Analysis of Speech Recognition Techniques for use in a Non-Speech Sound Recognition System

Michael Cowling, Member, IEEE and Renate Sitte, Member, IEEE

Griffith University
Faculty of Engineering & Information Technology
Gold Coast, Qld, Australia 9726

Abstract

This paper discusses the use of speech recognition techniques in non-speech sound recognition. It analyses the different techniques used for speech recognition and identifies those that can be used for non-speech sound recognition. It then performs benchmarks on two of these techniques (LVQ and ANN's) and determines which technique is better suited for non-speech sound recognition. As a comparison, it also gives results for the use of these techniques in speech recognition.

ACRONYMS

ANN	Artificial Neural Network
BPA	Back Propagation Algorithm
FFT	Fast Fourier Transform
LVQ	Learning Vector Quantization
HMM	Hidden Markov Models
NN	Neural Network

1. Introduction

It has long been a goal of researchers around the world to build a computer that displays features and characteristics similar to those of human beings. The research of Brooks [1] is an example of developing human-like movement in robots. However, another subset of this research is to develop machines that have the same sensory perception as human beings. This work finds its practical application in the wearable computer domain (e.g. certain cases of deafness where a bionic ear (cochlea implant) cannot be used.)

Human beings use a variety of different senses in order to gather information about the world around them. If we were to list the classic five human senses in order of importance, it is generally accepted that we would come up with the sequence: vision, hearing, touch, smell, taste.

Vision is undoubtedly the most important sense with hearing being the next important and so on. However, despite the fact that hearing is a human beings second most important sense, it is all but ignored when trying to build a computer that has human like senses. The research that has been done into computer hearing revolves around the recognition of speech, with little research done into the recognition of non-speech environmental sounds.

This paper expands upon the research done by the authors [3, 4]. In these papers, a prototype system is described that recognises 12 different environmental sounds (as well as performing direction detection in 7 directions in a 180° radius). This system was implemented using Learning Vector Quantization (LVQ), due to the fact that LVQ is able to produce and modify its classification vectors so that multiple sounds of a similar nature are still considered as separate classes. However, no testing was done to ensure that LVQ was the best method for the implementation of a non-speech sound classification system.

Therefore, this paper will review the various techniques that can be used for non-speech recognition and perform benchmark tests to determine the technique most suited for non-speech sound recognition. Due to lack of research into non-speech classification systems, this paper will focus on using speech and speaker recognition techniques in the domain of environmental non-speech sounds.

The remainder of this paper will be split into four sections. The first section will discuss techniques that have been previously used for speech recognition and identify those techniques that could also be applied to non-speech recognition. The second section will show the results of benchmarks on two of these techniques (LVQ and ANN) and also give comparative results for speech recognition. The third section of this paper will discuss these results. Finally, the fourth section will conclude and suggest areas for future research.

2. Sound Type Comparison

Research into speech recognition began by reviewing the literature and finding techniques that had previously been used for speech/speaker recognition. It was found that six techniques are commonly used for speech/speaker recognition or have been used for this domain in the past. These were:

- Dynamic Time Warping (DTW)
- Hidden Markov Models (HMM)
- Vector Quantization (VQ)
- Ergodic-HMM's
- Artificial Neural Networks (ANN)
- Long-Term Statistics

Comparison tables were then built (using [5, 7, 8, 9]) that compared the different feature extraction and classification methods used by each of these six techniques. These tables are presented on the following page as Table 6 and Table 7.

Looking at these comparison tables, we can begin to examine whether any of these speech recognition or speaker identification techniques can be used for non-speech sound recognition.

From looking at the comparison tables, it appears that some of these techniques, by their very nature, cannot be used for non-speech sound recognition. Any of the techniques that use subword features will not be able to be used for non-speech sound identification. This is because environmental sounds lack the phonetic structure that speech does. There is no set "alphabet" that certain slices of non-speech sound can be split into, and therefore subword features (and the related techniques) cannot be used.

Due to the lack of an environmental sound alphabet, all of the Hidden Markov Model (HMM) based techniques that are shown in the table cannot be used. Since HMM techniques are the main techniques now used in speech and speaker recognition, this leaves only a few other techniques.

After discounting HMM, the remaining five techniques were tested for their ability to classify non-speech sounds. This was done in two ways. First, benchmarking is performed, using these techniques, on non-speech sounds and data on the parameters, the resulting time taken and the final correct classification rate is recorded. Then, these results are compared with statistics and benchmark results reported in the literature for the performance of these techniques on speech. This demonstrates how these techniques perform against each other on speech and provide a comparison to the results for non-speech.

The five techniques left to be tested are:

- Dynamic Time Warping
- Long-Term Statistics
- Vector Quantization
- Artificial Neural Networks
- Other Statistical Features

Based on this information, in this paper we make a comparison between Vector Quantization and Artificial Neural Networks as possible techniques for non-speech sound recognition. As an initial test, eight sounds were used, each with six different samples. Data set size was kept as small as possible due to the time it takes to train larger data sets. The sounds used for this test are detailed in Table 1 and are some typical sounds that would be classified in a sound surveillance system.

TABLE 1. SOUNDS USED IN EXPERIMENTS.

Sound Type
Jangling Keys
Footsteps (Close)
Footsteps (Distant)
Wood Snapping
Coins Dropping
Footsteps on Leaves
Glass Breaking
Footsteps on Glass

These techniques will be tested using a jackknife method, identical to that used by Goldhor [6]. A jackknife testing procedure involves training the network with all of the data except the sound that will be tested. This sound is then tested against the network and the classification will be recorded. This jackknife procedure will be repeated with all six of the samples from each of the eight sounds.

3. Results

3.1 Vector Quantization

An implementation of VQ was tested in Matlab using the inbuilt Learning Vector Quantization (LVQ) routines. The LVQ network was trained using a simple statistical feature, being the frequency components of the signal (gathered by performing a Fast Fourier Transform (FFT)). Since non-speech sound covers a wider frequency range than speech (anywhere from 0Hz to 20,050Hz (the approximate limit of human hearing)), a 44,100 point FFT was performed and the results (22,050 unique features) were used to train the LVQ network. The network produced results as shown in Table 2.

TABLE 2. LVQ RESULTS - NON-SPEECH SOUND RECOGNITION.

Training Iterations (Epochs)	Training Time (Avg.)	Percent Correct
150	35 seconds	41%
300	42 seconds	58%
450	64 seconds	63%

This is in comparison to statistics for the use of LVQ for speech recognition. Results from Van de Wouver e.a. [10] are shown in Table 3 for both female and male voices. These results present statistics for both a standard LVQ implementation for speech recognition and an implementation of LVQ that then has fuzzy logic performed on it (FILVQ). As can be seen from the results, the use of LVQ for speech recognition produces quite low recognition results.

TABLE 3. LVQ RESULTS - SPEECH RECOGNITION [10].

Method	Female	Male
Standard LVQ	36%	29%
FI-LVQ	60%	64%

3.2 Artificial Neural Networks

An implementation of a Multi-Layer Perception (MLP) was tested in Matlab using the popular Back Propagation technique. As with LVQ, the network was trained using a simple statistical feature, being the frequency components of the signal (gathered by performing a FFT). Since non-speech sound covers a higher frequency range than speech (anywhere from 0Hz to 20,050Hz (the approximate limit of human hearing)), a 44,100 point FFT was performed and the results (22,050 unique features) were used to train the multi-layer perceptron. The network produced results as shown in Table 4.

TABLE 4. ANN RESULTS - NON-SPEECH SOUND RECOGNITION.

Training Iterations (Epochs)	Training Time (Avg.)	Percent Correct
150	145 seconds	8%
300	324 seconds	8%
450	483 seconds	13%

This is in comparison to results from speech recognition using an ANN. Results from Castro and Perez [2] are shown in Table 5 below. Their results were taken on a difficult isolated word recognition set, the Spanish EE-set. The MLP tested used the back propagation algorithm, contained 20 hidden neurons and was trained over 2000 epochs with various amounts of inputs. The figures given are the MLP's estimated error rate with a 95% confidence interval.

TABLE 5. ANN RESULTS - SPEECH RECOGNITION [2].

Number of Inputs	Error Rate
550 Inputs	19.7 (+ 2.9) %
220 Inputs	16.3 (+ 2.4) %

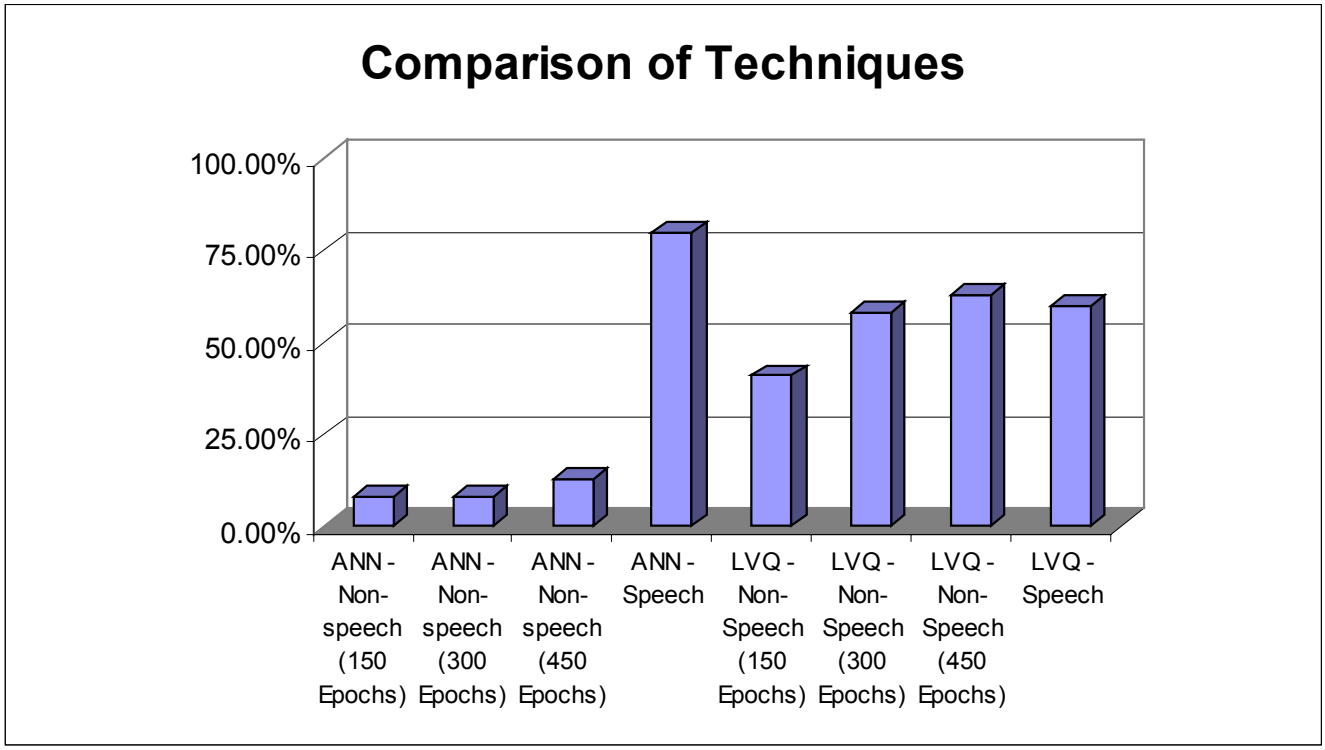

Figure 1. Comparison of Sound Classification Techniques.

TABLE 6: SPEECH RECOGNITION

Technique	Sub Technique	Relevant Variable(s) / Data Structures	Input	Output
Sound Sampling	ALL	Analog Sound Signal	Analog Sound Signal	Digital Sound Samples
Feature Extraction	Dynamic Time Warping (DTW)	Statistical Features (e.g. LPC coefficients)	Digital Sound Samples	Acoustic Sequence Templates
	Hidden Markov Models (HMM)	Subword Features (e.g. phonemes)	Digital Sound Samples	Subword Features (e.g. phonemes)
	Artificial Neural Networks (ANN)	Statistical Features (e.g. LPC coefficients)	Digital Sound Samples	Statistical Features (e.g. LPC coefficients)
Training and Testing	Dynamic Time Warping (DTW)	Reference Model Database	Acoustic Sequence Templates	Comparison Score
	Hidden Markov Models (HMM)	Markov Chain	Subword Features (e.g. phonemes)	Comparison Score
	Artificial Neural Networks (ANN)	Neural Network with Weights	Statistical Features (e.g. LPC coefficients)	Positive/Negative Output

TABLE 7: SPEAKER RECOGNITION

Technique	Sub Technique	Relevant Variable(s) / Data Structures	Input	Output
Sound Sampling	ALL	Analog Sound Signal	Analog Sound Signal	Digital Sound Samples
Feature Extraction	Dynamic Time Warping (DTW)	Statistical Features (e.g. LPC coefficients)	Digital Sound Samples	Acoustic Sequence Templates
	Hidden Markov Models (HMM)	Subword Features (e.g. phonemes)	Digital Sound Samples	Subword Features (e.g. phonemes)
	Vector Quantization (VQ)	Statistical Features (e.g. LPC coefficients)	Digital Sound Samples	Statistical Features (e.g. LPC coefficients)
	Ergodic-HMM's	Subword Features (e.g. phonemes)	Digital Sound Samples	Subword Features (e.g. phonemes)
	Artificial Neural Networks (ANN)	Statistical Features (e.g. LPC coefficients)	Digital Sound Samples	Statistical Features (e.g. LPC coefficients)
	Long-Term Statistics	Statisical Features (Mean and Variance)	Digital Sound Samples	Statisical Features (Mean and Variance)
Training and Testing	Dynamic Time Warping (DTW)	Reference Model Database	Acoustic Sequence Templates	Comparison Score
	Hidden Markov Models (HMM)	Markov Chain	Subword Features (e.g. phonemes)	Comparison Score
	Vector Quantization (VQ)	VQ Network & Codebooks	Statistical Features (e.g. LPC coefficients)	Distortion Value
	Ergodic-HMM's	Markov Chain	Subword Features (e.g. phonemes)	Comparison Score
	Artificial Neural Networks (ANN)	Neural Network with Weights	Statistical Features (e.g. LPC coefficients)	Positive/Negative Output
	Long-Term Statistics	Reference Model Database	Statisical Features (Mean and Variance)	Comparison Score

N.B. All techniques begin with an initial input of an analog sound wave.

4. Discussion

The results obtained are somewhat surprising. Even though the results from speech recognition suggest that the ANN will outperform LVQ, the opposite occurs for non-speech recognition. We propose that this is due to the closeness of the various environmental sounds presented to the two networks.

It is accepted that one of the main advantages of LVQ over ANN's is its ability to correctly classify results even where classes are similar. In this case, sounds such as footsteps (close) and footsteps (distant) appear the same but contain slightly lower or higher amplitudes. LVQ is able to classify these sounds properly where the ANN gets confused. Indeed, the detailed results of each test show that the ANN was classifying footsteps (close) as footsteps (distant) and vice versa. To prove this hypothesis, further tests were performed on the ANN using several higher epoch values (to allow more training time). The results are presented in Table 8 below.

TABLE 8. FURTHER ANN RESULTS - SPEECH RECOGNITION.

Training Iterations (Epochs)	Training Time (Avg.)	Percent Correct
200	184 seconds	0%
800	742 seconds	1%
1200	1087 seconds	13%

From these results, it can be seen that the ANN results remain the same regardless of the epoch value. This suggests that the ANN has problems training the sample sounds, most likely due to these sounds being non-linearly separable.

From the results, it can also be seen that an LVQ network seems to have an inherent ability to classify quicker than a corresponding ANN. We believe this is due to the algorithms used within each neuron to classify the input vectors. Further tests would be able to more adequately demonstrate the speed vs. classification accuracy of LVQ vs ANN's.

5. Conclusion

This paper has presented benchmarks for the use of both LVQ and ANN for classification of non-speech sounds. It has shown that, for sounds in close proximity to each other, an LVQ network will outperform an ANN. It has also shown that there is an inherent speed advantage in the use of LVQ for non-speech sound recognition.

Further research will endeavor to produce more benchmarks on the use of LVQ and ANN's for non-speech sound identification. Systematic comparisons will also be made of other classification techniques from speech recognition that we believe could be used for non-speech sound classification.

6. References

[1] R, Brooks; C, Breazeal; M, Marjanovic; B, Scassellati; M, Williamson "The Cog Project: Building a Humanoid Robot." C. Nehaniv, ed., Computation for Metaphors, Analogy and Agents, Vol. 1562 of Springer Lecture Notes in Artificial Intelligence, Springer-Verlag, 1998

[2] M. J. Castro; J. C. Perez, "Comparison of Geometric, Connectionist and Structural Techniques on a Difficult Isolated Word Recognition Task.", Proceedings of European Conference on Speech Comm. and Tech., ESCA, Vol. 3, pp 1599-1602, Berlin, Germany, 1993.

[3] M. Cowling, R. Sitte, "Sound Identification and Direction Detection in Matlab for Surveillance Applications", Proceedings of Matlab Users Conference, Melbourne, Australia, November, 2000.

[4] M. Cowling, R. Sitte, "Sound Identification and Direction Detection for Surveillance Applications", Proceedings of ICICS 2001, Singapore, October, 2001.

[5] B. Gold; N. Morgan, Speech and Audio Signal Processing, John Wiley & Sons, Inc, 2000, New York, NY.

[6] R. S. Goldhor, "Recognition of Environmental Sounds" Proceedings of ICASSP Vol. 1, pp 149 – p152, New York, NY, USA, April 1993

[7] C. H. Lee; F. K. Soong; K. Paliwal "An Overview of Automatic Speech Recognition", Automatic Speech and Speaker Recognition: Advanced Topics. Kluwer Academic Publishers, 1996, Norwell, MA.

[8] C. H. Lee; F. K. Soong; K. Paliwal "An Overview of Speaker Recognition Technology", Automatic Speech and Speaker Recognition: Advanced Topics. Kluwer Academic Publishers, 1996, Norwell, MA.

[9] R. Rodman, Computer Speech Technology. Artech House, Inc. 1999, Norwood, MA 02062.

[10] G. Van de Wouver; P. Scheunders; D. Van Dyck, "Wavelet-FILVQ Classifier for Speech Analysis", Proc. Int. Conference Pattern Recognition pp. 214-218, Vienna, 1996.