submitted to New Zealand English Journal, Aug 2002.

NZSED: building and using a speech database for New Zealand English[1]

Paul Warren, School of Linguistics and Applied Language Studies, Victoria University of Wellington.

Introduction

This paper provides a brief progress report on the development of NZSED, the New Zealand Spoken English Database. NZSED is being constructed in the School of Linguistics and Applied Language Studies at Victoria University of Wellington with support from a VUW Strategic Development Fund grant. NZSED is a Wellington-based equivalent to ANDOSL (Australian National Database of Spoken Language) being developed by a consortium of universities and other agencies in Australia (Millar et al. 1990). Both databases consist of spoken materials collected over recent years, representing speakers from a number of age, gender and ethnic groups, performing a range of tasks. These tasks include word and sentence reading, and two dialogue tasks: a map task and a discussion.

NZSED and ANDOSL differ from many other spoken language corpora in that they consist of digitally stored speech files, along with a range of time-aligned label files (e.g. for phonemes, syllables, words, stress groups, intonation units) and derived track files (including formant tracks, spectrograms and pitch tracks). The label files can be used to search for and retrieve acoustic data from the speech and track files. Using the Emu package developed at Macquarie University (Cassidy and Harrington 2001), and the statistical and graphical package R (http://www.R-project.org/), researchers can perform powerful searches and then carry out detailed analyses of speech data in the database. For instance, a researcher interested in the quality of the /æ/ vowel in late 20th century NZE will be able to use the search procedures to retrieve detailed acoustic information for instances of this vowel, and if desired the search can be constrained by context (e.g. before /l/, or in accented syllables) or by speaker type. By ensuring compatibility with ANDOSL, we hope also to allow ready comparison of NZE and Australian English.

---

[1] The author acknowledges the support of Victoria University of Wellington in providing seed funding for this project, and the assistance of Amy Austin, Frank Wilson, Tanya Davies, Laura Dimock, Andy Gibson and Leanne Singh-Levett in recording and labelling NZSED data. Please address all correspondence to paul.warren@vuw.ac.nz. There is a web-site for NZSED at http://www.vuw.ac.nz/lals/nzsed/

Structure and use of NZSED

Speaker population

The initial goal for NZSED is to obtain speech data from 72 NZE speakers, distributed across three age ranges (18-30, 31-45, 46-60), two sexes (male, female) and two ethnicities (self-identification as Maori or Pakeha). Six speakers will be recorded for each of the 12 cells defined by those variables. It was decided at an early stage to limit the speaker sample to the Wellington area, in order to reduce any potential for regional differences to interfere with differences resulting from the three variables listed above. It is hoped that subsequent recordings in other centers will augment NZSED for other regions.

Materials

The materials for the database consist of read sentences, word lists and dialogues.

The sentences are essentially the same as in ANDOSL, and can be described as "phonetically rich" in that they have been carefully constructed so as to contain a very wide range of permissible sound sequences in English, in a variety of prosodic contexts. In addition to 200 such sentences, there is a small set of "reference sentences" which target certain aspects of pronunciation typically associated with variation in NZE, such as t-voicing, l-vocalisation, the EAR/AIR merger, etc.

The word lists consist principally of *hVd* tokens, such as *head, hid, hod*, etc. Other words such as *here* and *hair* are included to complete the set of vowels. This word list, using largely identical phonetic contexts (i.e. between /h/ and /d/) provides an indication of the qualities of the vowel sounds for the groups of speakers recorded. A further word list includes the digits from zero to nine.

The dialogue tasks comprise a map task and a conversation. The former is based on the HCRC Map Task (Anderson et al. 1991), and involves one participant giving the other directions in order to plot a route through a map. The participants have slightly different versions of the map, and have to engage in substantial negotiation in order to complete the task. Two maps are used, so that each participant can take the role both of leader and follower. The map task provides useful data for the analysis of discourse structure, including the use of intonation in asking questions and providing feedback.

For the conversation task the participants are given a topic (their opinions on drink-driving "advertisements" on television) that stimulates a good amount of free discussion.

Labelling

In order for the database to fulfil its potential as a powerful tool for investigating details of NZE pronunciation, it has to be appropriately labelled. This is perhaps the most time-consuming aspect of the whole process of establishing the database. To date, word and phoneme (speech sound) labels have been entered by trained researchers for a subset of the word list and sentence recordings. The positioning of labels is based on a mixture of listening to portions of the speech files and scrutinising the speech wave and spectrograms derived from the speech wave. Since speakers

typically run speech sounds into one another, it has been necessary to establish a set of reliable criteria for determining the boundaries between them. Subsets of speech files have been double-labelled by different researchers, to allow checks to be made on the use of the criteria.

The speed of the labelling process is now being increased by first generating a set of phoneme recognition models (using Hidden Markov Modelling techniques) based on the initial set of utterances that have been labelled by the researchers. These phoneme recognition models are then used in a computer program which produces a provisional set of labels for previously unlabelled utterances. Since the computer phoneme recognition models are not totally robust, human labellers subsequently correct the label files for error and the revised label files are saved as part of the database.

The complete set of label files in NZSED contains information about the presence and location of speech sounds, words and other relevant aspects of the speech files. Each label file is associated with the source speech file, and can be used to locate instances of the labelled speech phenomena. Figure 1 gives an illustrative example of a portion of a speech file, with the waveform at the top, the phoneme and word labels below that, and a spectrogram of the speech at the bottom.

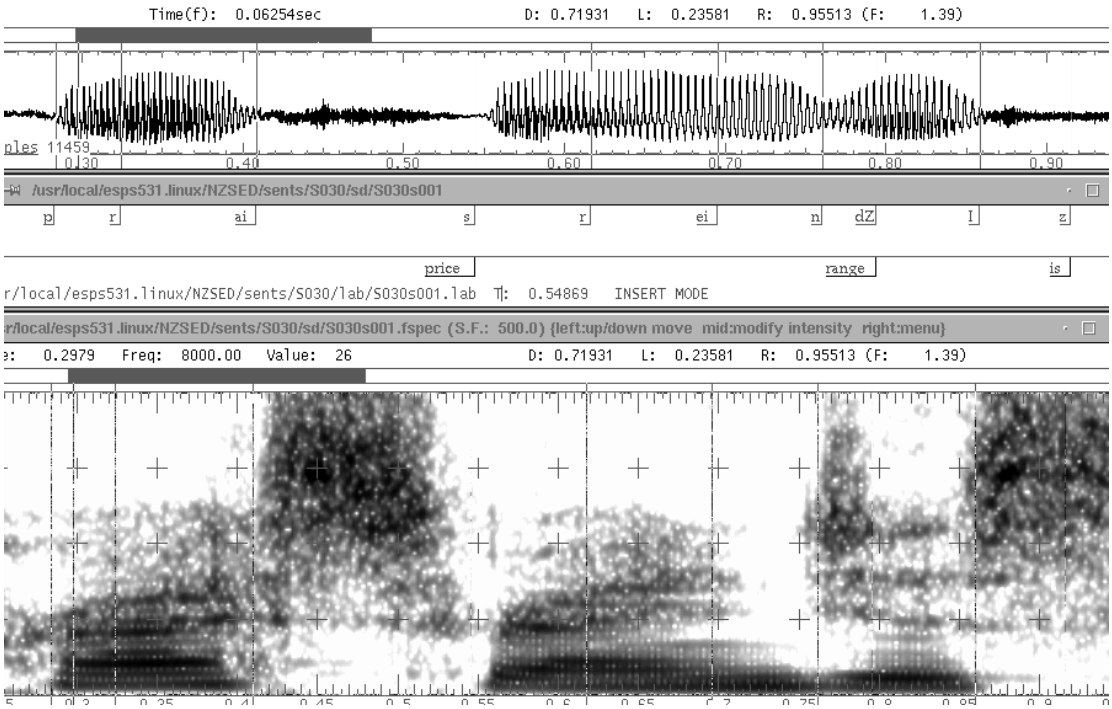

Figure 1: waveform, labels and spectrogram for the sentence fragment “price range is”.

Recent investigations using NZSED

Two examples of recent use of the database are outlined here. It should be stressed that the use of the database is constrained by the extent and accuracy of the label files, and that detailed research may require some checking that the use of the label files is in fact retrieving the relevant data and information.

The first example is from a recent Linguistics Honours project, which required students to compare the realisation of /e/ and /æ/ (i.e. the vowels in *bed* and *bad*) before /l/, as in the names *Ellen* and *Alan*, or the words *celery* and *salary*. One feature of NZE pronunciation that is discussed in the research literature is the neutralisation of the distinction between these two vowels in this particular context. Students were asked to evaluate the extent of this neutralisation using data from NZSED. Of the two students completing this assignment task, one chose to compare young Pakeha males and young Pakeha females, while the other compared the latter group with mid-aged Pakeha females. Using the label files generated for the sentence data, our students were able to retrieve speech segments containing each of these vowels before /l/, from which they then obtained acoustic data. The data for each vowel were compared with one another and with values for a range of vowels taken from the *hVd* contexts for the same speakers. These comparisons showed that before /l/ the /e/ and /æ/ vowels have largely overlapping distributions for each of the speaker groups, i.e. the vowels are essentially neutralised to a centralised open-mid vowel. However, there is some variation in the pronunciation of the neutralised form, with the younger speakers having a more open articulation than the mid-age speakers. Figure 2 shows ellipse plots derived using Emu and R, for vowel formants for /e/ and /æ/ before /l/ for the young female Pakeha speakers in the database (with centre points labelled as E–>l and A–>l, and with dotted lines for /e/, solid lines for /æ/), together with distributions for the /e/ and /æ/ vowels in all contexts (centre points labelled as E and A) for the same speakers.[2] Each ellipse shows the distribution of these vowels to within one standard deviation of the mean formant values. It is quite clear from the figure that the two vowels have near-identical distributions before /l/, and that these distributions are considerably centralised and lowered, compared to the overall distribution of /e/ and /æ/.

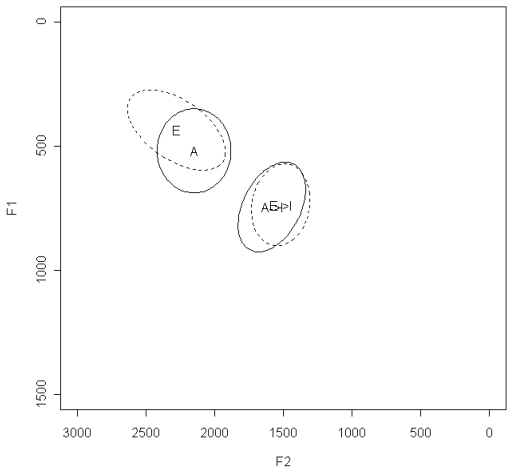

Figure 2: distribution of /e/ and /æ/ vowels in young Pakeha female data – see text for details

[2] Vowel formant plots indicate the position of the tongue in the mouth when the vowel is spoken. The first formant (or resonant frequency, labelled F1 in the figure) relates to the relative height of the tongue, and the second formant (F2) to its relative frontness. For this reason formant plots are usually displayed with the axes as in Figure 2, and the reader should imagine that the area in the figure relates to the inside of the mouth of a left-gazing head, so that high points on the plot are high positions within the mouth, and points to the left correspond to positions towards the front of the mouth.

A second example relates to the merger of EAR and AIR diphthongs in NZE. In this case the NZSED word list and sentence data provided background information for a study by Rae and Warren (in prep), confirming observations from other studies, namely that these diphthongs have overlapping starting points for young speakers of NZE. Figure 3 shows how the vowel formants change during the vowels in *here* and *hair* recorded for the word list task.[3] The data for the Figure come from two sets of female Pakeha speakers in NZSED, with the average of the six speakers from the "young" group on the left and that of the six "old" speakers on the right. The plots were generated using Emu and R. In each panel the dotted line shows the formant trajectories for the vowel in *here*, and the solid line shows the formant trajectories for the vowel in *hair*. The two panels show how the difference between the early portion of the two vowels is smaller for the younger speakers than for the older speakers.

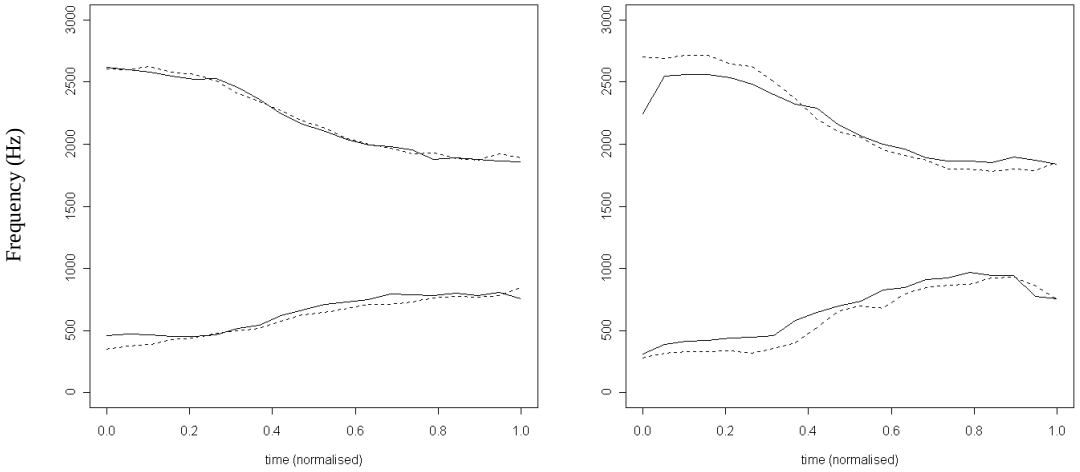

Figure 3: Average formant trajectories for the vowels in *here* (dotted line) and *hair* (solid line) from word-list recordings for young (left) and old (right) female Pakeha speakers in NZSED. See text and footnote 3 for more details.

Concluding remarks

NZSED has yet to realise its full potential. Largely this is because the labelling stage is a time-consuming part of the process of setting up the database, and at the time of writing this labelling is still some way from completion. The accuracy of the labelling process is clearly also crucial to the value of the materials collected for NZSED. However, the database has a strong potential in both teaching and research on NZE. In addition, the availability of digitally stored speech data and accompanying label files should be of interest to NZ-based companies designing products such as text-to-

[3] The formant trajectories in Figure 3 reflect the movement of the tongue during the articulation of the diphthongs. In each panel, the lower pairs of lines represent the first formant (F1) for EAR and AIR, with F1 indicating vowel height. The upper pairs of lines show the second formant (F2), relating to the relative frontness of the vowel. In each case F1 rises and F2 falls during the diphthongs, reflecting (respectively) the lowering and backing of the vowel from a fairly close front articulation towards the central schwa vowel position. The unevennesses in the plots reflect the fact that each line is an average of only six tokens of the relevant vowel.

speech systems, since one of the common complaints about computer-generated speech is that it frequently uses an accent that is not native to the end-users.

References

Anderson, Anne H., Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, Henry S. Thompson and Regina Weinert 1991. The HCRC Map Task Corpus. Language and Speech 34(4): 351-366.

Cassidy, Steve and Jonathan Harrington 2001. Multi-level annotation in the Emu speech database management system. Speech Communication 33: 61-77.

Millar, J. Bruce, Philip Dermody, Jonathan Harrington and Julie Vonwiller 1990. A national database of spoken language: concept, design, and implementation. International Conference on Spoken Language Processing (ICSLP-90), Kobe, Japan.

Rae, Megan and Paul Warren in prep. Asymmetry in the merger of EAR and AIR: psycholinguistic evidence. Wellington Working Papers in Linguistics.