representation (neither at the word nor the phonetic level) of the audio. For example, repetitions such as "yes yes yes" are transcribed as "yes," and other effects such as stutters etc. are not transcribed.

Techniques exist for the automatic or semi-automatic segmentation of speech waveforms to produce a phonemic labelling of the utterance (Schmidt and Watson,1991; Fujiwara, Komori and Sugiyama 1992). These techniques depend on the availability of a recogniser. In addition, a number of commercially available segmentation tools are now available, which will, given a transcript, produce a phonemic and word level segmentation of an utterance. The accuracy of these tools depends on both the quality of the recogniser, and the accuracy of the transcript (as the transcript is used to limit search in the recognition). In the case of court transcripts, it is clear that hand adjustment will be required to handle both the case where the transcript contains an inadvertent error, as well as the case where editorial policy requires a difference between the transcript and the utterance. The recognisers used with existing segmentation tools for English (both commercial and research versions) have been trained using corpora which provide coverage of American or British English, for example, the TIMIT database. It remains an open question as to how well such recognisers will perform with the English recorded in Australian court rooms. (This is particularly of interest as some 25% of Australian residents report speaking a language other than English at home.) Some techniques have also been described for the automatic labelling of prosodic events in the speech signal (Campbell and Sagisaka, 1992).

INITIAL LABELLING

As we did not have access to the particular automatic segmentation tools referred to above, we briefly describe a method of first-approximation automatic segmentation which is backed up by a number of simple techniques to provide semi-automatic segmentation of the speech to provide a word labelling of the utterance. A more complete description of this labelling process may be found in Kenne, O'Kane and Pearcy (1994). In general, a labelling provided by one of these techniques is not sufficiently accurate over the entire set of utterances to be used for a speech training/testing database, and human intervention is required to adjust the boundaries for the labels.

The method depends on (reliably) locating silences within an utterance and on having a transcript of the utterance available. It is also sensitive to errors in the transcript. The initial labelling proceeds by determining an average phone duration for the speaker based on the transcript (no attempt is made to take into account stressed or unstressed phones) and using a simple knapsack algorithm to provide the segmentation. The average phone duration is clearly speaker-dependent, however we have

found close agreement between the average duration for stressed vowels in English (van Ooyen, Cutler and Bertinetto, 1993) and the average phone duration. In general this produces a segmentation which is initially quite poor, and is refined by multiple-pass process to adjust the average phone duration to find a best fit for the labelling. This method however, does not require a speech recogniser.

THE EXPERIMENTS

Given the need for human adjustment of labels produced by an automatic segmentation process, we conducted a series of experiments in which both linguistically naive and linguistically trained subjects were presented with the results of the automatic segmentation, and were required to adjust the label boundaries. The time taken to segment an utterance was recorded, as well as a measure of the accuracy of the segmentation relative to a reference segmentation. In addition, we also measured the time taken to segment the utterance given the raw speech data and the transcript i.e. with no automatically produced segmentation. These tasks were performed using two different interfaces (described below) to measure to effect of the interface on these tasks.

The functionality required of a tool for this experiment included (as a minimum) the ability to

a select a region of speech,
b play a selected region of speech,
c create a label,
d edit a label,
e select a label, and
f move the end points of a label.

The tools used in this experiment were the toolset provided by the Oregon Graduate Institute (the OGI tools, Farly, Pochmara and Cole, 1993), and an interface developed locally using the Tcl and the Tk toolkit (Ousterhout, 1994).

All experiments were performed on a Sun Sparcstation 2, using a three-button mouse (with the left button being M1, the middle button M2 and the right button M3). The subjects were drawn from students and staff of The University of Adelaide and all were familiar with the use of computers, keyboards and the mouse. Approximately 60% of the subjects were familiar with at least one tool for the segmentation of a speech waveform, and had had experience with segmentation. Only 33% of these subjects had used the OGI tools previously. None of the subjects had used the Tcl/Tk interface prior to these experiments.

Figure 1 below shows an example of the OGI tools with a speech wave and a series of labels. With this set of tools, the functionality is achieved as follows:

a the start of a region is marked by clicking with M1 (or M3) and the other end of a region is marked with M3 (or M1) at the appropriate points in the wavetool window;
b clicking M2 anywhere in the wavetool window;
c clicking M3 anywhere in the Lola window. This creates an empty box for the label at the point at which the mouse was clicked;
d after the label is selected, the standard editing keys (backspace etc.) may be used to change the label;
e there are three methods - either by clicking anywhere in the box containing the label with M2 or by moving with the arrow keys (right or left) from a previously selected box. Initially no box selected. Clicking with M1 inside a box will also select that label, however, it does so in a 'destructive' way in that the boundary closest to the point at which the mouse is clicked is shifted to the point of clicking.
f clicking with M1 (described in e above), or by pressing M1 and holding it down, while dragging the appropriate boundary to the desired location. The boundary which is closest to the point at which the button is clicked (or pressed) is the one which is moved.

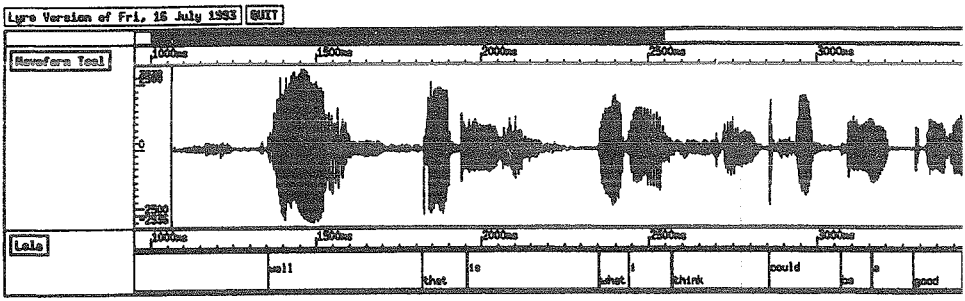

Figure 1

Figure 2 shows the interface of the Tcl/Tk tool. It should be noted that this tool was only developed for the purposes of the interface study, and does not have the complete functionality of the OGI toolset. The commands to perform the required functions in this tool are:

a same as the OGI tools;
b similar to the OGI tools - by clicking M2 either in the wave or the mark-up palette (note the contrast with the OGI tools);
c clicking with M3 in the mark-up palette;
d as the OGI tools - when a box is selected, the standard editing keys may be used to modify the label;
e click M1 in the label;
f as the OGI tools - by clicking or dragging with M1 (respectively M3). A difference is that the boundary marker remains visible while dragging (in contrast to the OGI tools).

The major differences between the two interfaces are that with the Tcl/Tk tool the user has the ability to listen to a selected region of speech without having to move the mouse between two regions in the window, and the Tcl/Tk tool also provides visual feedback when dragging the endpoint of a label.

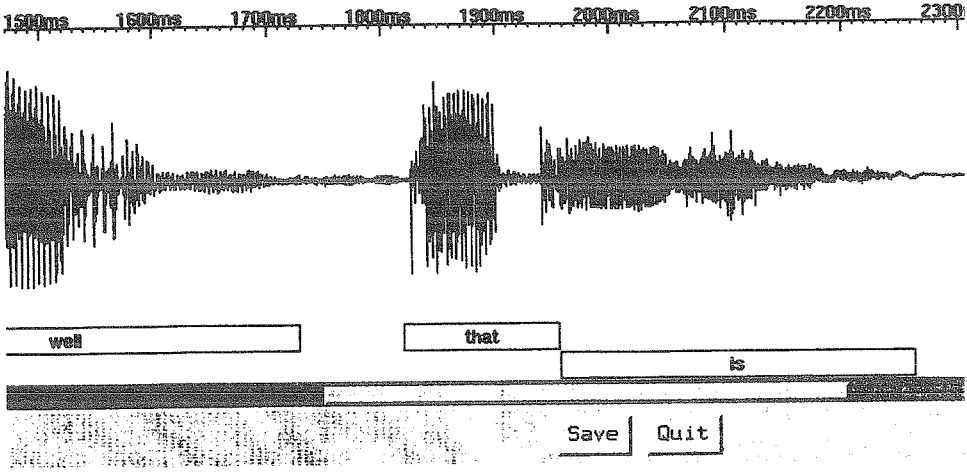

Figure 2

Subjects were able to spend as long as they liked familiarising themselves with each tool and the marking-up process. Assistance was available to them during this familiarization process. For each of the experiments, they were required to mark-up or adjust four short audio files. For each supplied segmentation, there were no errors in the labels.

In the case of adjusting the labels, the segmentation was assessed by summing the absolute values of the differences between the start and end times respectively of each word in the labelled speech, and the start and end times of the corresponding word in a reference segmentation. In the case where subjects were required to insert labels, subjects were supplied with a correct transcript, and were told that the transcript contained no errors.

Experiment 1

In this experiment subjects were supplied with a transcript and no automatically produced segmentation. They were told that the transcript was correct, and were required to manually label the utterances.

Experiment 2

In this experiment subjects were supplied with an automatically produced segmentation in which over 75% of the start and end points for all labels were over 50ms from their correct positions. Subjects were again told that all labels were correct and were required to adjust the label boundaries.

Experiment 3

As in experiment 2, subjects were supplied with an automatically produced segmentation. In this case approximately 20% of the start and end points for all labels were over 50ms from their correct positions. The remaining start and end points were within 50ms of their correct positions. Subjects were again told that all labels were correct and were required to adjust the label boundaries.

RESULTS

The tables below show the performance of the experienced (tables 1 and 2) and inexperienced (tables 3 and 4) subjects in each of the experiments for each of the interfaces. The performance is given as the time required (in hours and minutes) to adjust (or label) one minute of speech, and the error rate is the total error per minute in the start and end times of the labels relative to the reference labelling.

	Time (h:m) / minute	Error (%)
Experiment 1	1:56	8.9
Experiment 2	2:04	9.2
Experiment 3	1:35	8.8

Table 1:Performance for experienced segmenters, OGI tools

	Time (h:m) / minute	Error (%)
Experiment 1	1:48	8.8
Experiment 2	1:53	9.3
Experiment 3	1:24	8.9

Table 2:Performance for experienced segmenters, Tcl/Tk tools

	Time (h:m) / minute	Error (%)
Experiment 1	2:23	12.2
Experiment 2	2:04	12.5
Experiment 3	1:57	12.1

Table 3:Performance for inexperienced segmenters, OGI tools

	Time (h:m) / minute	Error (%)
Experiment 1	2:08	12.4
Experiment 2	2:04	12.1
Experiment 3	1:52	12.2

Table 4:Performance for inexperienced segmenters, Tcl/Tk tools

CONCLUSIONS

The error rates for both the experienced and inexperienced segmenters have been calculated relative to the reference segmentation and do not reflect an 'acceptable' segmentation.

It is interesting to note that the time required to insert labels is slightly less than the time required to adjust a badly segmented file for the experienced subjects.

The Tcl/Tk interface has decreased the time required to adjust (or insert) labels. We believe this is due to the elimination of a large number of mouse movements between windows. The mouse and keyboard events have been recorded for all subjects but have not yet been analysed.

REFERENCES

Campbell W.N. and Sagisaka, Y. (1992), *Automatic annotation of speech corpora*,SST'92, Brisbane, Australia, November 1992, 38-43.

Farly, M.A., Pochmara, J. & Cole, R.A. (1992) *An interactive environment for speech recognition research*, Proceedings International Conference on Spoken Language Processing 92, Banff, Alberta, Canada, October 1992, 1543-1546.

Fujiwara S., Komori, Y. & Sugiyama, M. (1992) *An integrated system for automatic labelling based on HMM and spectrogram reading knowledge*, ISSPA 92, Signal Processing and its Applications, Gold Coast, August 1992, 275-278.

Hirschman, L.(1992) *Multi-site data collection for a spoken language corpus*, DARPA Workshop on Speech and Natural Language Processing, February 1992, p.7-14,(Morgan Kaufmann).

Kenne, P., O'Kane M.J. & Pearcy, H. (1994) *On the derivation of a large speech and natural language database through the alignment of court transcripts,* Proceedings International Conference on Spoken Language Processing 94, Yokohama Japan, September 1994, 1819-1822.

van Ooyen, B., Cutler, A., & Bertinetto, P,M. (1993) *Click detection in Italian and English*, Proceedings Eurospeech 93, Berlin, Septermber 1993, 681-684.

Ousterhout, John K. (1994) *Tcl and the Tk Toolkit*, (Addison-Wesley:Reading).

Schmidt, M.S. & Watson, G.S. (1991) *The evaluation and optimization of automatic speech segmentation*,Proc. of Eurospeech 91, Genova, September 1991, 701-704.

AN *AB INITIO* ANALYSIS OF RELATIONSHIPS BETWEEN CEPSTRAL AND FORMANT SPACES

Simon Hawkins, Iain Macleod and Bruce Millar

Computer Sciences Laboratory
Research School of Information Sciences and Engineering
Australian National University

ABSTRACT – Building on earlier work (Hawkins & Clermont, 1990), this paper uses simple Artificial Neural Networks to learn how to map points representing a single speaker's vowels in a 12-Dimensional cepstral space into points in a 3-D formant space. We find that an ANN with only six hidden units can learn this mapping entirely on the basis of the supplied training data. We then analyse the operation of individual hidden units to try to discover the means by which the ANN is able to map input to output points so successfully. The suggestion from this analysis is that the F1 coordinate in the output space is estimated using a linear combination of input coefficients, but that the F2 and F3 coordinates are estimated by piecewise-linear mappings. It appears that selection of one or other of alternative linear mappings is selected in estimating the latter coordinates according to whether F2 is greater or less than 1350 to 1400 Hz or so, corresponding to the traditional front/back distinction.

Having gained some unbiased hints about the possible structure of the relationship between vowel representations in cepstral and formant spaces, we then proceed to evaluate these hints by means of multiple linear regression analysis. The overall results here confirm our somewhat limited analysis of the operation of the trained ANN.

INTRODUCTION

This paper addresses the nature of the relationship between a cepstral representation of a speaker's vowel system and its representation in terms of the first three formant frequencies. Broad and Clermont (1989) have suggested that each of the three formants of a vowel can be estimated from a linear combination of the low-order LPC cepstral coefficients. This implies the existence of a linear mapping between the cepstral and formant domains. If this mapping is in fact linear, then the monophthongal vowels uttered by a single speaker should form a piecewise-planar surface in cepstral space just as they have been observed to do in formant space (Broad, 1981; Broad & Wakita, 1977). The angle between the two planes may well differ in the formant and cepstral domains, but the vowel surfaces should still be piecewise-planar in form. However, Hawkins *et al.* (1994a) have found that the shape of a speaker's vowel surface is not necessarily the same in cepstral and formant spaces. This means that the mapping between these spaces must (at least to some extent) be nonlinear. In cepstral space, a speaker's vowel surface tends to be parabolic in form and symmetrical about the main axis – the front and central vowels predominantly lie on one side of this axis and the back vowels on the other. In formant space on the other hand, the shape of the vowel surface varies from speaker to speaker. Hawkins *et al.* found that some speakers had a curved vowel surface while others had a surface that was near planar in shape.

Broad and Clermont (1989) proposed that each of the first three formants could be estimated in terms of a linear combination of the low-order LPC cepstral coefficients. Broad (Broad, 1981; Broad & Wakita, 1977) had already observed that a speaker's monophthongal vowels form a piecewise-planar surface in formant space, with front and back vowels lying on different planes. F3 of a speaker's front and back vowels thus had to be estimated using different linear combinations of F1 and F2, meaning that by extension when estimating F3 from cepstral coefficients, different linear combinations would also be required for front/back vowels. Hawkins and Clermont (1990) found some evidence for a piecewise-linear relationship between the cepstral and formant domains. The current paper extends their work with the aim of clarifying the extent to which the relationship between these domains can be regarded as intrinsically nonlinear.

A NEURAL NETWORK WHICH PERFORMS CEPSTRAL-TO-FORMANT MAPPING

To investigate this question of nonlinearity, we took advantage of an ANN's ability to learn arbitrary input/output mappings purely on the basis of training data, given only that the ANN has sufficient complexity to encode the required mapping and that the volume of training data is adequate for that complexity. We presented the ANN with 12-D LPC cepstral input data, representing speech frames