Proceedings of the Institute of Acoustics

APPROACHING SPEAKER-DEPENDENT PERFORMANCE WITH SPEAKER-INDEPENDENT RECOGNITION

John N. Holmes

Speech Technology Consultant, 19 Maylands Drive, Uxbridge, UB8 1BH

ABSTRACT

Most "speaker-independent" recognizers do not exploit the fact that the whole of any particular man-machine conversation will normally only involve a single human being. This limitation inherently reduces the discriminative ability of such recognizers. This paper describes experiments which overcome this problem for a low-cost small-vocabulary connected-word recognizer. Many alternative sets of word models are provided, representative of a very wide range of voice qualities and regional accents of English. The goodness of match between any incoming word and all models of that word gives an indication of which is the best set of word models to use for that speaker. Within the available computational power, the recognition can start using a few sets of fairly general models in parallel, and switch to more specific models as the recognition results for each new word are obtained. The recognizer can thus select the optimum set of models for any speaker within a few words of the start of a conversation. This process is hidden from the user, who sees the recognizer as being fully speaker-independent. There need be no loss of performance with the most extreme of regional accent variations, provided models for all the regional accents are provided and the speech is clear and consistent.

1. INTRODUCTION

Over the last few years "speaker-independent" recognizers have been fairly successful, largely as a result of development of elaborate statistical models of the input speech, which are usually some form of hidden Markov model (HMM), and training on large amounts of data from a wide range of representative speakers. Some of the more recent work has achieved an extremely high accuracy [1,2,3] with recognition errors of less than 0.5% on a large data base of over 28,000 American English digits spoken in strings of various lengths from a total of 113 different adult speakers of both sexes. However, the sophistication of the statistical models used in these studies obviously required a substantial amount of computation, for both training and recognition.

In spite of the success of these speaker-independent recognizers, it is unavoidable that they should lose some discriminatory ability compared with a speaker-trained system of the same complexity. Speakers of any given language do in fact differ, both in their inherent voice quality (dependent on physiological differences) and in their learned speech habits. In the latter category, regional accent can make a very large acoustic difference to nominally identical words. In the United Kingdom in particular there are substantial phonemic differences between different regional versions of the same word. For example, most types of Scottish English have a significantly different vowel system from those which normally occur anywhere in England, and Northern Ireland is different from both of them.

These variations actually cause problems for human listeners also, and people often have some difficulty in understanding the same language from an unfamiliar geographical region. However, humans usually succeed in communicating in spite of these variations because of three important favourable factors:

1. There is usually enough linguistic redundancy in any message to severely limit the number of possible interpretations of any utterance.

2. Most people are reasonably familiar with other regional accents, partly because of present-day population mobility, and because of the widespread use of radio and television.

3. Very early in any conversation, listeners recognize the accent and therefore, completely subconsciously, they interpret subsequent words in relation to their knowledge of that accent. In this process they are relying on the fact that any particular speaker is not, under normal circumstances, suddenly going to change to a different accent or voice quality.

The speaker-independent approaches referred to earlier can certainly exploit the first of the three factors given above, either because of the very limited number of words being recognized in simple machines, or because of the language model which is always necessary for large vocabulary recognizers.

The second factor can be used to some extent by training on a suitable range of different voices, but not very effectively because the regional accents are not identified and kept separate in the training.

The third factor is completely ignored in the usual type of speaker-independent recognizer. Such machines would work equally well if, in any utterance, each word was spoken by a different speaker, and in the case of normal HMMs, they do not even properly exploit the fact that the whole of any one word will be spoken by the same person, with the same accent.

This paper describes a fairly simple technique that is able to exploit all the three factors that assist humans in coping with different speakers, and is thus able to give a fairly high accuracy of speech recognition for widely varying accents with very modest computational requirements.

2. GENERAL DESCRIPTION OF THE SPEAKER-INDEPENDENT TECHNIQUE

Although the differences between speakers have been emphasized in the introduction, it is very common for the speech of different people of the same sex and from the same geographical region to be acoustically quite similar. Listeners familiar with their speech will not usually have any difficulty in distinguishing between the voices of such people, but the essential features which characterize what is being said are not usually very different, so that a fairly simple recognizer trained on speech from a number of such similar speakers can provide "speaker-independent" performance within the restricted group that is not a lot worse than would be obtained from a speaker-trained machine. This performance can be achieved because the range of variation that occurs for any word or sub-word unit for any one speaker in the group is usually almost as great as between members of the group.

One way, therefore, of providing effective coverage of a large and diverse population is to collect a sufficient quantity of training data to cover all the distinct types of speech adequately, and to group the speakers according to the acoustic similarity of linguistically-equivalent events. Once this grouping has been done, models of the recognition units, whether words or sub-word units, can be made for each group. The only problem then, for achieving true speaker-independent recognition, is to ensure that the most appropriate set of models is in use at any time, in a way that is completely hidden from the user of the machine.

Once an appropriate set of models has been selected, it should not normally need changing within any one man-machine dialogue, because the human being involved will usually remain the same throughout. The problem, therefore, only occurs at the start of the dialogue, because nothing will be known about the speaker before the first few words have been spoken. However, in most man-machine dialogues, it is usual for the range of linguistically acceptable words at the beginning to be very small, and, if necessary, the dialogue can be artificially constrained to ensure that this is so. Thus discrimination between alternative words will be much easier than when the number of

choices is much larger, and accuracy will be fairly high even for word models derived from a wide range of different types of speaker.

If the range of word choices is small, the computation required in the pattern matching will be much less than when a larger vocabulary is involved, so it is possible to use spare computational capacity to do the pattern matching with models for many alternative types of speaker simultaneously. These models can be chosen to represent the total range of speaker types required as well as possible, given the total number of models that can be processed with the computing power available. As pattern matching algorithms normally give not only the identity of the best matching pattern, but also an indication of how well that best pattern matches to the input, if there is any disagreement about the identity of any word for the different models, recognition can be made more reliable by choosing the word or word sequence with the best-matching score. A by-product of making such a choice is knowledge of which group of speakers provided the best models, and this knowledge can be used to refine the recognition process for future utterances.

It would obviously be unwise to rely too heavily on the results for just one or two words to decide on precisely which group the current speaker belongs to, but if, for example, the first word matched quite well to a model derived from a group of Scottish female speakers, and fitted very badly to all available words from southern English men, it ought to be quite safe to remove the latter group from further consideration for the following words. The computational power so freed could then be used for a finer grouping of speaker types in the models, so that subsequent words could be used to identify the speaker type more closely. Provided a fair number of sets of models can be processed in parallel (say four or more), the latter process can very rapidly converge on the most appropriate set of well-matched speech models within a few words of the start of a dialogue, so that there should not at that stage be any great advantage for accuracy in processing more than a single set. It is then appropriate to use the spare computation to enlarge the vocabulary, if so required.

3. DESCRIPTION OF A RECOGNIZER USING THE SPEAKER-INDEPENDENT METHOD

The recognizer described in this paper uses an extension of principles described previously[4]. The current version is still a continuous connected-word recognizer intended for fairly small vocabularies, using the normal one-pass connected word recognition method[5]. It is already running in real time in C on an IBM-compatible PC, and it is currently being recoded for the whole recognizer to run on a single TMS320C25 DSP chip. This new implementation gives such an increase in computation over the 6502 version described earlier, that it is now possible to provide a larger feature vector, to recognize larger vocabularies, and also to process multiple sets of word models simultaneously for the hidden speaker-type selection. However, the emphasis has still been on developing an algorithm that is suitable for very modest hardware, while achieving as high a performance as possible within that constraint. Because the main application is expected to be on input from the telephone network, the acoustic analysis, although based on a sampling rate of 8 kHz, uses no information outside the 200 - 3200 Hz band. The recognizer still uses an HMM-like structure, with model topology selected to suit the expected phonetic content of each word[6]. The conventional HMM transition probabilities are not used, but explicit duration modelling is provided at the state and word levels.

Only the means of the single-Gaussian continuous probability density functions (PDFs) are trained for each speaker. The variances are set separately for each feature of each PDF, but the variances and the duration modelling parameters are kept the same for all speakers. The training is performed by initializing the word models by hand, and then iteratively using the recognition algorithm to label the input data in terms of model states, taking the statistics of such labelled input data to re-estimate the PDF means. It has been found that, provided the initialization of the models is fairly good, there is no obvious advantage in using the forward-backward algorithm for training, and the simple method adopted is extremely fast once the initialization has been done. For new speakers with voices that are generally similar to any for which models have already been prepared, it is normally

acceptable to use the best set of existing models as the starting point for training models for each new speaker. Because the models are designed and trained so that linguistically equivalent events from different speakers are always assigned to corresponding model states, it is possible to get models representing groups of speakers merely by averaging the model mean values of each feature of the PDF associated with each state of each word.

The system economizes on computation by using an unusually long frame period of nominally 32 ms, with an algorithm for varying the frame boundaries slightly so that they tend to occur at points of greatest acoustic/phonetic change. Use of this frame boundary algorithm avoids the blurring of sudden phonetic changes in the speech signal that would otherwise be caused by the long frame period. Within each frame, excitation-synchronous Fourier analysis is performed to produce a 32-point power spectrum. The power spectrum is further processed to derive five spectral amplitude features, three formant frequencies and the time-differenced values of these eight features. It is well-known that formant frequencies cannot always be reliably estimated from a single spectral cross-section, although for a high proportion of voiced frames the formant frequencies are obvious from the spectrum shape. For many other frames there may be two or more plausible ways of allocating formants to spectrum peaks, where the alternatives may sometimes have either one or two formants associated with the same peak in the spectrum. In these latter cases, all plausible alternatives are offered to the pattern matching algorithm, and the best-fitting formant allocation for each state is used in computing the HMM emission probability. Although the details are very different, the concept of using multiple formant hypotheses in speech recognizers had previously been suggested by Hunt[7].

4. SPEECH DATA

At the time of writing a data base is being collected of 103 potential vocabulary words spoken by nearly 200 adult speakers from all over the United Kingdom. Some of these words, such as the decimal digits, are being spoken in connected strings, and others, intended as command words, are being spoken in isolation. Initial experiments are concentrating on a smaller vocabulary of decimal numbers and 19 of the command words. The 13-word numbers vocabulary includes 'double', 'point' for decimal fractions, and the 10 digits, with both the letter name 'O' and 'zero' as alternatives for 0. Each speaker is recording 50 5-word number strings, in which all these 13 words are used approximately the same number of times. The positions of the words in the strings are also evenly distributed, except for 'double' and 'point' which would not make sense in some positions, and 'O' and 'zero' which are never both used in the same string. Care has been taken to ensure that all valid word-pair sequences are adequately represented. Unfortunately only a small portion of this proposed data base is so far available, and there has not yet been time to process more than 40 of the speakers. Because there are approximately equal numbers of the two sexes and the regional accents are very diverse, there are not yet any large groups whose voices are similar enough to form good composite models. Of the data so far available there are six males and ten females whose accents are a reasonable approximation to RP, and no other regional group contains more than three speakers. Preliminary experiments described here have therefore had to be restricted to the two RP groups. Even within these two groups there are other aspects of voice quality which are noticeably different, and better results would undoubtedly be obtained if more RP speakers were available.

5. SPEAKER SIMILARITY ASSESSMENT

An essential facility needed when preparing a recognition system using the principles described above is some method for measuring similarity of voices. As the results of the training algorithm used for each speaker are embodied entirely in the means for the sixteen features of each model PDF, a convenient method of comparing voices is just to compare the PDF means for corresponding model states. The comparison is made by summing the squared differences between corresponding means over all PDFs, weighted by the reciprocal of the variance for each feature of each PDF. The measure so given is very similar to the negative log "emission probability" calculated for

each feature vector for all states during the recognition process, except that another array of feature means is used instead of the feature vector derived from the input speech.

To illustrate that the speaker comparison measurement does in fact give a sensible indication of sex and accent, Table 1 shows the model dissimilarity (in arbitrary units) for a total of 13 speakers of both sexes and with different accents. The speaker categories chosen include all those in the set of 40 available speakers where there were at least two in the same category. In addition the male and female RP groups (3 of each are shown) there are 3 Yorkshire females, two Edinburgh males and two Belfast females.

	MR1	MR2	MR3	ME1	ME2	FR1	FR2	FR3	FY1	FY2	FY3	FB1	FB2
MR1	0	149	151	879	785	475	481	501	851	852	1087	1233	1220
MR2	149	0	135	940	862	517	520	548	900	894	1161	1304	1312
MR3	151	135	0	895	793	503	496	523	871	887	1121	1231	1238
ME1	879	940	895	0	203	1438	1441	1484	1093	1053	1344	1361	1346
ME2	785	862	793	203	0	1327	1323	1386	948	964	1213	1273	1266
FR1	475	517	503	1438	1327	0	51	91	623	650	690	936	937
FR2	481	520	496	1441	1323	51	0	64	639	667	706	949	961
FR3	501	548	523	1484	1386	91	64	0	662	678	727	969	995
FY1	851	900	871	1093	948	623	639	662	0	180	267	576	612
FY2	852	894	887	1053	964	650	667	678	180	0	231	634	690
FY3	1087	1161	1121	1344	1213	690	706	727	267	231	0	591	622
FB1	1233	1304	1231	1361	1273	936	949	969	576	634	591	0	138
FB2	1220	1312	1238	1346	1266	937	961	995	612	690	622	138	0

Table 1. Dissimilarity between sets of word models trained on speech from different speakers (arbitrary units). The first character of the speaker code indicates sex, and the second letter shows the accent (R = RP, E = Edinburgh, Y = Yorkshire, B = Belfast).

It can be seen from the table that the greatest dissimilarity within any one category is 267 units (between FY1 and FY3), and most other within-category pairs show differences of less than 200. Between categories the differences are very much greater. It is interesting to note that with the only accent where there are both males and females (RP) the difference between the sexes (averaging at about 500 units) is less than the difference between accents for the same sex. The difference between RP and Edinburgh males is about 850 units, between RP and Yorkshire females is about 650 units, and between RP and Belfast females is about 950 units. The largest difference of all is between the Belfast women and the Edinburgh men, at over 1300 units. Detailed examination of the components of the dissimilarity calculation showed, as might have been expected, that the main contributions to the larger differences are differences in the frequencies of formants 1 and 2 for many of the vowel sounds.

	MR1	MR2	MR3	MR4	FR1	FR2	FR3	FR4
3 closest (exclusive)	96	112	107	132	53	41	56	72
same sex (exclusive)	99	99	105	125	32	37	57	62

Table 2. Dissimilarity between word models trained on 8 RP speakers, and various combined models derived from speech from groups of speakers. The 3 closest are derived by combining the models of the other 3 speakers in the same group. The same sex (exclusive) entries are combinations of all RP speakers available of the same sex, excluding the one being compared.

The same dissimilarity calculation can also be used to compare any of the speakers with composite models obtained by averaging the means of models from several speakers. Because of the square-law nature of the calculation, the

distance from each member of any pair of speakers to models averaged between them will be about one quarter of the inter-speaker difference. The more interesting differences are derived when averaging three or more speakers, and comparing the composite models so obtained with other speakers in the same category who were not included in the averaging. This latter type of comparison, is of course, the only one that is directly relevant to the operational task of recognizing speech from previously unheard speakers. Some results of these comparisons are shown in Table 2. By comparing with Table 1, it can be seen that there is only one case where the dissimilarity from the composite models is more than from the nearest other single talker, and in most cases the dissimilarity from larger groups of models is less than from composite models derived from just the three most similar talkers, even though the additional members of the larger group were always more different from the models being compared than all members of the smaller group. This effect will be discussed later.

6. RECOGNITION TESTS

As the command words in the vocabulary are fairly different from each other there are normally very few errors on these words, so the experiments reported below are all on the fluently-connected number strings. Strings of this sort present an extremely difficult task for a very simple speech recognizer, mainly as a result of word juncture and coarticulation. For simplicity of the working recognizer, it has not been practicable to have different word models to deal with any coarticulation effects, even though some of these make a lot of acoustic difference to the sound patterns. For example the strong lip rounding and tongue movement that would be associated with the end of "three" in a sequence such as "three-one" has to be modelled by the same model that would be used in "three-two". It has been necessary in the models to make the final [t] release of the word 'eight' optional. When this word is followed by 'O', for several accents there is a significant danger of mis-recognition as 'eight-two', because of similarity of the vowels in the two interpretations of the second word. Also, when one word ends with a sound which is acoustically similar to the start of the following word, the matching algorithm might find an interpretation which puts all of the duplicated sound in either the first or the second word of the pair. Obviously some of these problems could be eased by using alternative, context-sensitive, models for many of the words, but this solution has been rejected because of the extra complexity and computational load that would be involved.

To obtain a base-line performance figure for the recognition algorithm, the first experiment used the algorithm for speaker-dependent recognition. For this experiment the training was done using the first 30 strings for each speaker, and the testing was done on the remaining 20 strings (100 words). The results of these tests for the most similar four male and most similar four female RP speakers are given in Table 3, and show an average accuracy of 99.75%. In general this level of performance in user-trained mode is typical of most of the careful speakers that have been tried, with any accent. The results are, of course, worse for those speakers who are very fast, or have sloppy articulation.

Experiment 2 is relevant to the model selection method described here. In this case composite PDF means were prepared for all possible combinations of three of the four talkers in the group. Recognition for each talker was then tested on the same 20 number strings, in each case using the composite model which did not include the talker's own training data. The results are also given in Table 3, and show an average word recognition accuracy over all eight speakers of 97.75%. Although the accuracy of this recognition is quite high for this demanding task, the performance is not as high as for the speaker-trained condition. Experiment 3 repeated the tests using models derived from all speakers of the same sex and accent, excluding the one under test. The female models were all derived from 9 speakers, whereas the males were derived from only 5. The results for the male speakers are somewhat worse than for experiment 2, but for the females they are slightly better. It seems likely that a large part of the difference between the results for the two sexes may be found in the dissimilarity measurements shown in Table 2. It can be seen from that table that the extra models used in the combination for females greatly reduced the dissimilarity from the talker under test in most cases, whereas for the male talkers the changes were smaller. This difference between the male and female models could be partly explained by the much smaller number of available

talkers in the case of the males, and partly because the various male models differed from each other by a greater amount than for the females. It would be expected that the overall performance would be substantially worse when models were combined from very different types of talker, and experiments 4 and 5 were intended to illustrate this effect. Experiment 4 combined all the RP male and female models (except for the talker under test) into a single composite set. The surprising thing about these results is that all except one of the male speakers achieved a higher accuracy than in experiment 3, although all the female speakers gave noticeably worse results. The final experiment used models derived from all 13 speakers used for table 1, with widely diverse accents, and in this case the general performance was worse for most speakers. Even so, the results for speakers MR1 and MR4 were better than the results for experiment 3 even with such inappropriate models.

It is very difficult to draw strong conclusions from such small scale experiments. However, from examining the detailed components of the model differences, it seems that these fall into two categories. On the one hand, there are differences that are a direct consequence of phonetic variation. These differences are mostly in the frequencies of F1 and F2, and to a lesser extent F3. The other differences are in the amplitude features, and these seem to show much less systematic variation between different accents. The amplitude effects seem to be mostly idiosyncratic differences between speakers, e.g. some speakers tend to have more intense fricatives, or have a voice source with a different average spectrum. Averaging a large number of speakers with the same accent tends to even out these idiosyncratic effects, while maintaining the formant frequency differences that are a genuine consequence of the accent. The result seems to be that increasing the number of similar speakers generally increases the robustness of the models, and it can be hypothesized that for the male RP speakers, who were so few in number, the improvement in amplitude models from including many more speakers, even though they were of other accents and many were of the other sex, largely offset the damage to the formant frequency features.

	Errors in 100 words for experiments 1 - 5							
	MR1	MR2	MR3	MR4	FR1	FR2	FR3	FR4
Experiment 1	0	0	0	1	0	1	0	0
Experiment 2	2	3	0	3	2	2	2	4
Experiment 3	3	3	1	3	1	2	2	3
Experiment 4	0	4	0	2	7	11	5	11
Experiment 5	2	15	2	2	9	10	9	11
Experiment 6	1	5	1	1	0	2	3	3

Table 3.

Experiment 1: using word models derived from the speaker under test.
Experiment 2: using word models derived from 3 speakers, excluding the test speaker.
Experiment 3: using word models derived from all RP speakers of the same sex, excluding the test speaker.
Experiment 4: using word models derived from all RP speakers of both sexes, excluding the test speaker.
Experiment 5: using word models derived from all 13 speakers of Table 1.
Experiment 6: speaker independent recognition, using automatic model switching.

With the available data it has not yet been possible to demonstrate the full amount of automatic switching between word models that the algorithm will allow, but it has been possible to test a small scale version. In all the tests described below the talker under test was not included in any of the available models. The recognizer has sufficient memory and computational capacity to test four sets of models in parallel on this 13-word vocabulary. The initial set used the following 4 sets of combined models:

1. Mixed RP and Edinburgh males.
2. Mixed RP males and females.
3. Mixed RP and Yorkshire females.
4. Mixed Belfast and Yorkshire females.

Each of these model sets is associated with a short list of more specific models, which are loaded in priority order as soon as the cumulative mismatch score for any other sets of models exceeds the score for the best models by some threshold. Model set 1 has the combination of all RP males as its top priority, the combination of the two Edinburgh males second, and two sub-clusters of RP males as third and fourth priorities. Models set 2 has all RP males as choice 1, RP females as choice 2, and one sub-cluster from each as choices three and four. Model set 3 uses RP females as choice 1, Yorkshire females as choice 2, and two RP female sub-clusters as choices 3 and 4. Model set 4 only has two items in the short-list, which are the Belfast females and the Yorkshire females. At the next level down, only the RP speakers have any further subdivision: for each sex four different sub-clusters from the available population are provided.

The 100 words for the same eight speakers were tested in experiment 6. Although there are variations in the scores for individual speakers it can be seen that the average recognition accuracy for both males and females was as good as the best of experiments 2 and 3. The recognizer has facilities for displaying which models it is using for the recognition of every word, and also which model sets are available for every utterance. In every case the most unsuitable models were removed at the end of the first utterance, and after about three or four utterances the recognizer was mostly using the four sets of male or female RP sub-clusters, as appropriate. The speakers who actually gave better results than both experiment 2 and experiment 3 achieved this performance because the recognizer could choose whichever was the best of the four sets of available models for every utterance, or could even choose different models from word to word in within an utterance. For speaker MR2, two of his five errors occurred on the first two strings recognized, before his most appropriate models had been loaded. This type of effect is always a danger with this recognizer, but it is not usually serious because even for the first utterance one of the four sets of models available should be reasonably suitable.

7. CONCLUSIONS

The experiments described in this paper demonstrate that a recognizer with very modest computation can give a fairly high recognition accuracy for the very demanding task of connected word recognition of number strings. The results indicate that each set of word models will need to be derived from many similar speakers, and it is clear that many composite models will be needed to cover all the accent variations of English adequately. The performance should then be comparable for any regional accent of English for which there is sufficient speech data to train the word models properly. This performance would certainly not be possible for diverse local accents using the more conventional speaker-independent approach.

8. REFERENCES

[1] G. R. DODDINGTON: 'Phonetically sensitive discriminants for improved speech recognition', *Proc. IEEE ICASSP*, Glasgow, pp.556-559 (1989)

[2] J. G. WILPON, C. -H. LEE and L. R. RABINER: 'Connected digit recognition based on improved acoustic resolution', *Computer Speech and Language*, 7, pp.15-26 (1993)

[3] Y. NORMANDIN, R. CARDIN and R. DE MORI: 'High-performance connected digit recognition using maximum mutual information estimation', *IEEE Trans. on Speech and Audio Processing*, 2, pp.299-311 (1994)

[4] J. N. HOLMES: 'A very-low-cost connected-word recognizer for small vocabularies': *IEE Colloquium Digest* 1991/066 (March 1991)

[5] J. S. BRIDLE, M. D. BROWN and R. M. CHAMBERLAIN: 'Continuous connected word recognition using whole word templates', *The Radio and Electronic Engineer*, 53, pp.167-175 (1983)

[6] J. N. HOLMES: 'Use of phonetic knowledge when designing and training stochastic models for speech recognition', *Proc. Eurospeech '91*, Genoa, pp.1257-1260 (1991)

[7] M. J. HUNT: 'Delayed decisions in speech recognition - The case of formants', *Pattern Recognition Letters*, 6, pp.121-137 (1987)