The Application of Speech Recognition Techniques to Radar Target Doppler Recognition: A Case Study

E. J. Hughes, M. Lewis
Cranfield University at The Defence Academy of the UK
Shrivenham, Swindon, SN6 8LA
Flt., Lt., E. Reid, RAAF
ejhughes@iee.org, m.lewis@computer.org

This paper reports some preliminary results of an examination into the feasibility of recognising the Doppler signatures of targets using speech recognition processing techniques. The rationale is that human operators typically listen to the Doppler audio output from the surveillance radar to detect and possibly identify targets. A feature of speech recognition is that pre-processing is used that takes account of the voice mechanisms that produce speech and the characteristics of the human ear. Three different recognition techniques, with identical pre-processing, were implemented. After validating the recognition algorithms with speech the recognisers were retrained with Doppler signals from a number of sources. It was found that the best of the speech recognisers, HMM-GMM, was also the best of the Doppler recognisers with 88% recognition. The work has been compared with that of others using a similar technique and a good agreement has been found. Some recent discoveries in neuroimaging are quoted that suggest that the human brain and that of several other mammals performs visual recognition in a manner common in speech recognition..

Introduction

Automatic Target Recognition and designation is often seen as a 'holy grail' that allows true 'fire-and-forget' beyond visual range target engagement. However the perceived confidence in practical recognition accuracy is low, although automated systems may be more reliable than human operators in some contexts.

The motivation for the development presented is accounts from surveillance radar operators of being able to distinguish between target types through listening to the radars Doppler audio output. There is anecdotal evidence of experienced operators being able to discriminate between males and females walking within the radar coverage area. Although this claim may be questionable, if a distinction between targets can be made by the human ear, then it can be hypothesised that a suitable speech recognition implementation may be capable of providing the same or superior recognition performance as that of a human operator as well as having the capacity to reduce operator workload. The possibility of distinguishing between target types based on analysis of Doppler audio signals thus warrants investigation.

In this case study, two pieces of research are described and contrasted. The first is results from a MSc., project [1], and the second is a contemporary conference publication [2]. The MSc., thesis compared three methods used for speech recognition, and then used the methods with radar data. The conference publication also addresses the radar data with a very similar technique to the best of those described in the MSc. report.

Speech recognition theory and techniques

The majority of speech signal analysis is performed in the frequency domain and this can be understood when the quantity of discriminating frequency features is examined [3].

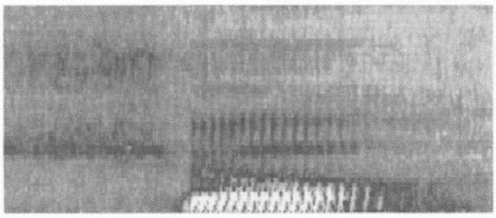

Figure 1. Spectrogram of the word 'shop'

Figure 1 provides a spectrogram of the waveform for the word 'shop'. In the spectrogram the structure of the word can be easily observed, such as the unvoiced fricative at the beginning of the word (sh), changing to the voiced vowel (o) and decaying into the unvoiced plosive (p). The fundamental frequencies and harmonics which constitute each sound are also evident.

Mel-Frequency Cepstral Coefficients

The most commonly utilised feature vector in Speech Recognition is composed of Mel-Frequency Cepstral Coefficients (MFCCs) [4].

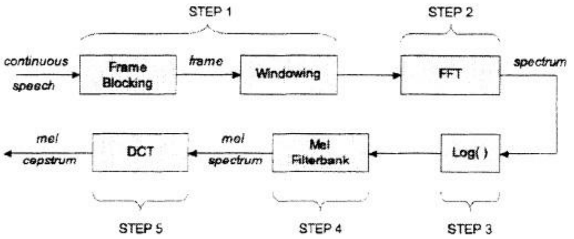

Figure 2. MFCC generation

Figure 2 is a block diagram of the MFCC generation process from a continuous speech source.

Within human speech, there are two methods employed to form words. These sounds are categorized into the voiced and unvoiced. For the voiced part, the vowel sounds, the throat acts like a transfer function. The unvoiced part describes the "noisy" sounds of speech. These are the sounds made with the mouth and tongue, such as "f", "s" and "th" sound.

Mathematically, the two parts are described in the following time domain convolution integral:

$$x(t) = \int_0^t g(t)h(t-\tau)dt$$

By the Convolution Theorem the frequency domain representation of the signal is:

$$X(\omega) = G(\omega)H(\omega)$$

The two frequency components can be separated by taking the logs of both sides:

$$\log|X(\omega)| = \log|G(\omega)| + \log|H(\omega)|$$

Since the two components exist in distinct bands they may now be filtered and optimised for speech content. The mel-frequency scale is applied to obtain an appropriate signal representation since psychophysical studies have shown that human perception of the frequency content of sounds does not follow a linear scale. The recognition model thus reflects the behaviour of the brain in this respect and is equally applicable to both speech and radar Doppler.

Finally the discrete cosine transform (DCT) of the output signal is formed. Application of the DCT approximates to principal components analysis (PCA), which decorrelates the components of the feature vectors and reduces the dimension of a given feature vector by projecting the original feature space onto a smaller subspace [5].

Speech Recognition Methods

To achieve recognition of continuous speech, which must address the issues of speech complexity, variability and ambiguity, the processing topologies typically employed are:

- Dynamic Time Warping (DTW) [7].
- Hidden Markov Models (HMMs) [8] with Gaussian Mixture Models (GMMs) [2]
- Artificial Neural Networks (ANNs).

In order to process normal conversation requires the recognition of 10 to 15 phonemes per second [6].

The ANN used one hidden layer, allowing convex decision boundaries to be created. Both Dynamic Time Warping and HMM's with GMMs could also create non-convex decision boundaries.

Recognition Task

The task of an isolated word recognition problem was chosen in order to test the methods created. In order to limit the scope of the recognition problem, the task of identifying the spoken digits 'zero', 'one', 'two',..., 'nine' as well as 'oh' was chosen. The recognition task is to distinguish between 11 possible digit utterances for different speakers.

Speech Recognition Test Case 1
Sufficiently Trained

The Average Percentage Correctly Recognised Words for the three implementations were:

DTW	90.9
HMM	100
ANN	100

This case represents a typical correctly trained scenario where numerous samples of speaker 1 are used to train the system and a different speaker 1 sample is used as the test sample. In this case four sample sets from the male speaker were used to train the system and the fifth set used to test it.

Speech Recognition Test Case 2
Not Trained

DTW	27.3
HMM	21.8
ANN	17.3

Test Case 2 examined the performance of the methods when presented with training samples from speaker 2 when only speaker 1 data has been used to train the method. This Test Case illustrates the generalisation ability of the recognition method.

In this case the trained networks of Test Case 1 were tested with the one of the sample sets from the female speaker.

Speech Recognition Test Case 3
Minimal Training

DTW	27.3
HMM	79.1
ANN	12.7

This scenario examines the ability to recognise a speaker 2 sample when the system has been trained with predominately speaker 1 information.
The networks were trained using three sample sets from the male speaker plus one sample set from the female speaker and tested with a second sample set from the female speaker.

Speech Recognition Test Case 4
Partial Training

DTW	45.5
HMM	99.1
ANN	22.7

The final Test Case examines the situation where the system has been trained with an equal number of speaker 1 and speaker 2 samples. It should be noted however, that training with only two samples from each speaker is hardly likely to be sufficient for recognition due to the variability between speech utterances made by the same speaker. The testing was performed using one of the male sample sets not used for training.

Speech Recognition Discussion

In the case of the DTW method, the result is uniquely mathematically determined for each test to template match, hence multiple runs for each Test Case are unnecessary as exactly the same computational result is achieved each time.

In the case of the HMM and ANN methods, the performance of each is dictated by how effective the training was. Consequently, for both of these methods, ten iterations for each Test Case were conducted to obtain a limited statistical average of the recognition performance.

Examining Test Case 1 across each of the three methods, it is evident that for this isolated word recognition problem, any of the approaches can provide acceptable recognition performance with a preference for use of either the HMM or ANN methods. Occurrences of Test Case 2 should be avoided.

The HMM, and especially the ANN, implementation rely upon a sufficient quantity as well as variety of test data to be capable of classifying inputs effectively. When the training data is insufficient or not provided as in Test Case 2 2, the techniques perform poorly with the DTW method, surprisingly, displaying the best performance.

In general, the performance of the HMM approach is the most impressive, especially in cases when only a small amount of training data was provided.

As the ANN worked well, it is hypothesised that a convex decision boundary will provide sufficient classification accuracy. It is likely that the ANN generated an approximately linear separation plane when the amount of training data available was limited.

Radar Doppler and Speech Signal Comparison

The radar signals examined are those from a CW X-band radar module that outputs Doppler signals in the audio band.

In addition to the frequency shift due to relative motion between the target and receiver, the reflected wave can be modulated due to periodically moving parts on the target including wheel and track motion for the vehicular targets and arms and legs swinging for the human targets. In addition, target physical features, such as grilles and other items having a periodic structure will also produce modulation of the return. When the modulations are periodic, the spectrum of the received signal will have a line spectrum distributed about the Doppler-shifted transmission frequency. The modulation effect has the benefit of producing potentially distinguishable spectrum fluctuations for each target type encountered.

Implemented Radar Methods

In each of the three methods considered, the target recognition task is treated as an isolated word recognition problem. The assessment was conducted using Doppler audio samples of five-seconds duration, as this time interval appeared to capture most of the signal variation due to target motion during the data collection process.

The feature vector utilised for the recognition task consisted of mel-frequency cepstral coefficients as utilised in the speech recognition implementation.

As a result of conducting a feature vector comparison an assessment of the performance using both the typical 'default' feature vector consisting of 12 MFCCs as well as a more detailed 20 MFCC feature vector was undertaken. Additionally, as the HMM solution provided the best speech recognition results, an extra assessment of this method using a feature vector consisting of 20 MFCCs as well as the 0th order coefficient, energy features, delta and delta-delta features was performed. This assessment was included to examine the benefits of additional feature vector descriptive content as well as to attempt to ascertain the upper limit of recognition performance for the method.

Radar Recognition Test Cases

A wider set of samples was used for the radar tests than had been used for the speech recognition. The sample sets were

- Car approaching
- Car receding
- Tank approaching
- Tank receding
- 1 person approaching, running
- 1 person receding, running
- 1 person approaching, walking
- 1 person receding, walking
- 2 people approaching, walking
- 2 people receding, walking
- 3 people approaching, walking
- 3 people receding, walking
- Clutter

The percentage correct identifications are as follows

Feature Vector	DTW	HMM	ANN
12 MFCCs	38.5	73.8	28.5
20 MFCCs	61.5	88.5	30.8
20 MFCCs + e0dD		87.7	

All Test Cases used 4 five-second sample sets per class for templates or training and 1 one-second sample set per class for testing.

Comparing the overall performance of each method it can be seen that the HMM-GMM approach provides the best recognition result, achieving an 88.5% recognition rate when using a 20 MFCC feature vector. Considering that this has been achieved using only four training samples per class, the result is very encouraging. Examination of the detailed results, however, shows that 8 out of the 13 target classes achieve 100% recognition performance with another two classes having 90% performance. The clutter performance is of concern only achieving a 50% correct recognition result.

The performance of the ANN was rather disappointing, achieving around 30% recognition performance for both the 12 and 20 MFCC feature vectors. The poor results can be attributed to the manner in which the neural network was applied to the recognition problem as opposed to being a weakness of neural network based classification. When processing five-second duration samples, the number of processing frames distributed across the samples must be severely limited in order to restrict the size of the neural network. As a consequence, the spectral detail across the sample was captured using only 10 frames resulting in a very low resolution MFCC representation of each sample. As distinction between the MFCC representations using hundreds of processing frames across each sample is challenging for the DTW and HMM implementations, the ANN results are understandably poor. In order to use the same resolution feature vectors as the other methods, the number of network inputs would be in the order of 5000 for the 12 MFCC case and 8500 for the 20 MFCC case. Training such a network is a painstaking task that it is not considered feasible and the ability of a network to linearly separate the resulting problem space is truly questionable.

No occurrence of ANNs being used to classify long sample durations was encountered in the literature and it can be understood why this is the case. Even when recurrent and time delay neural networks are applied, they are still utilised to process short duration frames of data. The approach performed reasonably well in the speech recognition task due to the short duration utterances enabling a limited number of processing frames to capture sufficient detail for word discrimination.

The results appear curious when the Test Case with extra data is compared with the case without it. It seems that the addition of energy features, the 0th order coefficient, delta and delta-delta features to the feature vector has slightly decreased the recognition performance of the HMM method. Although this is true in an overall sense, examination of the detailed results highlights some desirable benefits of using the extra components. The most important is a 100% classification result for the clutter class meaning no false alarm reports. Furthermore, the only errors (with the exception of a one off tank direction error) are confusion between 2 and 3 people walking targets, the 1 person walking targets have been correctly identified.

Fundamentally, as the dimensionality of the feature vector increases, there is more opportunity to separate the classes, but conversely more training examples are required in order to define the decision boundaries accurately. All transformation processing chains configured using training data as examples require an optimisation process. Optimisation becomes more difficult as the dimensionality of the problem increases, and with few training examples, the error metric for optimisation is inexact. With non-linear processing structures, often the optimisation process is multi-modal, with many locally-optimal decision boundaries possible, and accounts for conflicting classification performance when compared to an alternative architecture or training set.

Alternate Class Definition

The most common area of confusion amongst the methods was distinguishing between the 2 and 3 people walking target classes as a result of walkers falling in and out of step. Although it would be desirable to determine the number of people, it may be acceptable to distinguish between '1 person' and 'multiple people' classes. Consequently, classes such as 2 people walking away and 3 people walking away could be combined into the class 'multiple people walking away'. If the results presented are re-calculated using the multiple people classes the performance changes to that shown in the following table.

Feature Vector	DTW	HMM	ANN
12 MFCCs	61.5	92.3	33.1
20 MFCCs	69.2	95.4	37.7
20 MFCCs + e0dD		99.2	

This example highlights the issues associated with modifying the problem to suit the classification algorithm and it has been long suspected that many ATR publications in the literature have used similar practices to enhance their results. Unfortunately, these enhanced results often do not hold when tested under realistic conditions.

Comparison with other workers

Results from a similar radar study are available in [2]. The approach was to utilise an HMM-GMM architecture with the Gaussian mixture models trained using the reportedly superior Greedy algorithm as opposed to the usual EM algorithm used in this work. The paper states that LPC and cepstrum coefficient feature sets were used, but does not provide any detail.

The reported performance of the implementation for the broad target class headings of 1, 2 and 3 persons, wheeled vehicle, tracked vehicle, clutter and animals was 88% when using the Maximum-Likelihood (ML) decision scheme and 96% for the 'majority voting' decision scheme. Contrasting this performance against the HMM results presented above (88.5%) it is encouraging to note a close agreement with the 88% performance reported using the ML decision scheme, the decision method employed in this work.

The implication is that a superior performance is possible and suggests that the decision boundaries are sensitive to small changes in shape. We hypothesise that the ML method will tend to be drawn to convex decision surfaces, while the non-linear behaviour of the majority voting is more likely to discover more non-convex structures.

The paper also reported the equivalent human operator performance. Once operators were 'trained' with samples from the training database, their recognition performance on the test samples was 37%. Evidently, the approaches presented in [2], and in this paper, are a vast improvement on human classification performance.

Human Brain Action

Recent research in neuroimaging demonstrated that recognition related activity developed significantly earlier in the orbitofrontal cortex (OFC) than in object areas in the visual cortex. Further study showed that this early OFC activity was driven by low spatial frequency (LSF) components in the image [9].

It was proposed that a LSF representation of the input image is projected directly to the OFC and activates information that subsequently sensitizes the representation of the most likely candidate objects in the temporal cortex as an ''initial guess.''

The parallel with techniques where an ANN is used to classify the signals prior to identification by a HMM is obvious. A target return in pulsed radar is the convolution of the transmitted pulse with the target's range profile within the beamwidth. The return will have low frequency components due to bulk features of the target and high frequency components superimposed by target fine detail. The low frequency components may be separated and processed to cue potential models prior to detailed classification.

Conclusion

The target recognition performance achieved via the most successful of the three techniques developed was an 88.5% recognition rate. The method which produced this result was a Hidden Markov Model implementation using Gaussian Mixtures for probability distribution modelling and feature vectors consisting of 20 mel-frequency cepstral coefficients per processing frame.

The work has clearly shown that speech recognition processing techniques can be applied to the recognition of radar targets from their Doppler signatures and additionally, acceptable recognition performance can also be achieved.

References

1. Reid, E., *Application of Speech recognition Techniques to Radar Target Recognition.* MSc. Thesis, RMCS, July, 2005
2. Bilik, I., *et al., Target Classification Using Gaussian Mixture Model for Ground Surveillance Doppler Radar,* IEEE Int., Radar Conf., 2005.
3. Holmes, J.N., *Speech Synthesis and Recognition*, Van Nostrand Reinhold (UK) Co. Ltd., Berkshire, 1988.
4. Antoniou, C., *Modular Neural Networks Exploit Large Acoustic Context Through Broad Class Posteriors For Continuous Speech Recognition*, Proc., Int., Conf., on Acoustic Speech and Signal Processing, 2001.
5. Bischof, H., *et al., Adaptive Combination of PCA and VQ Networks*, Technische Universität Wien, 1996.
6. Scavone, G.P., *Speech Recognition and Synthesis*, http://ccrma.stanford.edu/CCRMA/Courses/150/speech_recognition.html. (14 Aug 06).
7. Harris, J.G., *Isolated Word Speech Recognition Using Dynamic Time Warping Towards Smart Appliances*, at http://www.cnel.ufl.edu/~kkale/dtw.html (2 Apr 05)
8. Cheok, A.D., *et al., HMM Modelling for Audio-Visual Speech Recognition*, IEEE Int., Conf., on Multimedia, 2001.
9. Bar, M., et al., *Top-down facilitation of visual recognition.*, Proc., NAS, Vol., 103, No., 2, pp.449-454