Improving Accuracy of Telephony-based, Speaker-independent Speech Recognition

Daniel Azzopardi, Shahram Semnani, Ben Milner, Richard Wiseman
[daniel.azzopardi, shahram.semnani, ben.milner, richard.wiseman]@bt.com

Speech and Image Processing Unit, BT Laboratories, Martlesham Heath, Suffolk, IP5 3RE, UK

ABSTRACT

A combination of techniques for increasing recognition accuracy has been developed for an automated corporate directory system with 120,000 entries. Using a traditional recogniser an accuracy of around 60% has previously been obtained for both a 156 town name task and 1108 road name task. Techniques presented in this paper comprise front-end modifications, context dependent models, improved lexicon and noise modelling. This resulted in an increased recognition accuracy of around 90%.

1. INTRODUCTION

Good recognition performance of isolated words over the telephone is a major requirement to enable BT to provide useful voice activated services. This paper describes a combination of techniques which were used to obtain a large improvement in recognition accuracy of town names and road names.

The main components of a speech recogniser are shown in Figure 1. The first stage is speech capture which includes sampling and digitally encoding the signal. The signal is filtered, and the filter outputs are transformed to feature vectors. The speech recogniser then uses hidden Markov models (HMMs) to match the sequence of feature vectors to phonemes which make up the English language. The output of this pattern matching stage is a number of hypothetical sequences of phonemes which then have to be matched to words in the vocabulary of the recogniser.

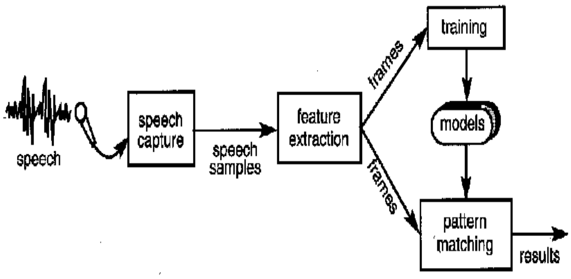

Figure 1: Main components of a speech recogniser

To improve the accuracy of recognition, more robust feature extraction, context dependent models (HMMs), a lexicon containing pronunciation variations, and improved noise modelling were used. The organisation of the paper is as follows: the first three sections introduce the theoretical aspects of robust feature extraction, triphone models and continuous speech effects. The next section describes the use of noise modelling and the following sections describe the experiments performed and results which were obtained.

2. ROBUST FEATURE EXTRACTION

A speech signal is generated by convolving the excitation signal from the lungs with the frequency response of the vocal tract. For speech recognition it is usually the vocal tract component which gives best discrimination between speech sounds. In most feature extraction methods, cepstral analysis is used to extract this vocal tract component.

There are essentially two routes for extracting the cepstrum of a speech signal - via a discrete Fourier transform (DFT) or via linear predictive (LP) analysis [1]. More recently, modifications to conventional cepstral processing have attempted to include attributes of the pyschophysical processes of human hearing into the analysis. For example, the DFT cepstrum has been modified to incorporate a Mel-scaled filterbank giving the so called in Mel-frequency cepstral coefficients (MFCCs). A similar process to the Mel filterbank is used in perceptual linear predictive (PLP) analysis, where a set of critical-band filters are convolved with the speech spectrum. These modify the spectrum according to perceptual measurements of human hearing and lead to PLP cepstrum.

The original front-end configuration consisted of conventional 8-D MFCC speech features augmented by a velocity vector and velocity log energy term. This gave a 17 dimensional feature vector and proved to be adequate for simple tasks. However for the town name and road name tasks the performance and robustness was deemed to be unsatisfactory.

Telephony speech is subject to a number of degradations such as background acoustic noise and channel distortions. It is thus important for any systems to be as robust as possible to these distortions. The MFCC derived feature described earlier has no implicit or explicit robustness. To improve this situation a two stage improvement of the feature vector was employed and resulted in the so called RASTA cepstral-time matrices.

2.1 Cepstral-Time Matrices (CTMs)

The cepstral-time matrix provides an alternative framework to differential parameters for encoding the temporal variations of speech into the feature vector [2]. Differential parameters take into account the speech dynamics by taking a difference or regression across static cepstral vectors. The cepstral-time matrix is computed by taking a discrete cosine transform (DCT) across a stack of typically 7 cepstral vectors, with the resulting columns of the matrix representing the different temporal regions. Static components of the speech signal are contained in the zeroth column, with successive column containing faster and faster speech dynamics.

For speech recognition, the lower right portion of the matrix is usually retained as the speech feature (as illustrated by shading in Figure 2), with the remainder discarded.

2.2 RASTA Filtering

The RASTA filter (RelAtive SpecTrAl) was first proposed by Hermansky in 1991 [3] and is an additional front-end operation which simultaneously reduces communication channel effects and noise distortion by bandpass filtering the time series of feature vectors. The RASTA filter, H(z), is implemented as an IIR bandpass filter,

$$\mathrm{H}(z) = 0.1z^4 \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{1 - 0.98z^{-1}} \quad \ldots (1)$$

The filter can be applied to either the log filter bank vectors or to the MFCCs. In this implementation RASTA processing is applied to the MFCC vector stream.

2.3 RASTA-CTMs

The combination of RASTA filtered MFCCs and cepstral-time matrices is referred to as RASTA-CTM. These features are generated using the conventional CTM approach with the MFCC vector stream being RASTA filtered prior to stacking - shown in Figure 2.

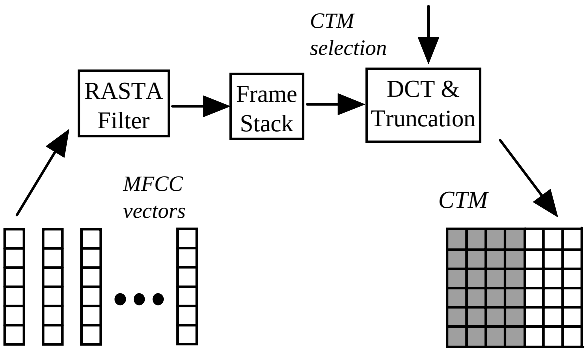

Figure 2: Generation of CTMs

3. CONTEXT DEPENDENT MODELS

Initial attempts at building a recogniser were based on monophone speech models. Although the system can provide an acceptable level of performance for small tasks, it suffers performance degradation when used for large vocabulary tasks.

This is mainly due to the inadequacy of monophones to model context variations in an utterance. It has been shown that context dependent models [4], provide better performance than monophone models. However using context dependent models, such as triphones, the number of distinct contexts can lead to a prohibitively large number of models. Hence the main problem in building context dependent models is maintaining a balance between model complexity and the available training data. To overcome this problem model based Decision Tree Clustering (DTC) [5] was utilised. A model based decision tree is a binary tree in which each node in the tree (except the terminal nodes) is associated with a phonetic question - Figure 3.

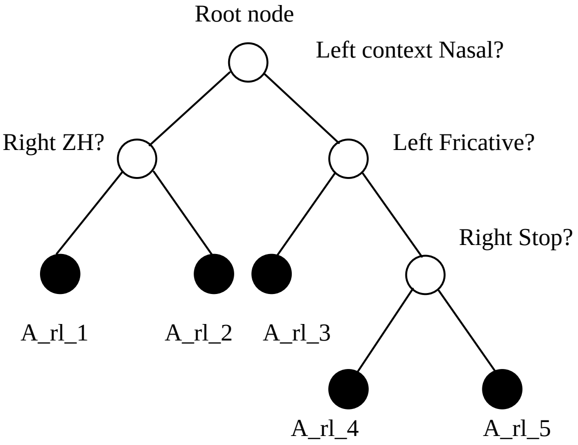

Figure 3: Decision tree for phoneme “A”.

DTC uses a set of phonetically based questions that allows the grouping of acoustically similar models, providing a more robust estimates of speech models. The question set generally determines the phonetic group, for example fricatives, vowels, stops, glides, etc.

A decision tree was constructed for each of the monophones in the phoneme set. The process of building the tree was as follows.

1. For each monophone, collect appropriate amount of acoustic data and build a single mode HMM representing the root node of the tree.
2. Ask all the questions at terminal nodes and for each question, initially split the data into two child nodes.
3. Use the data at terminal node to build a single mode HMM, and hence calculate the likelihood of the model.
4. For the node under consideration, select the question that provides the best splitting likelihood.

Continue steps 2 to 4 until a predetermined likelihood score is reached or a threshold indicating the number of examples in the terminating node is passed.

4. CONTINUOUS SPEECH EFFECTS

People rarely pronounce words as a concatenated list of canonical baseform transcriptions; even the most careful speaker deviates from the canonical as a result of articulatory limitations. Since the recogniser relies on an accurate representation of the words it is to recognise, knowledge of the kinds of changes made to these baseform transcriptions would be of great use. Two general approaches exist for the generation of alternate phonemic transcriptions: data-driven and rule-based; experiments detailed in this paper utilise the latter method. A total of six general rules [6,7] were used. Table 1 shows an example of the effect of each rule:

Rule	***Example***
Assimilation	"tin can" /T I **N K** AA N/ → /T I **NG K** AA N/
Coalescence	"would you" /W OO **D Y** UU/ → /W OO **J** UU/
Consonant Elision	"old man" /O **L D M** AA N/→/O **L M** AA N/
Phonemic Elision	"run along" /R U **N A L** O NG/→/R U **N L** O NG/
Intrusive 'r'	"far away" /F **AR A** W AI/→/F **AR R A** W AI/
Allophonic Variation	"how old" /H **OU O** L D/→/H **AA O** L D/

Table 1: An example of each of the six rules used

Although the rules are well suited to application across word boundaries, as suggested by the examples in Table 1, they can also usefully be applied within words and particularly between syllables. This is demonstrated by the directed graphs of town name pronunciations in Figures 4, 5 and 6, where the phones of the canonical transcription are shaded:

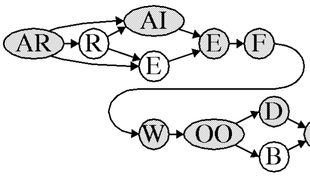

Figure 4: Pronunciation directed graph for R.A.F. Woodbridge

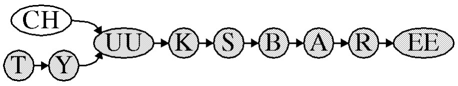

Figure 5: Pronunciation directed graph for Tewkesbary

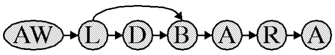

Figure 6: Pronunciation directed graph for Aldeburgh

Baseform transcriptions of the road and town names were checked for places where any of the rules in Table 1 could be applied, and for each of these places, a new transcription was generated. New transcriptions were generated until no further rules could be applied. Figure 4 has a number of places where rules can be applied. To enable rules to cascade, with one rule allowing the application of another, rules are applied from the end of the transcription forwards. This has proven to be most effective. This is illustrated in Table 2.

Baseform:		/AR AI E F	W OO D B R I J/
Assimilation?	✘	/AR AI E F	W OO D B R I J/
	✔	/AR AI E F	W OO B B R I J/
Allophonic Variation?	✘	/AR AI E F	W OO D B R I J/
	✘	/AR AI E F	W OO B B R I J/
	✔	/AR E E F	W OO D B R I J/
	✔	/AR E E F	W OO B B R I J/
Intrusive 'r'?	✘	/AR AI E F	W OO D B R I J/
	✘	/AR AI E F	W OO B B R I J/
	✘	/AR E E F	W OO D B R I J/
	✘	/AR E E F	W OO B B R I J/
	✔	/AR R AI E F	W OO D B R I J/
	✔	/AR R AI E F	W OO B B R I J/
	✔	/AR R E E F	W OO D B R I J/
	✔	/AR R E E F	W OO B B R I J/

Table 2: Sequence of rules applied to "R.A.F. Woodbridge"

The first rule which may be applied is assimilation, which causes the /D/ to sound more like a /B/; this rule is applied (✔) to one copy of the baseform and suppressed (✘) in another. Each of these new transcriptions may then undergo allophonic variation, through which the /AI/ may be pronounced as an /E/; again, for each of the two transcriptions, the rule is applied to one copy and suppressed in another. Finally, the intrusive 'r' rule — /R/ inserted between certain vowels or diphthongs — is applied to one copy of each transcription from the previous stage, and suppressed in another. This yields the eight different transcriptions described by figure 4.

5. NOISE MODELS

Analysis of wrongly recognised utterances revealed that a variety of noise sounds preceded and followed the utterance. The noise sounds included breath noise, clicking noise, mains hum, and a variety of other background noises. To recognise the various noise sounds, the following models were trained:

• BRT	breath noise
• IMP	impulsive noise
• PSN	pre-speech noise
• LIN	line noise
• EXS	extra speech
• OTN	other noise

These models were included into a noise network as illustrated in Figure 6. This better models the various noise sounds that may precede or follow the utterances and hence improves the recognition accuracy.

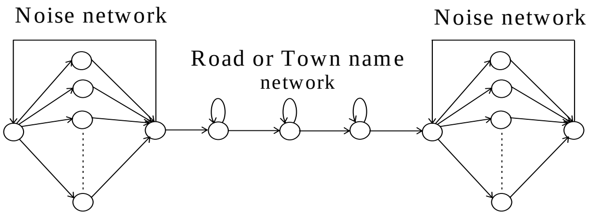

Figure 6: A noise network preceding and following the road or town name.

6. EXPERIMENTS

The effectiveness of the approaches described in this paper was evaluated on a town name task and a road name task comprising 156 Suffolk town names and 1108 road names respectively.

The BT TADS database was used to perform the decision tree clustering. This is a database of telephony-based utterances including town names and road names. The talkers were recruited from regions around the UK to provide speaker-independent data. By using the TADS database during DTC, the resulting set of context dependent models was made application specific. The training data for the models came from both the TADS database and the Subscriber database [8] which contains 4300 sentences.

Baseline performance on the two tasks was obtained using a speech feature comprising MFCCs 1-8 augmented by their velocity and a velocity log energy term, resulting in a 17-D feature vector. Monophone-based 3 state, 12 mode, diagonal covariance HMMs were used to model the speech. This resulted in an accuracy of 61.5% and 64.7% for the town and road name tasks respectively (shown as experiment 1 in table 2).

The lexicon described in section 4 was then added to the baseline set-up. This increased respective recognition accuracy to 76.9% and 75.4% for the two tasks (shown as experiment 2 in table 1).

The context independent monophones were then replaced by the context dependent triphones of section 3 and noise networks of section 5. About 450 triphones were used and were of the same topology as the monophone models. This resulted in accuracies of 83.2% and 79.9% for the two tasks respectively (experiment 3 in table 2).

Finally the RASTA-CTM robust front-end replaced the original 17-D feature. This increased performance to 91.4% and 86.0% respectively (experiment 4 in table 2).

Experiment	Town names	Road names
1. Baseline	61.5 %	64.7 %
2. Baseline + new lexicon	76.9 %	75.4 %
3. MF1 triphones + new lexicon + noise	83.2 %	79.9 %
4. RASTA-CTM triphones + new lexicon + noise	91.4 %	86.0 %

Table 2: Experiment results

Conclusions

The combination of techniques described in this paper have improved the performance of a baseline recognition system significantly. This has enabled the development of an automated corporate directory system with 120,000 entries.

The original system contained no inherent robustness and as such attained only 60% for the two tasks. The four techniques described each attempt to improve robustness by tackling a different problem encountered when dealing with real speech data. These have included background noise and channel distortions, variations in pronunciation, modeling different background noise conditions and taking into account the context within words. Adding this robustness to the system has been shown to dramatically increase performance to around 90% on the two tasks as illustrated in Figure 7.

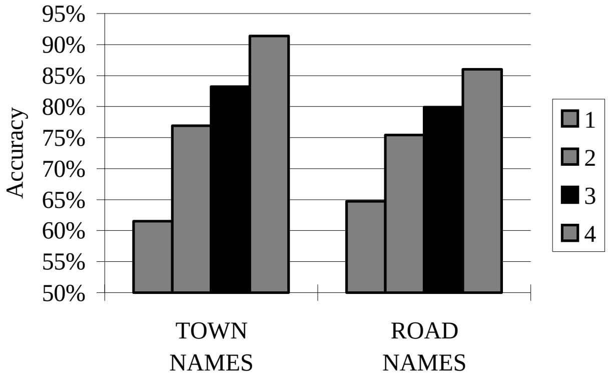

Figure 7: Accuracy for town and road name

REFERENCES

1. Rabiner L. and. Juang B-H, *Fundamentals of speech recognition*, Prentice-Hall, 1993.

2. Milner B., "Inclusion of temporal information into features for speech recognition", *Proc. ICSLP, pp. 256-259, 1996.*

3. Hermansky H. and Morgan N., "RASTA processing of speech", *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 4, pp. 578-589, October 1994.

4. Young S.J, "A review of large vocabulary continuous speech recognition", *IEEE Signal Processing magazine*, Vol. 13, No. 5, pp. 45-57, September 1996.

5. Bahl LR. et al, "Decision trees for phonological rules in continuous speech", *Proc. ICASSP, pp185-188, 1991.*

6. Wiseman, R M and Downey, S N, 'Dynamic and Static Improvements to Lexical Baseforms', pp157–62, *Proc. ESCA Workshop on Modeling Pronunciation Variation for ASR, Rolduc 1998.*

7. Gimson, A C, *An Introduction to the Pronunciation of English*, 4th Edition, Edward Arnold publishers 1989.

8. Simons A. and Edwards K., "Subscriber - A phonetically annotated telephony database", *Proc. IOA, Vol. 14, part 6, pp. 9-16, 1992.*