SPEAKER INDEPENDENT TELEPHONE SPEECH RECOGNITION USING SUBWORD UNITS

R Y John & D J B Pearce

GEC-Marconi Limited, Hirst Research Centre, Elstree Way, Borehamwood, Hertfordshire, WD6 1RX, England

1. INTRODUCTION

The development of speech recognition applications in telephony is a highly active area of work. Some applications are now available on certain commercial telephone networks. These seek to enhance or automate operator services and also telephone facilities which were previously totally controlled by touch-tone codes - for example, with the voice interactive phone only a single touch tone code is required to reach the facilities menu, from which point voice commands may be used to access or activate facilities. The speech recognition methods used in these applications employ whole word modelling techniques as the vocabulary involved is small. A subword approach extends the range of applications. Voice dialling by name is one application which is dependent on subword recognition. Another example, for which the first field trials have just come to an end, is the use of speech recognition in telephone based market research. This is an application which the GEC-Marconi Hirst Research Centre (HRC), in collaboration with Marconi Speech and Information Systems (MSIS)[1] and AGB Taylor-Nelson, has been working on. A detailed analysis of field data is not yet available, but the response of users has been very favourable and the viability of the technology for this application has been demonstrated.

The objective of this paper is to report work carried out to produce subword models for speaker and vocabulary independent telephone speech recognition. In Section 2 the telephone speech corpus, a key element of the work, is discussed. Then, in Section 3 the subword method employed is described. Results of experiments are presented in Section 4. Finally, concluding remarks are made in Section 5.

2. TELEPHONE SPEECH CORPUS

The conditions encountered in telephone speech recognition are generally not ideal [1]. One problem is the noise associated with the transmission and switching equipment of the network. Another is the spectral distortion of the voice signal which occurs in two places, namely, at the speaker end and over the telephone channel. At the speaker end room

[1] the collaborators in MSIS are now part of GEC-Marconi Secure Systems Ltd.

acoustics (reverberation) and the microphone transducer [2] affect the voice spectrum. Signal bandwidth over the telephone channel is limited to between 300 and 3400 Hz within which there is a variation in spectral attenuation. In addition this variation differs from channel to channel. To explicitly take into account all the above factors would make the speech recognition problem even more complex.

Results of experiments in noisy conditions have been reported [1]. A major finding was that good performance was obtainable if training and test conditions were exactly the same. More important, from a practical point of view, was the finding that even if the conditions were not identical a reasonable performance could be obtained if training conditions only approximated test conditions. This has led to the use of telephone speech corpora in speech recognition over the telephone network: by collecting training speech utterances using several telephone lines over a period of time the criterion of having training conditions which approximately match test conditions is met.

The establishment of the telephone speech corpus was a multi-faceted task which was carried out in collaboration with MSIS and AGB Taylor-Nelson [3]. The Flexicall System - an interactive voice response system developed by MSIS - was used in automating the telephone collection process and structuring collected utterances into individual computer speech files.

The corpus consists of two sections, namely, training and testing (see Figure 1). Speakers in the training section belonged to either a primary or secondary set. Speakers comprising the former consisted of eight males and females from the London and South-East (L&SE) region and were all employees of the three project collaborators. All primary speakers generated a single utterance of each and every word of the training vocabulary. The secondary speakers (about four hundred in total) were drawn uniformly (approximately) from both sexes and from two dialectal regions, i.e., L&SE and Lancashire and Yorkshire (L&Y). The ages of speakers fell into three categories: 16-34, 35-54 and 55 plus. Speakers here only produced utterances of subsets of the training vocabulary. The vocabulary subsets were chosen such that the total number of utterances of each word for the secondary set was equal to that of the primary set.

The training vocabulary was designed with the aim of enabling vocabulary independent training of subword models. A list of words was chosen to provide good coverage of all major phonetic contexts whilst limiting the total size to be around 2000 words - the final size was 2090. The objective was to provide sufficient training data for both moderately rare and common phonetic contexts, which in turn would ensure the availability of trained models for most contexts in a typical (English) application vocabulary. It should perhaps be mentioned here that phonetic transcriptions for words in the vocabulary were based on Southern English pronunciations as specified in the Longman Pronunciation Dictionary [4].

The test section of the corpus is made up of utterances from independent speakers drawn uniformly from both sexes and dialectal regions. These utterances represent five different application vocabularies, the average size being about 65 words. The majority of these words do not occur in the training vocabulary.

The collection process was spread over a period of time, with each speaker being required to utter only 55 words per recording session. This was necessary to avoid problems associated with voice characteristics changing over a session and human boredom. The time window for each utterance was three seconds and a sampling rate of 8 kHz was used. The corpus thus represents about sixty hours of voice recordings and occupies 3 Gbytes of disk space.

3. THE SUBWORD APPROACH

3.1 Subword units

Telephony applications of speech recognition generally deal with isolated word responses to pre-recorded questions. A possible framework for such applications is shown in Figure 2. When activated the system would play to the telephone user the first of a series of previously recorded questions. The (isolated word) answer would determine the next question to be played, and subsequent question and answer processes would determine the route taken by the user through the application. The facility to repeat questions or end the session at any point would be a feature of such an application. The use of subword units in recognition is desirable for two reasons. Firstly, it provides flexibility and allows the creation of new vocabularies quickly and easily. And, secondly, it permits an increased vocabulary range, well beyond the practical limits associated with a whole word approach.

A variety of subword units have been proposed over the years. The most successful of these have involved the use of context-dependent phone models. One popular unit has been the triphone which models each phone by taking into account both its left and right neighbours. Achieving an inventory which contains all triphones naturally occurring in speech is not feasible: an impractically large corpus would be required to ensure that all triphones are adequately trained. Thus, for a system which uses a finite training corpus, the probability of encountering phone contexts in test data that are not present in the training data is fairly high. One way to overcome this limitation is to use a decision-tree clustering approach which makes use of the fact that some groups of phones can have similar effects on neighbouring phones. Thus for contexts in test data that are not present in the training data the tree can be traversed to find the most appropriate context model.

An alternative to the original triphone approach is employed by the HRC subword recogniser [5]. Triphones are built up from an appropriate sequence of units called phonicles (from *phon*etic par*ticle*). With this approach phonicle models may be shared across different triphones, greatly reducing the number of possible context dependent units. Memory requirements are therefore reduced and the problem of undertraining lessened.

The phonicle system splits each phone into two phonicles, or "biphonicles", with the first being dependent on the left context and the second on the right context. Triphones may be produced by concatenating appropriate pairs of biphonicles. A hidden Markov modelling (HMM) method is used. To take some account of the anticipatory nature of coarticulation in English the left and right biphonicles are assigned one and two states respectively.

3.2 Training and recognition procedures

The recognition system [6] employs HMMs with a single diagonal covariance multivariate Gaussian probability density function, the covariance matrix being pooled over all states and models. The models incorporate a simple left-to-right topology with self transitions and no skips. A special single-state silence model is used. Model training is performed by an "embedded" Baum-Welch re-estimation procedure, using only orthographic transcriptions, a pronunciation dictionary and unmarked training data. It should be mentioned that no hand-labelling of training data is performed either at the word or subword level.

Training is carried out in three stages with each stage progressively becoming more context specific. The process starts off by initialising a set of context-independent models to identical values computed from the centroid and variance of the entire training data set. Five iterations of the Baum-Welch (BW) re-estimation procedure are then carried out yielding a trained set of context-independent models. The next stage involves models at the first level of context-dependence: these models depend on the broad class of the adjacent phoneme, the broad class being defined by the place of articulation. Initial estimates for these models are obtained from the trained set of context-independent models. Two iterations of the BW procedure are performed. These trained broad class models are then used to initialise models at the final level of context-dependence where biphonicles are dependent on neighbouring phonemes. Again re-estimation is carried out twice. It may be noted that at each stage a model was trained only if the number of occurrences of the associated biphonicle in training data exceeded a threshold, namely, three.

Word models for a given test vocabulary are formed by concatenating the appropriate sequence of biphonicles using a pronunciation dictionary and a list which has the frequency of occurrence of all biphonicles at the different (context-dependent) levels - trained models from all levels are retained and the frequency list enables the most context-dependent biphonicles to be used.

4. EXPERIMENTS

Experiments were performed along the lines of earlier work [6]. An FFT-based Mel cepstrum analysis was carried out on speech files. The time window used was 32ms and it was applied at 16ms intervals. The speech coefficients, in frame sets, arising from the analysis were augmented with time derivatives. These were obtained for each frame by computing the difference between following and preceding frames.

Results reported here are from trials performed to find out, firstly, how well the 'baseline' system would perform for the different vocabularies, and secondly, to observe the significance of including secondary speakers in the training process. The decision was made to preclude speakers from the Lancashire and Yorkshire region from these (initial) experiments as phonetic transcriptions for the training vocabulary were based on Southern English pronunciations. Separate experiments were carried out for male and female speakers, as gender-dependent models are known to enable better recognition [7]. For each gender, recognition performance was recorded for models trained on utterances from the primary set, the secondary set and that formed by pooling the two (see Table 1). In each recognition run with male models the Shares and Market-Research tasks consistently achieved the best and worst scores respectively. With female models the corresponding tasks were Cities and Objects. That this should be so is perhaps linked to the fact that the Shares and Cities vocabularies consist of poly-syllabic names e.g *Nottingham*, *Ultramar*, etc, whereas the Market-Research and Objects vocabularies are made up of mainly monosyllabic words, e.g., *me*, no, nought, etc.

The influence, if any, exerted by secondary speakers should be reflected in the way average recognition error rates change. The average error rates obtained by using models trained on male primary and secondary speakers are 8.5% and 6.6% respectively. Models trained by pooling primary and secondary speakers yield a figure of 6.6%. With models trained on female speakers the corresponding rates are 7.0%, 8.4% and 6.8%. Even though models produced using pooled primary and secondary speakers generate the best error rates for both male and female trials, the improvement in performance is not vastly better. This is similar to previous findings made with studio quality recordings [7].

In speaker and vocabulary independent isolated word recognition, where no language model is employed, good performance is dependent on accurate phoneme recognition. There are no standard telephone speech tasks around against which performance may be rated. A way forward might be to make a comparison against human listener performance for the same test sets. It is worth noting, however, that the sufficiency of the above performance figures for a market-research application has been demonstrated by recently concluded field trials - the figures are "sufficient" in the sense that a well designed or structured dialogue interface can be used to avoid or reduce problems of misrecognition.

5. CONCLUDING REMARKS

An inventory of HMMs has been generated for the development of speaker and vocabulary independent speech recognition applications in telephony. This has been achieved by applying a modelling technique which uses subword units (biphonicles) to a fairly large telephone speech corpus. The best average recognition error rates for five different task vocabularies were 6.6% and 6.8% for male and female speakers respectively.

6. ACKNOWLEDGMENTS

The above work was partially funded by the Department of Trade and Industry. This support is gratefully acknowledged.

7. REFERENCES

[1] L. Rabiner and B. Juang, Fundamentals of Speech Recognition, PTR Prentice Hall, Englewood Cliffs, New Jersey 07632, 1993.

[2] Wang et al, "A Novel Approach to the Speaker Identification over Telephone Networks", Proc. ICASSP, Vol II pp 407-410, April 1993.

[3] W. Blyth and H. Piper, "Speech Recognition - A New Dimension in Survey Research", Annual Conference of the Market Research Society, March 1994.

[4] J.C. Wells, Longman Pronunciation Dictionary, Longman UK, 1990.

[5] W.J. Holmes and D.J.B. Pearce, "Sub-word Units for Automatic Speech Recognition of any Vocabulary", GEC Journal of Research, Vol II, No 1, Nov 1993.

[6] W.J. Holmes et al, "Allophone Modelling for Vocabulary-independent HMM Recognition", Proc ICASSP, Vol II pp 487-490, April 1993.

[7] F. Kubala and R. Schwartz, "A New Paradigm for Speaker-independent Training", Proc ICASSP, Vol I pp 833-836, 1991.

Table 1 Recognition error rates (%) for subword recogniser (dialectal region: L&SE)

MALE

Application	Size	Primary	Secondary	P+S
Banks	71	8.7	7.2	6.7
Cities	65	6.5	4.7	4.7
Market-Research	55	14.4	9.5	11.3
Objects	64	7.4	8.5	8.0
Shares	68	5.4	3.0	2.4
Average	65	8.5	6.6	6.6

FEMALE

Application	P	S	P+S
Banks	5.7	7.6	5.2
Cities	4.1	7.3	4.1
Market-Research	7.8	6.9	7.1
Objects	9.7	12.4	10.2
Shares	7.9	7.9	7.5
Average	7.0	8.4	6.8

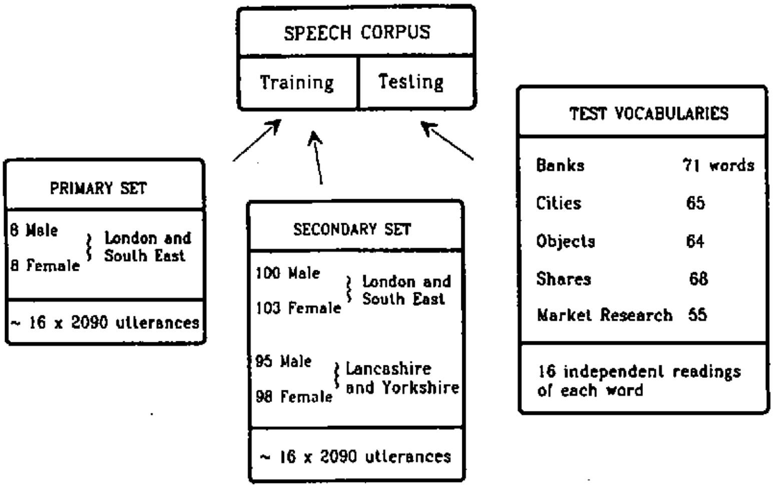

Figure 1 The Telephone Speech Corpus

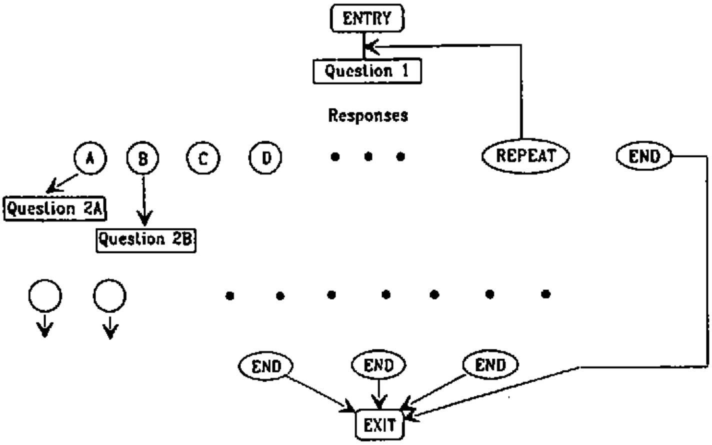

Figure 2 Generalized framework for telephony applications