Proceedings of the Institute of Acoustics

TWO-STAGE UNSUPERVISED MODEL ADAPTATION TO IMPROVE THE PERFORMANCE OF CONNECTED-WORD RECOGNITION

John N Holmes

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

ABSTRACT

If HMM word models for automatic speech recognition have not been trained by the current speaker, the recognition performance can be improved by adapting such models during use to make them better represent the observed properties of the input speech. There are two quite distinct types of inadequacy in word models that may cause recognition errors. The first type is a result of a general difference in spectral properties of the speech of the current user from those of the speech used as training data. Such differences can be caused, for example, by different microphones and differences in typical glottal source spectrum. The second class of difference is specific to particular phonetic events, depending on articulatory detail of individual words. Correction of general spectral trends is much more robust than adaptation of individual p.d.f.s, so it is advantageous to spend the first minute or so of the adaptation correcting the spectral trend effects only. This process by itself should normally improve the recognition accuracy, thus increasing the effectiveness of subsequent unsupervised adaptation for modifying the individual p.d.f.s. Results are presented for several speakers using a simple connected-word stochastic recognizer, comparing this two-stage adaptation with simple adaptation of p.d.f.s only.

1. INTRODUCTION

For many applications of automatic speech recognition the ability to recognize the speech from previously unheard speakers is essential, and therefore the models of the recognition units stored in the recognizer cannot be derived from the speech of the current user. However, the statistical properties of speech from different people normally have significant differences, caused by a combination of differences in physiology, articulatory habits and electroacoustic speech input systems. It is therefore obvious that in most cases the accuracy or robustness of recognition would be improved if the recognition models were specifically set up to suit the current input speech. One way of achieving this improvement is to gradually adapt the parameters of the recognition models during use to make them closer to the speech properties they are meant to represent. In speech recognition the term "adaptation" is normally used to refer to a system in which previous knowledge is modified in the light of new speech data. With typical stochastic recognizers a straightforward method for the adaptation is to use the dynamic programming of the recognition process for labelling each frame of input data to identify its most likely model state, and then to modify the p.d.f.s associated with these states to incorporate the statistics of the new input.

It is most convenient for the user if the adaptation can be done without supervision (i.e. such that even a wrong recognition changes the models). Provided the adaptation for each occurrence of a recognition unit only changes the model parameters by a very small proportion of the parameter difference, occasional misrecognitions do little harm. The essential requirement is merely that, for each state of each recognition unit, correct recognition occurs most of the time and incorrect recognition is comparatively rare.

Of the three types of difference in properties of speech signals mentioned earlier, the electroacoustic differences will affect the spectrum of all speech sounds in a similar way, whereas the differences in articulatory habits will be specific to the phones being uttered. Although some physiological differences may cause effects that are specific to particular types of phone, many of these differences, particularly those affecting the spectrum of the glottal source, will apply generally to a high proportion of the speech. Thus electroacoustic and some physiological effects will manifest themselves mainly as an apparent variation of the frequency response of the input circuit.

For those differences that are not phone-specific, it would obviously be possible to adapt models much faster if such differences changed all models to which they could apply, for every frame of the input speech. It would then be possible to achieve all the benefits of adapting the general properties on a comparatively small amount of speech, even when many of the recognition units have not been used at all. Such a principle was described in 1981 by Hunt [1], and has since also been investigated by other workers [2,3].

Most speech recognition systems produce some representation of the logarithmic short-term spectrum envelope of the speech signal during the initial analysis process. A convenient way of applying general spectral shape correction is therefore simply to add the current estimate of the spectral correction function to the spectral intensity derived at each frequency, rather than to modify the p.d.f.s of the individual models. If this general stage of adaptation is done first, the adaptation of the individual p.d.f.s will then only have to correct for speech properties which are not covered by the general correction.

2. DESCRIPTION OF SPEECH RECOGNIZER

The adaptation scheme described in this paper has been applied to a very-low-cost small vocabulary connected word recognizer, implemented entirely in software on a 6502 8-bit microprocessor. More detailed descriptions of various aspects of this recognizer have been given elsewhere [4,5,6], so only the main features will be included here.

The speech analysis splits the speech into three bands with simple bandpass filters, roughly corresponding to the ranges of the lowest three formants. The F3 band spreads up to about 3.6 kHz, and will thus often also contain F4. Within the F1 and F2 bands the logarithm of the short-term total power and the frequency of the spectral centroid are measured every 32 ms. Although earlier

versions of this recognizer also made similar measurements on the F3 band, the F3/F4 overlap made the frequency measurement for this band very unreliable, so the current system measures only the log power in the top band. There are thus five raw features measured every 32 ms, corresponding very roughly to F1, A1, F2, A2, and A3. The recognition process uses the two frequency features, F1 and F2, and amplitude features derived from the three basic amplitude measurements, as follows:

i. MAX = maximum of A1, A2, A3
ii. D1 = A1 - A2
iii. D2 = A2 - A3
iv. DA1 = A1 - A1 for the previous frame

The amplitude features are transformed in this way to reduce the off-diagonal terms of the feature covariance matrix, and DA1 is included because time differences have been widely found beneficial in speech recognition. Other time differences could not be included because computation limitations in the 6502 prevented the total number of features from being increased above six.

The recognition algorithm itself is a fairly conventional implementation of the one-pass continuous recognition algorithm [7], using Viterbi decoding of an HMM-type structure with single Gaussian continuous-density p.d.f.s. and diagonal covariance matrices. However, explicit state duration penalties are used instead of the conventional HMM transition probabilities. The topologies of the word models are chosen separately to suit each word in the vocabulary, on the basis of a priori phonetic knowledge, modified by experience with analysing the causes of recognition errors.

The means of the state p.d.f.s are stored within the recognizer's read-only memory for a large number of different types of speaker. For most applications, a new speaker speaks a few specified word sequences in a very short enrolment session, to enable the recognizer to choose the model set which fits best to his/her speech. The variances of the six features are the same for all sets of models, but were chosen separately for each state after analysing the speech from many talkers.

As currently implemented this recognizer has been trained for a vocabulary of 14 words, comprising the 10 English decimal digit names and the words "point", "enter", "delete" and "cancel". In spite of the crude analysis process, the small number of features and the low frame rate, fairly careful speakers typically achieve about 98% word accuracy for connected digit recognition.

3. THE TWO-STAGE ADAPTATION PROCESS

During the first stage of adaptation, it is necessary to collect data to estimate the error in general spectral properties. As each recognized word is output, the traceback path through the states is used to label each frame of input data with its corresponding state number. For each frame the values of MAX, D1 and D2 of the chosen p.d.f. means are decoded to get the corresponding

values of A1, A2 and A3. Adaptation is only allowed where the amplitudes of the features and the models both represent medium- or high-level speech sounds. For the top channel, which has no frequency parameter, the differences between the A3 input feature and its corresponding p.d.f. mean are accumulated while the data collection is in progress. A similar process is carried out for A1 and A2, except that the differences are accumulated separately for each of the possible values of the corresponding frequency feature. The average correction at each frequency is calculated after 256 frames of data have been collected, excluding low-level periods and non-speech sounds. This much data is usually obtained from about seven 5-word strings. The frequency-dependent correction functions for A1 and A2 are then smoothed in the frequency domain by convolving with a window seven quantization intervals wide (175 Hz for F1 and 350 Hz for F2).

The calculated correction functions are applied to all subsequent values of A1, A2 and A3 before they are used for the recognition. A single correction value is used for A3, and the A1 and A2 corrections are chosen according to the values of F1 and F2 measured in each frame. For speakers for whom general spectrum trends differ substantially from those of the models, the allocation of frames to states is likely to be better after this correction is applied. For this reason, the general adaptation process is then repeated for another 256 frames, to get an improved correction characteristic. The final frequency-response correction functions so derived generally have noticeably larger values than are obtained from the first 256 frames. As soon as the second group of 256 frames has been collected, the general adaptation is terminated and the model adaptation is started, with the final values of the general correction in use.

The adaptation of the individual p.d.f.s is applied to the means only. It uses exactly the same state labelling system as the general adaptation, but the differences between the model means and the input features are used differently. For each frame, the values of the state means are modified by an amount equal to 1/64 of the difference between their old values and the new feature values. After a sufficient amount of input speech has been received each p.d.f. mean will thus tend to the exponentially weighted average value of the corresponding feature for the current speaker, with a weighting time constant of 64 frames (i.e. about 2 s of actual speech for each state). However, as the adaptation must be applied to all states of all word models (a total of 144 states), several minutes of speech utterance time are needed for a substantially complete adaptation for all words.

4. EXPERIMENTS

The speech data used for training the word models used for these experiments contained a fairly even mix of adult male and adult female speakers, with a variety of voice qualities and many regional accents of English from various parts of the UK and Ireland. This speech was recorded using an electret microphone mounted about 10 cm from the speakers' lips. The frequency characteristic of this microphone was substantially flat within the 200 Hz to

3600 Hz band used by the recognizer. Separate sets of models trained using data from 65 speakers were collected into 15 groups by an automatic clustering algorithm. The p.d.f. means for corresponding states were then averaged within each cluster to provide 15 sets of word models for use in the recognizer.

Some additional speech data was also available that had been recorded using a telephone handset fitted with a rocking-armature microphone. To a first approximation the handset microphone gave a 6 dB/octave lift over the frequency band of interest, and this difference was corrected in the experiments. However, the handset also had some noticeable departures from this simple characteristic which were not corrected, with peaks and troughs of at least 6 dB in some frequency regions.

For the adaptation experiments 3 male and 3 female speakers were chosen for each type of microphone. The tests were all done on recognition of five-word connected number strings, using eleven of the vocabulary words (i.e. the decimal digits and the word "point", because some strings included decimal fractions). A total of 40 strings were used for each talker, in which each of the digits occurred 19 times and "point" occurred ten times. The strings had been chosen to get a fairly even distribution of all possible word pairs, and all digits occurred an equal number of times in initial and final position. The 40 strings for each speaker were divided into two groups of 20, such that in every group of 100 words each digit occurred either 9 or 10 times.

For each adaptation run a set of clustered models was chosen which was appropriate for the voice quality and regional accent of the speaker. For all the handset data the models used were those given as the best fit by the automatic enrolment process. For two of the runs using the electret data the best set of models was also chosen, but for all others the second best was used, to provide a more difficult test for the adaptation. In the two best-set cases the talker under test had contributed to the model generation, but only as one of at least 6 members of the cluster. In all the other cases the talkers made no contribution to the unadapted models used in the tests.

The adaptation was always performed using only half of the available data for each speaker, to leave the remaining unused data available for testing. Because only 100 words of adapting material was thus available, it was necessary to re-use the data a number of times to get sufficient adaptation. This re-use would be expected to give adaptation inferior to that which would have been derived from new speech material, because any non-typical effects of specific utterances would be magnified.

As a result of chance variation in timing of the frame boundaries relative to the input speech, repeated analyses of the same speech recordings will in general give different feature vectors for most frames. As these differences make quite large chance differences to the recognition errors, all the experiments described in this paper have been done using files of feature vectors, analysed during a single playback of each speech recording.

The error results given show the sum of word substitutions, insertions and deletions. For each speaker the number of errors are presented for each group of 100 words, with the following adaptation conditions:

i. No adaptation;
ii. General adaptation only;
iii. Model adaptation only using 200 words;
iv. General adaptation followed by model adaptation using 200 words;
v. Model adaptation only using 600 words;
vi. General adaptation followed by model adaptation using 600 words;

For adaptation using 600 words, the p.d.f. means of the more frequently used states would have been adapted to somewhere near their asymptotic values, but the others would still have had quite a long way to go.

For each condition all the results are presented separately for the two groups of 100 words. The second number (words not used for the adaptation) is obviously the only one that is relevant to the operational situation, but comparison with the first figure is useful to give some idea of the extent to which 100 words from this eleven-word vocabulary can be regarded as representing the properties of the words for any given speaker.

5. RESULTS

The results of the experiments are shown in Table 1. As can be seen from the table, there are wide differences between talkers in the number of errors, and in the improvement being given by the adaptation process. Some of the talkers were fairly slow and careful in their speech, and others were quite fast with a lot of coarticulation between words. In some cases the speaker's regional accent was not well represented in the few sets of word models available. The very simple structure of the recognizer, combined with the low frame rate and approximations caused by the 8-bit arithmetic, mean that chance factors will have influenced the results considerably. In fact, even playing the same 200-word audio recording of a typical speaker repeatedly into the recognizer often gives error scores ranging between 0 and 5 on different runs. It is therefore more useful to look at general trends obtained by combining the results from several speakers. The column sums show that the right-hand column for each condition usually has more errors than the left-hand column. However, as the ratio of errors in each pair of columns is in most cases fairly close to the ratio for the condition with no adaptation, it seems likely that the cause is merely that the second 100 words for most speakers were spoken a little less carefully than the first 100 words. There seems to be no evidence to suggest that the reduction of error rate after adaptation is generally greater when measured on the adapting words than on the test set. In fact, in the "mod.600" condition the improvement is substantially greater in the test set.

When talker M1 was used with his best set of models (set 6), the models fitted very well to his speech and the general adaptation gave very little alteration to the spectrum trends (no more than 1 dB over most of the range). Not surprisingly, therefore, the general adaptation made no difference to the

spkr.	model set	no adap.		general only		mod.200 only		gen. + mod.200		mod.600 only		gen. + mod.600	
Electret													
M1	6	1	2	1	2	1	1	1	1	0	0	0	0
M1	3	9	6	4	5	4	3	3	2	5	2	2	2
M2	1	6	10	7	7	7	7	6	3	4	3	2	0
M3	3	11	10	7	8	6	11	8	9	5	6	4	5
F1	8	1	3	2	6	1	3	1	4	2	2	1	3
F2	9	9	5	6	2	3	1	2	3	5	1	3	1
F3	8	11	12	9	8	8	11	5	6	6	5	4	3
Handset													
M4	6	8	7	6	1	6	4	6	2	6	2	5	3
M5	6	12	20	4	12	9	12	2	9	6	10	2	5
M6	2	22	25	12	19	20	23	10	19	21	23	10	14
F4	8	3	2	2	1	2	1	1	0	2	0	1	0
F5	8	6	6	5	3	4	4	6	2	1	3	5	3
F6	8	12	19	8	15	8	15	7	13	7	12	5	15
sums		111	127	73	89	79	96	58	73	70	69	44	54

Table 1. Number of errors given in the adaptation experiments for 12 talkers. In each condition the left-hand column gives scores for the 100 words used in adaptation, and the right-hand column is for the words used in testing only.

results. With a sub-optimum set of models (set 3), adaptation gave a considerable improvement for the same speech data. Speaker F1, using her best set of models, gave a low error rate without adaptation, and the general spectrum trend was presumably about right. In this case, chance spectral mismatches for many states actually caused the general adaptation to make the results noticeably worse; it was not until the end of the model adaptation that the results were as good as they were originally. Speakers M6 and F6 from the handset data gave much lower accuracy than all the other speakers shown. In the case of M6, this appeared to be because of an unusual accent, not well represented by any of the available word models. Model adaptation on its own seemed superficially to make very little improvement to the recognition score. However, analysis of the 44 errors after 600 words of model adaptation revealed an interesting result. Recognition of most words was greatly improved, but 17 of the 19 utterances of "two" were misrecognized as "three", which was much worse than before adaptation. When using both types of adaptation the total number of errors was reduced to 24, which were fairly well spread throughout the vocabulary. In this condition, therefore, a useful degree of recognition could be achieved for all words. Speaker F6 was by far the fastest of all speakers used for these experiments, and extreme coarticulation appeared to be the main cause of her residual errors.

6. DISCUSSION AND CONCLUSIONS

The results presented here clearly show great benefit from unsupervised model adaptation, and a general advantage from splitting the adaptation into two stages. However, because the adaptation depends critically on how input frames are allocated to the model states, it is important that the basic recognition should be fairly good. Although the adaptation process will usually make mediocre recognition a lot better, the results for speaker M1 show that the better state allocation from a better initial set of word models will give a better end result. In almost every case shown, adding the general adaptation stage has improved the ultimate performance. In the case of talker M6, the improvement given by the general adaptation prevented the large number of errors on one word from getting a lot worse during the model adaptation.

In the results given, the same amounts of model adaptation were used both with and without the general adaptation, so more time was needed to include both types. However, the extra time needed for the first stage is a very small fraction of the time needed to adapt the p.d.f.s to anywhere near their asymptotic values, so this extra time would be of little importance in practice.

In an operational situation there are two main reasons why the adaptation should be even more effective than reported here:

i. The adaptation would always be from new material, so reducing chance effects from non-typical utterances;
ii. Immediate feedback would cause users to speak more carefully for words that caused frequent errors, so making the adaptation work better.

Although the very crude recognizer described in this paper shows a great improvement from the two-stage adaptation scheme, a more conventional recognizer, with many more features to represent the overall spectral cross section, seems likely to gain even more benefit.

7. REFERENCES

[1] M J HUNT, 'Speaker Adaptation for Word-Based Speech Recognition Systems', J Acoust Soc Am 69 p S41 (1981)
[2] S J COX & J S BRIDLE, 'Unsupervised Speaker Adaptation by Probabilistic Spectrum Fitting', Proc IEEE ICASSP, Glasgow, pp294-297 (1989)
[3] C-H LEE, C-H LIN & B-H JUANG, 'A Study on Speaker Adaptation of Continuous Density HMM Parameters', Proc IEEE ICASSP, Albuquerque, pp145-148 (1990)
[4] J N HOLMES, 'A Very-Low-Cost Connected-Word Recognizer for Small Vocabularies', IEE Colloquium Digest 1991/066 (1991)
[5] J N HOLMES, 'Methods and Apparatus for Spectral Analysis', British Patent Application 9002852.3 (1990)
[6] J N HOLMES, 'Use of Phonetic Knowledge when Designing and Training Stochastic Models for Speech Recognition", Proc. Eurospeech '91, Genoa (1991)
[7] J S BRIDLE, M D BROWN & R M CHAMBERLAIN, 'Continuous Connected Word Recognition Using Whole Word Templates', The Radio and Electronic Engineer, 53 pp167-175 (1983)