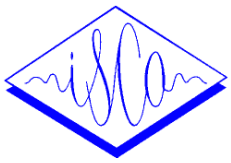


2nd European Conference on
Speech Communication and Technology
EUROSPEECH '91
Genova, Italy, September 24-26, 1991

ACCENT SPECIFIC MODIFICATIONS FOR CONTINUOUS SPEECH RECOGNITION BASED ON A SUB-WORD LATTICE APPROACH

K. Edwards, F. R. McInnes and M. A. Jack

Centre for Speech Technology Research, Edinburgh

Abstract

Speech research has to date been dominated by either General American or RP British English accents. This limits the applicability of continuous speech recognition systems based on sub-word modelling to a small percentage of the English speaking population and makes true speaker independence more elusive.

The modular speech recognition system discussed here allows experimentation at different stages in the recognition process, and is well suited to the implementation of different accents and languages. The 'front end' processing which uses Hidden Markov Models allows information about the phonetic and phonological details of an accent or language to be incorporated with the minimum of adjustment to the whole system. The 'back end' uses a lexicon which contains word transcriptions in terms of the units modelled by the HMMs.

This paper describes the adaptation involved in enrolling speakers of a Scottish accent into the system, which was previously based on the assumption of RP accent users.

A Continuous Speech Recognition System

For the continuous speech recognition system used in this work [4], initialisation of the HMM training process requires speech data, appropriate transcriptions, a set of 'seed' HMMs and a vector quantisation codebook. The HMMs reflect a given phonemic system, and include various sub-phonemic and allophonic details e.g. clear /l/ vs. dark /l/. Silence, consonant clusters and other phoneme sequences are also modelled, giving 98 APUs (Acoustic Phonetic Units) for the RP accent of English. The sentence transcriptions used in training and evaluation are in terms of these units and are produced automatically from a phonemic lexicon and various rule-based conversions.

Speaker specific refinement of the seed HMMs depends on the iteration of two stages, segmentation and training.

Segmentation

This stage includes the signal processing (for the first iteration) where cepstral analysis and formant estimation are applied to the speech data to give acoustic feature vector files. The vectors are then transformed into a discriminant space and quantised using the codebook, to give VQ label files. These are used with the sentence transcriptions and the HMMs to segment the speech into time-aligned sequences of APUs. Following this, sub-segmentations based on the original segmentations are derived.

Training

Training involves linear discriminant analysis of the acoustic feature space and sub-segmentations, to update the transform coefficients file, and class-based k-means analysis to update the VQ codebook. New VQ label files are then produced which are used, along with the segmentations, to re-train the HMMs.

There are several iterations of these stages. The 'seed' HMMs are used to give initial segmentations of the data. From these, the HMMs are re-trained into the first speaker-specific HMMs. These are then used to re-segment the data, producing a more accurate segmentation than the 'seed' HMMs. Another training stage produces further refined models. This is augmented by the separate training of sonorant and non-sonorant HMMs.

Finally, the trained models and VQ codebook can be used to 'decode' some unknown utterances. Once the signal processing and VQ labelling is applied, the input data is segmented and probabilistically classified using the universal and broad-class VQ labels and HMMs. The HMM segmentation is also guided by APU sequence probability scores.

'Segments' are hypothesised from sequences of HMMs which best match the input data, and the log probability for all APUs within each segment is calculated, to give the complete APU lattice. The lattice, then, contains many scored, complete segmentations of the input.

The performance of the front end can be evaluated using an entropy measure [3,4], which is based on correct sentence transcriptions of the test data. These transcriptions are generated in the same way as the training sentence transcriptions. Various results can be obtained, including APU alignment statistics, insertion and deletion probabilities, utterance entropy and phoneme entropy. For a more detailed description of the system, see [4].

Adaptation to Scottish Accent

From the above outline, the two key areas of the speech recognition system which are accent specific are the set of APUs modelled by the HMMs, and the sentence transcriptions which are matched to the training speech.

If the only differences between accents of English were phonetic, realisational differences, then the iterative training of HMMs reflecting one phonemic system would successfully model them, and regional inter-speaker differences would cease to be a problem for HMM based recognition. However, the systemic and distributional differences which do exist across accents, restrict the applicability of a single set of models and transcriptions.

To cover Scottish accents in a way which reflected their phonemic systems, a Scottish APU set was devised. Although new symbols were introduced to represent vowels not present in RP, similarities with RP were preserved where possible, with the same symbol being used for distributionally different, but phonetically similar vowels. (e.g. /ɑ a ɔ ɒ/)

Some of the most important differences between Scottish and RP are exemplified in table 1.

Example	RP	Scottish
bright	/braɪt/	/brʌɪt/
fire	/faɪə/	/faeər/
bored	/bɔd/	/bɔrd/
report	/rəpɔt/	/rəpɔrt/, /rəpɒrt/
foot	/fʊt/	/fut/
goose	/gus/	/gus/
cure	/kjʊə/	/kjur/
deer	/dɪə/	/dir/
where	/wɛə/	/wer/, /ʍer/
hand	/hand/	/hand/
path	/pɑθ/	/paθ/

Table 1. Examples of RP and Scottish accent.

These changes also have implications for the 'back end' processing in the recognition system, and for the sentence transcriptions: Scottish transcriptions and lexica had to be generated.

Rather than write Scottish transcriptions by hand, the existing lexica (for the training material and recognition task) were enriched to a level which contained phonemic details of both accents. Rules were then written to allow this 'rich' level to be converted into the phonemic representation of the lexica in each accent. The phonemic details of accents are widely agreed although the Scottish rules had to generate enough optionality in the transcriptions to represent several common Scottish systems [1,2]. Extracts from the lexica are given in table 2.

'rich' form	RP form	Scottish form
'dwarf'		
d w o-o r-r f	d w oo f	d w o r f
		d w oo r f
'bear'		
b e@ rr	b e@ r	b ee r
	b e@	
'argue'		
aa r-r g y uu	aa g y uu	a r g y uu
		aa r g y uu
'earth'		
e- r-r th	@@ th	@ r th
		e r th
'entice'		
i n t ui- s	i n t ai s	i n t ui s
'entire'		
i n t ai @ rr	i n t ai @ r	i n t ae @ r
	i n t ai @	
'full'		
f u l	f u l	f uu l
'fluke'		
f l uu k	f l uu k	f l uu k

Table 2. RP and Scottish lexical entries.

After Scottish and RP sentence forms were produced, they were filtered through rule programs which added options to:

- account for likely word-boundary co-articulatory effects of assimilation and elision.
- introduce the extra units from the extended set. (98 units compared to the 'standard' 44 phonemes of RP)

Training the system for a Scottish Speaker

As well as being accent specific, the system is also speaker specific although it is initialised on universal 'seed' HMMs. The successive iterations of training models and segmenting speech data produce speaker specific HMMs and a VQ codebook which are then used in the recognition task. The seed HMMs have been trained on data which was hand-segmented by professional phoneticians, so they are a good choice as start-up HMMs for the training of other speakers within the same accent group.

Since no hand-segmented data existed for any Scottish speaker, the RP seed models had to be used. They were copied and relabelled to reflect the Scottish APU set as follows:

- The /iə ɛə ʊə/ models were discarded.
- The /ei/ model was renamed the /e/ model
- The /oʊ/ model was renamed the /o/ model
- The /aɪ/ model was copied to two models /ae/ and /ʌi/
- The /f/ model was copied to a /ʍ/ model
- Otherwise the RP models were used without alteration

The foundation for training and recognising Scottish speakers is clearly less well defined than for RP: the models must match not only a different speaker but also different phonemes.

If the initial segmentations with seed models are good enough, then repeated trainings and segmentations should produce usable Scottish accent HMMs for the recognition task, and for seed models for other Scottish speakers.

Experimental Results

Table 3 shows front-end entropy results for several RP speakers enrolled into the continuous speech recognition system, and those for two Scottish speakers. The entropy results for Scottish talkers lie with the overall range of RP entropy values. This demonstrates that other inter-speaker factors are more significant to the result than regional accent factors.

Speaker	Accent	Entropy
gsw0	RP	2.467
pms0	RP	2.671
hxb0	RP	2.621
jmr0	RP	3.022
frm0	SCOT	3.006
exd0	SCOT	2.942

Table 3. Front-end entropy results for 6 speakers.

For a more rigorous test of the accent specific modifications made to the system, the same training speech data that were used in generating Scottish HMMs for speaker frm0 were used to generate RP HMMs. Once the models were trained, the same recognition task was performed with speaker frm0 using an RP lexicon. An equivalent entropy result from this experiment was 2.901, suggesting that for an identical task within the same system, the RP HMMs produced a significantly lower entropy than the Scottish set. Entropy values for individual utterances were also in favour of the RP HMMs. Front-end lattices for part of a typical utterance in each experiment are shown in figures 1 and 2. The lattices contain segments made up of APUs plotted against scaled negative log probabilities. The lattices are well-formed in the sense that all of the speech is segmented, and if the correct transcription can be traced from low scoring APUs, then the segmentation has been accurate.

Closer examination of the test transcriptions and lattices clarified where errors could have occurred, and suggested possible improvements.

The first possibility was that the automatically generated sentence transcriptions were prone to error. If they contained sequences of phonemes which differed greatly from the realisations in the training data, then the HMMs would have produced a poor initial segmentation. The models would then be re-trained on this segmentation and would thereby incorporate some of the inconsistencies. The Scottish transcriptions contained, by design, more optionality than the RP ones, but they did not contain any unlikely phoneme sequences, so this cause was ruled out.

It was also possible that the Scottish HMMs were generally undertrained compared to the RP set. The Scottish set did not contain the APUs which were undertrained in the RP case (i.e. / iə ɛə ʊə /) but some Scottish models were more poorly trained. All the occurrences of the APU for /ai/ in RP were split into two categories for the Scottish vowels /ʌi/ and /ae/, so clearly those Scottish vowels were individually less well trained.

The most likely explanation for the higher entropy in the Scottish accent experiment is that the RP 'seed' models when restructured into Scottish seed models did not produce a good initial segmentation due to having to model not only a different talker but also different phonetic realisations. Several results point to this general problem.

Entropy values for each APU were calculated and were generally higher for the Scottish APUs, especially vowels. The vowel results are summarised in table 4.

APU	Scottish	RP	APU	Scottish	RP
a	1.408	2.024	ɒ	2.085	1.587
aa	-	1.731	ɔ	-	2.147
ae	2.270	-	ɔɪ	5.961	2.052
aɪ	-	1.436	o	1.442	-
aʊ	4.068	2.727	oʊ	-	1.516
ɛ	2.189	1.832	ʊ	-	6.125
e	1.325	-	u	3.335	3.507
eɪ	-	1.617	ʌ	-	6.125
ɪ	2.860	2.743	ʌi	2.783	-
i	1.472	1.393	ə	2.297	1.990

Table 4. Entropy values per APU (vowels).

The Scottish APUs /e o u/ have lower entropy values than their RP counterparts but they also occur more often in the Scottish transcriptions and so were slightly better trained. The Scottish vowels /i/ /e/ and /u/ each cover more vowels in RP: /i/ and /iə/; /eɪ/ and /ɛə/; /u/, /ʊ/ and /ʊə/. The converse effect is shown in the entropies for /ae/ and /ʌi/ which are worse than for RP /aɪ/. This supports the previous suggestion that splitting the same training data over two models instead of one degrades recognition performance.

Another indication of how accurately the HMMs matched the speech can be derived from the APU 'confusions'. A confusion takes place when the APU which best matches a segment of speech is not one suggested by the transcription. For consonants and vowels, the best-matching APU reflected the transcription more often in the RP experiment, although the APUs of the monophthongs /e/ /o/ and /u/ were more often correct in the Scottish case. This suggests that the RP HMMs were better trained.

The totals for insertions and deletions of APUs are less for the RP case than the Scottish. (79 vs. 115 insertions and 310 vs. 320 deletions) One explanation for this lies in the possibility of inaccurate sentence transcriptions. Alternatively, the deletion counts may indicate poorer initial segmentations from the Scottish seed models, building inaccuracy into the subsequent training of the Scottish HMMs.

It is also possible that the phonemic systems in the Scot-

tish transcriptions do not correctly represent that used by the speaker who provided the training and test data. Given that the Scottish systems were based on independently recorded observations [1,2] and that the transcriptions contained options, this possibility can be largely discounted. If Scottish seed models based on hand segmentations were available, then the matter could be settled because both RP and Scottish experiments could start from exactly the same basis.

One final factor which contributes to the result is that the Scottish speaker used in the experiment spoke with a typical Edinburgh accent of Standard English. Although systemically it is a Scottish accent, in terms of phonetic realisations of common vowels, it is in fact perceptibly close to RP. This may explain why the enrolment of speaker frm0 within the RP framework was so successful. If the experiments were undertaken for a speaker with a 'broader' accent, the enrolment for the RP accent might be worse than that for Scottish. However, the underlying initialisation problem would still affect the results.

Conclusions

Results from a modular sub-word based continuous speech recognition system have demonstrated the effective enrolment of Scottish accent speakers as well as the original RP accent speakers. An enriched form of the phonemic lexica allows transcriptions in both accents to be produced, and this could be expanded for other accents with appropriate rule sets.

Performance results for two Scottish speakers fall within the range of results for RP speakers, demonstrating the effective enrolment of speakers of different accents with one set of 'seed' HMMs. However, a comparative experiment, recognising the same Scottish speech data using both accent systems, suggests that the HMMs used to initialise the iterative training and segmenting process are, in fact, less successful when modelling different accents.

ACKNOWLEDGEMENT

This work was supported by the Information Engineering Directorate/Science and Engineering Research Council as part of the IED/SERC Large Scale Integrated Speech Technology Demonstrator Project (SERC grants D/29604, D/29611, D/29628, F/10309, F/10316) in collaboration with Marconi Speech and Information Systems and Loughborough University of Technology.

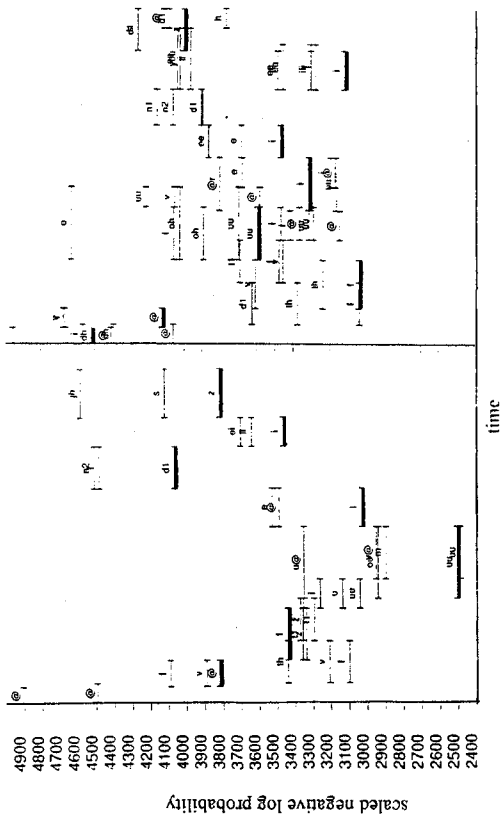
References

1 Abercrombie, D. 1977. The accents of Standard English in Scotland. *Edinburgh University Dept. Linguistics Work In Progress 10*, 21-32.

2 Wells, J. C. 1982. *Accents of English 2: The British Isles.* Cambridge University Press.

3 McInnes, F. R., Y. Ariki and A. A. Wrench. 1989. Enhancement and optimisation of a speech recognition front end based on hidden Markov models.*Proc. Eurospeech 89, 2.* 461-464.

4 McInnes, F. R., D. McKelvie and S. M. Hiller. 1990. The structure, strategy and performance of a modular continuous speech recognition system. *Proc. Inst. Acoust. 12.10* 173-181.



Figures 1 and 2. Phoneme lattices for 'The fluid is' by subject frm0 enrolled as an RP and Scottish speaker respectively.