LIP READING FROM THERMAL CAMERAS

STEVEN JAMES ANDERSON

A thesis submitted to Auckland University of Technology in partial fulfilment of the requirements for the degree of

Master of Computing and Information Sciences (MCIS)

School of Computing and Mathematical Sciences

2012

ABSTRACT

A constructive research methodology has been used to explore the use of thermal images for improving Automatic Speech Recognition (ASR) performance. Previous research has shown that the addition of a visual modality for speech recognition improves ASR performance in both clean and noisy environments. However, Audio-Visual Automatic Speech Recognition (AVASR) performance can be greatly affected by changing lighting conditions. Conversely, thermal cameras are highly invariant to changes in lighting conditions as such the use of thermal video may be beneficial to AVASR.

An AVASR system was created for testing the effect of adding a third modality to AVASR. Mel-frequency Cepstral Coefficient (MFCC) based speech recognition was used for audio speech recognition. For visual recognition, the standard video and thermal video were processed using a method derived from Wai Chee Yau’s (2008) proposed Motion Templates method of feature extraction.

A custom audio visual database was created for this project. Eleven participants were recorded in audio, standard video and thermal video repeating ten words each (the numbers zero through nine) fifteen times to create a database of 1650 words for testing. Testing was completed using a speaker dependent, isolated word recognition system. For each participant 14 samples of each word were used for training Hidden Markov Models (HMM) with the remaining sample used for testing with Gaussian white noise added to the audio signal at 20, 10, 0, -10 and -20 decibel signal to noise ratios (SNR). This test was repeated five times with different samples selected for each test and the results averaged to reduce sample bias.

It was successfully shown that combining audio, standard, and thermal video for ASR can improve performance by increasing recognition rates a relative 11.8% over audio and standard video combined ASR and a relative 38.2% over audio only ASR when averaged over all noise levels.

CONTENTS

ATTESTATION OF AUTHORSHIP

"I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person (except where explicitly defined in the acknowledgements), nor material which to a substantial extent has been submitted for the award of any other degree or diploma of a university or other institution of higher learning."

Steven Anderson ________________________________

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

AV – Audio-Visual

ASR - Automatic Speech Recognition

AVASR - Audio-Visual Automatic Speech Recognition

DCT - Discrete Cosine Transform

DFT – Discrete Fourier Transform

DOF - Difference of Frame

FFT - Fast Fourier Transform

FPS – Frames Per Second

HMM - Hidden Markov Model

IR – Infra Red

MFCC - Mel-frequency Cepstral Coefficient

MT – Motion Template

ROI – Region of Interest

SNR – Signal to Noise Ratio

1. INTRODUCTION

Recently in New Zealand there has been much controversy over the funding of wages for transcription services for New Zealand's first deaf MP Mojo Mathers. Due to Ms Mathers hearing difficulties, live captioning of parliamentary debates was required for her to follow the discussions (Romanos, 2012). Imagine if such services could be performed automatically. Not only for the ability to automatically transcribe meetings, improved speech recognition performance could benefit society in a large number of ways and open the door for a wide range of consumer applications and new human computer interaction techniques: live captioning of the news, automatic translation, improved accessibility for the disabled, robotics, replacement of television remotes, military aircraft, vehicle accessories, such as navigation systems, and could be used in a wide range of consumer products.

Automatic Speech recognition has long been a goal of computer science research. In 1950 Davis, Biddulph, and Balashek (Davis, Biddulph, & Balashek, 1950) of Bell Laboratories built a system for isolated digit recognition for a single speaker recognising each of the ten digits, one through nine and oh for determining identity of an unknown spoken digit (Juang and Lawrence, 2004; Davis, Biddulph, & Balashek, 1950). Since then much research has been devoted to ASR and now it is in common usage though out our daily lives with the proliferation of Bluetooth devices and mobile phones. However, the reliance on sound alone makes speech recognition unsuitable for use in noisy environments and even in clean environments state-of-the art ASR systems perform well below that of humans (Potamianos, Neti, Gravier, Garg, & Senior, 2003). As such the use of ASR is more often than not the source of frustration than it is a convenience.

It is well known that combining audio and visual analysis improves speech recognition accuracy in both noisy and clean environments. According to G. Potamianos, Neti, Gravier, Garg, & Senior, 2003) the visual modality benefit to speech intelligibility in noisy environments has been quantified by as far back as 1954 (Sumby & Pollack). A famous example of the visual modality's benefit to speech is the *McGurk Effect.* This is where sound superimposed over the video of a speaker reciting a different sound can result in an observer hearing a third sound. For example when the sound played is /ga/ and the video is of a person saying /ba/, most observers will identify the word as /da/. Audio Visual Automatic Speech Recognition (AVASR) attempts to use the known benefit of a visual channel as a means to improve ASR.

The first AVASR system was developed by Petajan (1984) who showed that addition of the visual modality can improve the performance of ASR (Potamianos, et al., 2003). Since then, a large amount of research has been conducted into AVASR and the majority of research projects have shown improvements to audio only ASR in a variety of conditions. However, these systems are still highly affected by a number of factors that make them impracticable for real world use. One of the biggest problems encountered changes in lighting conditions, which greatly reduces accuracy making AVASR unsuitable for use in uncontrolled environments.

This thesis investigates the use of adding a thermal camera as a third modality to AVASR. The proposed advantage of using a thermal image camera is that they are insensitive to changes in lighting levels and immune to the differences in skin tone that can make lip-reading difficult. It is hoped that this research can show that these properties of thermal cameras can be utilized to improve the performance of AVASR.

1.1 RESEARCH OBJECTIVE

The aim of this research is to test the use of thermal video for lip-reading with the aim of improving ASR performance. This is because thermal cameras offer two main advantages over standard cameras: thermal cameras detect light in the emissive portion of the infrared spectrum are thermal images are highly invariant to changes in lighting conditions and also as humans emit infrared light whereas most inanimate objects do not it is easier to separate people from background clutter. To complete this research a solution as to how to lip-read from thermal video needed to be constructed.

An AVSR system was created to test all the combinations of the three modalities: thermal video alone; standard video alone; audio alone; thermal video and audio combined; standard video and audio combined; standard video and thermal video combined; standard video, thermal video and audio video combined with the hypothesis that the addition of thermal video to AVASR will lead to an improvement of ASR. The AVASR system was tested in and office environment, with the intention that more challenging environments such as in car speech recognition be tested in later research. Tests were taken on recognition rates with the hypothesized results to be (in order of best to worst):

1. Audio, standard video, and thermal video combined
2. Audio and standard video combined
3. Audio and thermal video combined
4. Audio
5. Standard video and thermal video combined
6. Standard video
7. Thermal video

1.2 Structure of the Thesis

This thesis is split into six different chapters. The first chapter gives an introduction to the project and describes its goal. The second chapter describes the methodology used in this research. The third chapter is the literature review and gives an overview of the differences between thermal cameras and visible light cameras, the performance of thermal images in face recognition, and the basic steps used in creating an AVASR system. The fourth chapter covers the creation of the AV database for this project and covers the criteria, equipment, and the collection for creating the audio-visual database used in this project. The fifth chapter covers the audio and visual system data extraction and HMM classification techniques used in the AVASR system as well as test results. The sixth chapter is the conclusion and discusses the limitations of this work and further research to be done. The appendix contains the individual results for speakers.

2. Methodology

This chapter discusses the research methodologies used in this thesis and is split into three parts. The first section gives a brief overview of the two main paradigms of research: positivist and interpretivist research. The second section discusses the constructive methodology, and the third section discusses the test instrument used in this research.

2.1 Positivist and Interpretivist Research

Research may fall into two main paradigms: *Positivist* and *Interpretivist.* Positivist research is intended to produce an exact unbiased representation of reality. Findings should be replicable by other researchers and is equated with the scientific method, where knowledge is discovered by controlled means, such as experiments and studies are frequently hypothesis based. Conversely, interpretivist research aims to find new interpretations or underlying meanings and adheres to the ontological assumption of multiples realities, which are time-and context dependent. Interpretive research is value-related as interpretivism leads to subjective findings which may differ between researchers and is an appropriate view for studies of complex human behaviour and social phenomena. (de Villiers, 2005).

Positivism primarily relies on quantitative methods, where data comprises mainly numbers and measurements and analysis is done using statistical methods, whereas interpretivism mainly uses qualitative studies. However, the two are not mutually exclusive and often studies require a mixture of the two. Qualitative research can often be used as exploratory research and can be used to formulate hypothesis and questions, setting the foundation for quantitative research. (de Villiers, 2005)

2.2 Constructive Methodology

A constructive approach was taken towards this research. According to Crnkovic (2005) the key idea of constructive research is the construction, based on the existing knowledge used in novel ways, with possibly adding a few missing links. The constructive research method implies building of an artefact that solves a domain specific problem in order to create knowledge about how the problem can be solved. The methodology of engineering and Software Engineering as a research field is fundamentally constructive with most Software Engineering research predominantly constructive and inventing new models and tools. Due to the scope of this research and the problem of arriving at a solution as to how lip reading from thermal cameras can be achieved, a constructive research approach was considered appropriate.

Kasanen, Lukka, and Siitonen (1993) describe the constructive method as requiring four additional components above constructing a solution to a problem to ensure a contribution is valid research and it is not just an exercise in problem solving. These are shown in Figure 1.

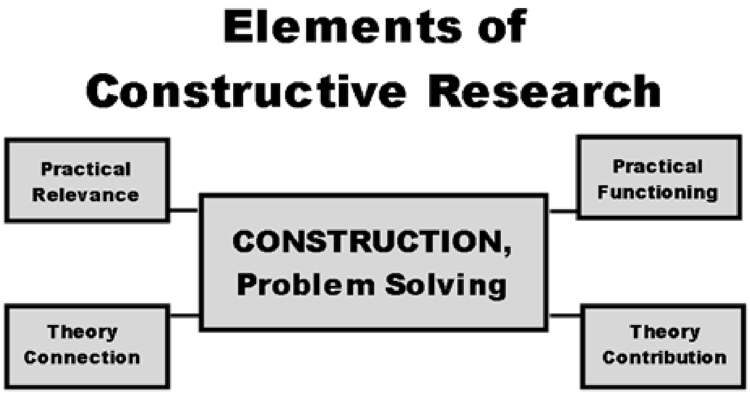


FIGURE 1 ELEMENTS OF CONSTRUCTIVE RESEARCH. ADAPTED FROM KASANEN ET AL (1993)

As such, this project needs to not only construct a solution to the problem of lip-reading from thermal video, but also to have practical relevance, practical functioning, a connection to theory, and a contribution to theory. The aim of this project is to improve ASR provides the practical relevance. Practical functioning is shown through the construction of an AVASR system incorporating thermal video. A connection to theory will be used in the construction of the solution. The contribution to theory will be completed during the construction of the solution.

The steps in constructive research as outlined by Kasanen et al (1993) are:

1. Find a practically relevant problem which also has research potential.
2. Obtain a general and comprehensive understanding of the topic.
3. Innovate, i.e., construct a solution idea.
4. Demonstrate that the solution works.
5. Show the theoretical connections and the research contribution of the solution concept.
6. Examine the scope of applicability of the solution.

The constructive approach is iterative in nature and steps 3 through 6 can be repeated with different solutions found.

For the first step of the constructive approach, the problem is how adding a third modality of lip-reading from thermal video to AVASR can be achieved. The second step will be completed with a literature review. For the third step, construct a solution, an ASR system was constructed for all the different combinations for the three modalities: thermal video, standard video, and audio. The fourth step, to demonstrate that the solution works, was completed by testing the solution and comparing recognition rates for the various combinations. The fifth step of showing theoretical connections to the solution as well as research contributions in the solution was completed during the construction of the solution and the analysis of the results. The sixth step of examining the scope of the solution will be completed by both critiquing the choices made in construction of a solution and the results.

The nature of constructive research is interpretivist in nature. Constructive research emphasises the fact that scientific knowledge is constructed by with help of cognitive tools. It is the opposite of the positivism where scientific facts are discovered and the connection between the world and the fact is unique. Constructive research requires acknowledging that multiple solutions to a problem may be arrived at and there is no guarantee for a consensus. (Shaw, 2001).

2.3 DEMONSTRATING THE SOLUTION WORKS

An essential component of constructive research is demonstrating that the solution works as this can wither confirm or deny the validity of the solution arrived at. There are a number of techniques that can be used to validate software engineering research, as listed by Shaw (2001) in Table 1 below:

Technique	Character of validation
Persuasion Technique Design Example	I have thought hard about this, and I believe that… …if you do it the following way, then… …a system constructed like this would… …walking through this example shows how my idea works
Implementation System Technique	Here is a prototype of a system that... ...exists in code or other concrete form ...is represented as a set of procedures
Evaluation Descriptive Model Qualitative Model Empirical quantitative model	Given these criteria, here's how an object rates... ...in a comparison of many objects ...by making subjective judgments against a checklist ...by counting or measuring something
Analysis Analytic formal model Empirical predictive model	Given the facts, these consequences… ...are rigorous, usually symbolic, in the form of derivation and proof ...are predicted by the model in a controlled situation (usually with statistical analysis)
Experience Qualitative or descriptive model Decision criteria Empirical predictive model	I evaluate these results based on my experience and observations about the use of the result in actual practice and report my conclusions in the form of... …prose narrative …comparison of systems in actual use …data on use in practice, usually with statistical analysis

TABLE 1 TECHNIQUES USED IN VALIDATION OF RESEARCH. ADAPTED FROM SHAW(2001).

While persuasion was used in the literature review and discussion of system construction, the main method of demonstration was chosen to be an empirical quantitative model in the form of an experiment where the testing was unbiased and repeatable. The measurement was chosen to be the word recognition rates, which are a common measurement in ASR systems. To achieve the system was required to be implemented and exist in code and a data collection of participants filmed reciting words while recorded in thermal video, standard video and audio was required to test and demonstrate the solution.

While the research involves constructing a solution as to how to lip-read from AVASR, the purpose of doing so is to improve ASR performance. Therefore it is not sufficient to simply show the addition of thermal video to AVASR is possible but to also demonstrate an improvement over audio only ASR and AVSR. Therefore the results between the three systems need to be compared.

2.4 Section Summary

A constructive research approach was used in this research. As such, the main objective of this project was to construct a solution as to how thermal video can be used for lip reading and show that the addition of thermal video to AVASR can be used to improve ASR performance. The contribution to theory was made during the construction of the solution and the testing of the solution to verify the validity of the theories used.

3. LITERATURE REVIEW

Little research has been undertaken into the use of thermal video for lip reading. Searches of the IEEE, ACM, Springer-Link, Scopus and Science Direct online databases, as well as the Google Scholar and Microsoft Academic search engines only uncovered one research paper investigating the use of thermal images for lip-reading Saitoh and Konoshi (2006). While this research gave low recognition rates of 44% for thermal video and only a small improvement in recognition rates for combined thermal video and standard video 80% compared to 76% for standard video alone, this research did not explore the effects of combining the signal with audio.

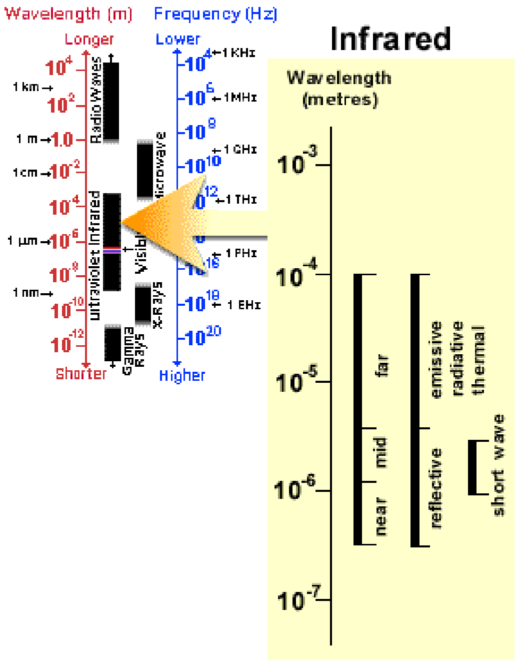
Huang, Potamianos, and Neti (2004) developed a headset with an infrared for use in ASR. While this processes images that use light in the infrared spectrum, it is in a different part of the infrared spectrum to thermal light where light is emitted rather than reflected. This means that under variable lighting conditions the system will be affected by the same problems that plague standard cameras.

The reason so little research has been conducted into using thermal video for lip-reading is likely due to the same reasons Socolinsky and Selinger (2002) gave for the lack of research into using thermal images for face recognition. Those reasons are: much higher cost of thermal sensors versus visible video equipment, lower image resolution, higher image noise, and lack of widely available data sets.

As little research has been conducted into the use of thermal imagery for automatic speech recognition, relevant research into face recognition is also included in this literature review, which is split into three sections. The first section discusses the differences between thermal imaging cameras and visible light cameras. The second section investigates research into thermal imaging cameras and visible light cameras for use in facial recognition. The third section looks at various methods to explore different methods used in the construction of audio visual speech recognition systems.

3.1 Comparing Thermal Imaging Cameras with Visual Spectrum Cameras

The difference between thermal cameras and standard cameras is that standard cameras measure light in the visible spectrum range (0.4 µm –0.7 l µm in wavelength), whereas thermal cameras respond to radiation in the thermal ranges (medium-wave infrared and long wave infrared). Light in the visible spectrum, near infra-red 0.7 µm –0.9 l µm, short-wave infrared, (0.9 µm –2.4 l µm) ranges is largely reflected. However, in the medium-wave infrared (3.0 µm –5.0 µm) and long-wave infrared ranges (8.0µm –14.0µm), light is largely emitted. This is shown in Figure 2.



© CCRS / CCT

FIGURE 2 THE INFRA RED SPECTRUM ("THE ELECTROMAGNETIC SPECTRUM," 2008)

Because light in the thermal imaging ranges is largely emitted and not reflected, images captured using thermal IR sensors are nearly invariant to changes in ambient illumination and are less subject to scattering and absorption by smoke or dust. The amount of thermal radiation emitted by an object depends on the temperature and emissivity of the material. The human body emits thermal radiation in both the medium-wave infrared range and the long-wave infrared range with more thermal radiation emitted in the long-wave infrared range than in the medium-wave infrared range, making long-wave infrared camera more suitable for measurements. (Kong, Heo, Abidi, Paik, & Abidi, 2005).

The fact that thermal imaging cameras detecting emissive light and not reflective light makes them suitable for situations where lighting conditions are variable, and they also perform well where subjects need to be separated from background images as many objects do not emit thermal radiation. Examples of this are shown in Figures 3 and 4.

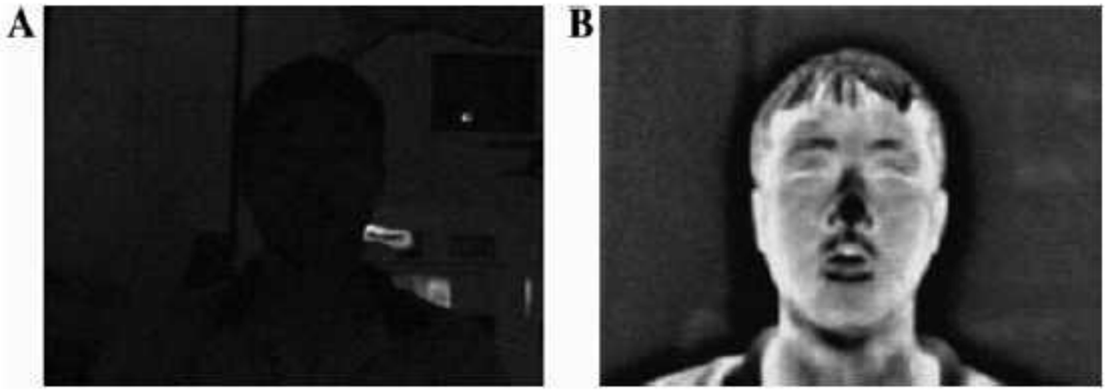


Figure 3 Comparison of images taken under low lighting. (A) Visual image. (B) Corresponding thermal image (Kong, et al., 2005).

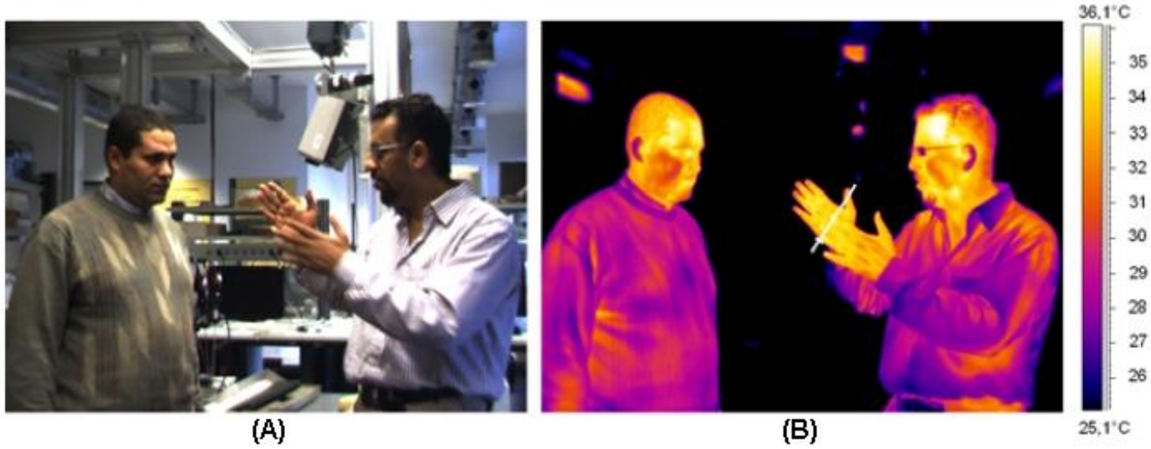


Figure 4 Comparison of images taken with a cluttered background. (A) Visual image. (B) Corresponding thermal image (Appenrodt, Al-Hamadi, Elmezain, & Michaelis, 2009)

However, while thermal images have advantages due to their invariance to light, they also come with disadvantages. Edges are more blurred, the lips are in many cases undistinguishable and therefore the mouth is hardly distinguishable if it is closed. While the face aspect is not affected by illumination changes, the images can change because of the heat in the room. However, these changes are subtle and not comparable to those introduced by variations in lighting in the case of colour images (Martinez, Binefa, & Pantic, 2010). Another disadvantage of thermal imagery is that thermal infrared radiation cannot penetrate glass (Morris, Avidan, Matusik, & Pfister, 2007).

Martinez et al’s (2010) description of the mouth in thermal images indicates that there are difficulties in detecting the mouth in thermal images, which may indicate difficulties in lip-reading from thermal images: “When the mouth is closed, the problem becomes very challenging since the mouth is sometimes barely distinguishable even to the naked eye.” (p.49). However, they go on to state, “there is a big difference in the temperature inside and outside the mouth … especially for the case of expressive faces, and the fact that the interior of the mouth can appear as hotter or colder than the skin. Furthermore, teeth, tongue, and the rest of the interior of the mouth can present different temperatures as well.” (p.51). Therefore there may be some difficulty locating the lips in thermal images negatively impact on tracking movement for lip-reading. However, this may be offset by the temperature variations as there are large variations in the temperature of the mouth. This suggests that instead of tracking lip movements in thermal images, a better approach would be to track changes in mouth temperatures by looking at changes of pixel values in the vicinity of the mouth.

3.2 Thermal Images for Face Recognition

Face recognition faces suffers many of the same issues as speech recognition with visible light and variance in lighting conditions known to be one of the major factors limiting performance. However, numerous studies have shown that face recognition using thermal imagery is more reliable than face recognition in the visible light spectrum under variable lighting conditions.

Socolinsky & Selinger (2004) conducted a comparison on the visible light images and thermal light images, and fused thermal visible light images in both controlled and uncontrolled environments. While visible light images performed better in controlled environments, it was found that thermal images produced superior results for outdoor environments. However, it was found that the fusing of visible light images and thermal images for face recognition produced the best results for both indoor controlled lighting environments and in outdoor uncontrolled environments; with results for outdoor environments nearing that of indoor environments. While face recognition using thermal images outperformed face recognition in variable lighting conditions, thermal images were affected by fluctuations in skin temperature, which means that when used in outdoor environments, performance decreases. This is because, according to Socolinsky & Selinger (2004), when someone is exposed to cold or wind, capillary vessels at the surface of the skin contract, and reduce the blood flow to the face, therefore reducing the skin temperature. The opposite effect can be seen when transitioning from cold to warm environments, with capillary vessels at the surface of the skin expanding and increasing the skin temperature. Additional factors such as physical exercise can also cause fluctuations in skin temperature. On top is this, breathing can cause large changes in temperature as the mouth and nose become cooler when inhaling, and warmer when exhaling. These variations can be seen in Figure 5.

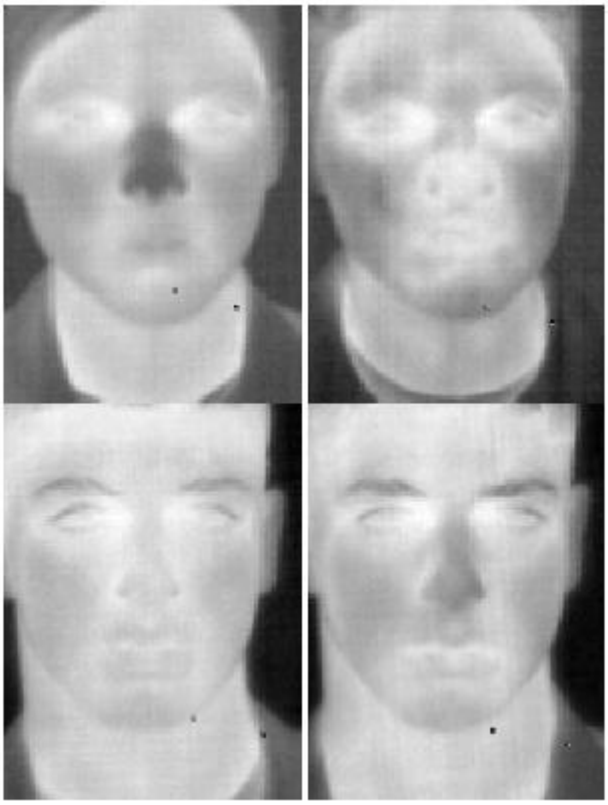


FIGURE 5 DIFFERENCES DUE TO VARIATIONS IN SKIN TEMPERATURE (SOCOLINSKY & SELINGER, 2004).

Hermosilla, G., Ruiz-del-Solar, Verschae & Correa (2006) conducted research comparing facial recognition rates for thermal image cameras with visible light images using a variety of different algorithms. The comparison included aspects designed to reduce recognition accuracy such as variable illumination, different facial expressions, facial variations observed when speaking, and the use of eyeglasses. The results showed that several algorithms were able achieve a higher recognition rate using thermal images than with visible light images and the best results were from face recognition using thermal images. Another factor to take into consideration is that although thermal imaging cameras received better results overall, visible light images gave better results than thermal images with several different algorithms. This illustrates that care must be taken when choosing the method to be used for lip-reading with thermal images, as the results of the study will depend largely on the techniques used. As some techniques may yield better results in the visible light spectrum than they do in the thermal range and vice versa, simply applying methods developed for the visible light spectrum will not suffice.

Appenrodt et al (2009) conducted research into the use video for gesture recognition. As such they explored the use of standard 2D video, thermal video, and stereo video cameras for gesture recognition. Their results showed that 3D Cameras produced the best results for gesture recognition, followed by thermal video, followed by 2D video. However, the test did not include conditions with variable lighting which may have had an influence on results.

Due to thermal images detecting emissive light rather than reflective light they are more resilient to changes in illumination. However, thermal images offer different challenges to visible light images in that they are not as clear, and although resilient to changes in illumination, thermal images can be affected by changes in body temperature brought about by ambient temperature conditions, physical activity and other physiological conditions.

Having looked at previous research into the use of thermal images for facial recognition and gesture recognition, it can be seen that thermal imaging cameras give better performance than visible light cameras under variable lighting conditions. However, that result may depend on the algorithms used. Overall the best results for face recognition can be achieved when the output of visible light cameras is combined with that of thermal range cameras. It is expected that these findings will carry the same relevance for automatic speech recognition.

3.3 Techniques used in Audio Visual Speech Automatic Speech Recognition

This section explores those techniques used in AVASR for consideration as to which would be most suitable for thermal images. The construction of an audio-visual speech recognition system can be split into several parts, as well as the Audio ASR, the video side of the equation incorporates image pre-processing, region of interest extraction, data extraction, image post-processing and classification.

3.3.1 Image Pre-processing

Various image pre-processing techniques are used lip-reading systems. This can range from changing hue to histogram flattening, balancing of the left-to-right brightness distribution, all of which can minimise the effect of variable lighting conditions. For geometric feature extraction the image may also undergo thresh holding to reduce the image to binary format to simplify the process of feature extraction. (Scanlon & Reilly, 2001). As the data collection for this experiment was filmed under bright lighting with no variation, no pre-processing was done at this stage.

3.3.2 Region of Interest Selection

The ROI (Region of Interest) is the area of the image where the visual speech information is extracted from. Before the visual speech features can be extracted, the ROI has to be detected and tracked. In AVASR systems the ROI can vary from focusing solely on the lips to including, the jaw, the cheeks, and even extending to include the entire face. While having a smaller region of interest may decrease data demand and require more processing, a larger region of interest will impart more information. (Potamianos, et al., 2003). One factor that may have influenced Saitoh and Konoshi's results is that they focused their region of interest tightly on their lips. However, when people speak, there is more movement involved than the lip region therefore it was chosen to use a larger ROI for this research.

One of the difficulties Saitoh and Konoshi (2006) faced in their research was that they did not have an effective method for detecting the mouth region. In thermal images, the mouth can be difficult to detect, especially when closed, however Martinez, Binefa, and Pantic (2010) showed that once the eyes and nostrils have been located, the position of the mouth can be reliably estimated. As such, similar methods can be used to locate the region for lip reading.

According to Lucey, Dean, and Sridharan (2005) for AVSR to be effective, it is essential that the visual front-end is highly accurate, otherwise these errors will cascade through and reduce the ability of the AVSR system to reliably recognise speech.

Due to the difficulties involved in implementing accurate ROI selection and the limited research that has previously been conducted into mouth detection in thermal images, it was decided to extract the ROI manually.

3.3.3 Data Extraction

Once the ROI has been detected, data from that region is extracted for analysis. Methods for extracting information for lip-reading fall into three categories: Appearance based, shape based, or a combination. Appearance based lip-reading assumes that all pixels within a region of interest can be used to obtain information for speech recognition. Shape based lip-reading follows the movement of contours and curves such as lips and jaw movement. Combination based methods incorporate aspects of both shape based and appearance based methods. (Potamianos, et al., 2003).

In pixel based approaches every pixel in the region of interest is considered a feature. This approach ensures that no information is lost but the dimensionality of the feature vector is very high and contains a large amount of redundant information. To reduce the size of the files and separate the relevant information, images transforms are applied to the image. Image transforms can include methods used in this implementation include Discrete Cosine Transform (DCT), Hadamard Transform , Haar Transform, and Eigenfeatures . The DCT outperforms all other methods mentioned above for AVASR. (Scanlon & Reilly, 2001).

Shape based methods can fall into two categories: point tracking and model based tracking. Point tracking works by tracking the movement of points around the mouth, such as on the corners and along the lip lines. Point tracking offers the advantage of a large reduction in data requirements. The disadvantage of point tracking is the requirement for sophisticated edge detection algorithms which can be greatly affected by changing illumination conditions. (Rothkrantz, Wojdel, & Wiggers, 2006).

Model based lip reading is an extension of point tracking. Model based lip reading assumes the points lay on a predefined curve and limits the tracker to predefined models of the mouth. As the output of the tracker is limited to predefined models, it is more robust than point tracking alone as impossible situations can be eliminated. (Rothkrantz, et al., 2006).

According to Rothkrantz et al (2006) , appearance based methods come with two disadvantage over shaped based methods. Firstly, appearance based methods produce a large amount of data that requires heavy computation to process and storage of large files. Secondly, two slightly different images will produce greatly different datasets – making appearance based techniques highly susceptible to changes in lighting conditions with the datasets gathered for the same person repeating the same word being greatly different in slightly different lighting conditions. While these effects can be reduced by using large training sets and applying pre-processing to compensate for shadows, these techniques detract from the advantage of simplicity that appearance based methods have. However, according to Lucey et al (2005), Potamianos, Graf, and Cosatto (1998) conducted a review of shaped based and appearance based recognition methods using visual only speech recognition. This showed that appearance based recognition obtained superior performance as and more robust to visual noise and compression artefacts.

Saitoh and Konoshi (2006) implemented an appearance based approach, analyzing the region of interest by pixel value based on eigenvalues. However, they found that the results were limited and conjectured that the lower recognition rates were due to the pixels area within the region of interest not changing as much on the inside of the mouth with thermal images as they do with visible light images.

It was chosen to use an appearance based method in this research because of the simplicity of design, and superior performance. As the lips are not well defined in thermal images, this approach was chosen over model based methods as a shape based model would be difficult to implement.

3.3.4 POST-PROCESSING

Before classification, the visual signal undergoes post processing. This can include image normalisation to allow for variance in facial proportions for different speakers as well as differences in distance from the source camera. Post processing can also include dynamic time warping to allow for different speaking rates. (Potamianos, et al., 2003). No post-processing was used in this research. The AVASR system is a speaker dependent system because of this. For a speaker independent system, image normalisation taking into account speakers different facial proportions would need to be added.

3.3.5 COMBINING VISUAL AND AUDIO SIGNALS

The visual and audio signal may be combined before (low level fusion) or after signal classification (high level fusion). Low level fusion assumes there is a direct dependence between the visual and audio modalities; therefore combination occurs at feature level by combining or concatenating features to use a single classifier. High level fusion assumes complete independence between the visual and audio modalities; therefore separate classifiers are used for each with only the outputs of the classifiers combined.

Low level fusion offers the advantage that less complementary information is lost. However, low level fusion requires accurate synchronisation of signals and at lower levels of fusion. Therefore with low level fusion there is a higher risk of corruption between channels and factors such as noise in the acoustic channel can corrupt the visual information and poor ROI selection can corrupt audio information. Due to the additional information being put into the classifiers, low level fusion requires larger amounts of training. (Scanlon, 2005).

High level fusion suffers from the disadvantage that there is a greater degree of complementary information lost. However, as well as the classifiers requiring less training data, high level fusion is advantageous when it comes to benchmarking different algorithms as it means signals can be compared independently and allow for the redesign of the front end data extraction. (Potamianos, et al., 2003).

It was chosen to use high level fusion to combine the visual and audio signals because of the lower level of complexity required to integrate the signals and the ability to easily compare and benchmark the accuracy of different components.

When combining visual and audio channels, care must be taken to decide how much weight to give each signal as by giving too much weigh in classification to one modality may exacerbate incorrect categorisation. Scanlon and Reilly (2001) recommend using adaptive weighting the audio and visual outputs to the dispersion or variances of their output probabilities, which indicates reliability of the modalities. These adaptive weights can account for the confusion in both channels.

3.3.6 Speech Classification

Units of speech may be classified as either a whole word or split into smaller portions. In visual speech recognition systems, mouth movements are typically classified into groupings known as visemes which are matched to phonemes (the smallest unit of sound) for audio recognition.

As there are several sounds a given mouth shape can represent, one viseme may represent several different sounds. While there is no standard grouping for phonemes and visemes, Table 2 shows typical phoneme/viseme groupings.

Phoneme	Viseme	Phoneme	Viseme
P	/p/	K	/k/
B		G	
M		N	
EM		L	
F	/f/	NX	
V		HH	
T	/t/	Y	
D		EL	
S		EN	
Z		IY	/iy/
TH		IH	
DH		AA	/aa/
DX		AH	/ah/

W	/w/	AX	
WH		AY	
R	/ch/	ER	/er/
CH		AO	/ao/
JH		OY	
SH		IX	
ZH		OW	
EH	/ey/	UH	/uh/
EY		UW	
AE		SIL	/sp/
AW		SP	

Table 2 Viseme-phoneme groupings. Adapted from Lucey et al (2004).

The advantage of classifying sounds or visual images into phonemes and visemes is that once classified into different types different syllables can be strung together in a tree directory structure to match input to dictionary words – similar to predictive text on a mobile phone. This allows for a smaller database of images to be stored and reduces processing requirements; making matching of words a less process intensive task and easier to achieve than matching entire words. (Potamianos, et al., 2003).

The disadvantage of classifying sounds or visual images into phonemes and visemes adds an extra layer of complexity to the system in creating the problem of separating a word into separate visemes and phonemes. In this research due to the small sample size to be collected and the added complexities therefore it was chosen to classify by isolated word recognition.

While there are several methods that can be used for signal classification, including weighted distance in visual feature spaces, support vector machines, and artificial neural networks, Hidden Markov Models are the most commonly used recognition technique in AVASR and ASR (Potamianos, et al., 2003). As the goal of this research is not to explore different statistical methods, Hidden Markov Models were chosen to be used in this project because of their previous success and wide use in research.

Models to for classifying signal origins can be placed in generally two groups: Deterministic models and statistical models. Deterministic models operate by manipulating known properties of a signal, such as if the signal is known to be in the form of a sine wave, to estimate values for the parameters of a signal. In contrast, statistical models do not try and find the input parameters of a model, instead only working with the statistical properties of a signal. Along with Poisson processes, and Gaussian processes, Hidden Markov Models are examples of statistical methods (Rabiner, 1989).

According to Juang and Rabiner (2004), Hidden Markov Model theory was first published by Baum and Petrie (1966) in the in the late nineteen-sixties and early nineteen-seventies and was first used for speech recognition by IBM in the nineteen-seventies. HMM uses a Markov chain to represent the linguistic structure and a set of probability distributions to account for variability in signal. Given a set of known words (or visemes), and a sufficient collection of their representations an efficient estimation method, called the Baum-Welch algorithm, is used to obtain the “best” set of parameters that define the corresponding model or models. The resulting model is then used to calculate the likelihood that an unknown utterance is a word (or viseme) represented by the model. (Juang & Rabiner, 2004).

3.3.7 SECTION SUMMARY

While thermal cameras are less influenced by lighting than standard cameras, they may suffer from decreased performance due to changes in skin temperature.

Although standard images outperform thermal images in face recognition, thermal images have been shown to be more reliable than standard images for face recognition in uncontrolled conditions. Fused thermal and standard images outperform both.

The lips and mouth in thermal images can be difficult to detect and research is limited into this area is limited. Developing an accurate system for auto extraction of the ROI is a separate project in itself and would take away from the scope of this research. Because of this, it was chosen to manually extract the ROIs and an appearance based method for data extraction. DCT was chosen for image data extraction.

Classification was chosen to be HMM based on whole word recognition with high-level integration to reduce complexity and allow for comparison of the individual modalities.

4. Database Creation

To gather data for testing of the AVASR system, an audio visual database consisting of standard video, thermal video and audio was created. This section describes the process used to create the database and is split into three parts. The first section covers data criteria, the second covers the equipment used, and the third covers the data collection.

4.1 Data Criteria

The research instrument used to collect data for testing and demonstration is an essential element of this research project as the quality of the data collected can have a great impact on the results. As there are no available AVASR databases which include thermal video, a custom database needed to be constructed for this research. As part of this, the words used in testing are important as poor word choice could bias the results for example it would be easier differentiate between the words "hippopotamus" and "ten" that it would be to differentiate between the words "goat" and "boat". Before collecting data a review of existing methodologies and databases was conducted.

Millar, Wagner and Goecke (2004) pointed out the need for improved methodology in existing AV databases calling for a large number of speakers for statistical significance, a broad coverage of phonemes and visemes, different levels of acoustic noise starting with `clean speech' , whole-face images in colour, short words and continuous speech with transcription, and extensibility. Such a corpus would allow the database to be used for a variety of purposes and allow greater comparison of results between researchers.

While focused on audio only ASR, Becchetti and Ricotti (1999) describe different categories of speech databases the first type is *analytic-diagnostic* databases which are used to research the basic linguistic and phonetic elements of speech. A second type is *generic* databases which are suitable for a wide range of applications and have a non-specific vocabulary. The third type of database is *specific* databases which are designed for specific purposes such as information request systems.

Becchetti and Ricotti (1999) further describe the various types of speech that can be found within a database. These are *reading of isolated phonemes* where individual phonemes are recorded for comparison with natural speech. *Reading of isolated words* where the words can be either real words or the nonsense words. *Reading of isolated phrases* where words are recorded within a sentence to give a more natural pronunciation. *Reading of text fragments* where participants are asked to read from portions of text to enable a more natural pronunciation within sentences that are semantically tied. *Semi-spontaneous speech* where the vocabulary and syntax are controlled, an example of this would be the reading of telephone numbers. *Spontaneous speech on a predetermined topic* where the speech elicited is unrestrained speech however a topic of discussion is given to give some control and ensure specific words are repeated. *Speech elicited by the "Wizard of Oz" method* where participants are required to give instructions to a computer secretly controlled by a human to gather recordings of a participant acting as they would when interacting with a computer that has speech recognition. Lastly there is *spontaneous speech* where participants are asked to speak on any topic of their choosing.

The AVOZES database contains 20 speakers (10 male and 10 female) speaking consonant-vowel-consonant and vowel-consonant words in a carrier phrase covering the range of phonemes and visemes in Australian English, the digits zero through nine in a constant carrier phrase, and three sentences as examples of continuous speech (Goecke & Millar, 2004).

The CUAVE database is a corpus consisting of over 7,000 utterances of both isolated and connected digits from 36 speakers. The database contains five different datasets. The first set of data the speakers were filmed standing still speaking 50 isolated digits each. For the second dataset the speakers were filmed moving back and forth and side to side and tilting. The third dataset contains each participant speaking 20 isolated digits in profile. The fourth dataset contained the speaker facing the camera reciting 60 connected digits including telephone-number-like sequences – 30 while standing still and 30 while moving. The final dataset contains speakers in pairs speaking interconnected sequences one participant following the other as well as both reciting different speakers at the same time. (Patterson, Gurbuz, Tufekci, & Gowdy, 2002).

AVICAR is a corpus recorded in a car environment, the advantage of which is that it gives a more accurate simulation of real life scenarios. AVICAR consists of four datasets: isolated digits, phone numbers, isolated letters, and sentences. 100 speakers

(50 male and 50 female) are used and scripts were recorded at five different noise levels – with the car at idle, driving at 35mph with the windows closed, driving at 35mph with the windows open, driving at 55mph with the windows closed, driving at 55mph with the windows open.(Lee et al., 2004).

Patterson et al (2002) describe Tulips1 consists of twelve participants reciting the first four English digits. AVLetters contains the English alphabet repeated three times by ten participants. DAVID consists of 31 participants speaking digits, alphabets, vowel-consonant-vowel syllable utterances and, and some voice conferencing commands. Lee et al (2004) describe MOCHA as 78 participants speaking 48 isolated words.

After reviewing the available AV databases, it was chosen to create an analytic-diagnostic database recorded from a front only view containing the digits zero through nine and 15 isolated words to ensure the entire range of phonemes and visemes as listed by Lucey (2004) were covered. A data collection took place with these words, however as there were problems with the quality of the data a second data collection was taken and the words were cutback to only include the digits zero through nine. It was felt that this would be sufficient at this stage and is in line with other research in this area.

If further research were to be completed in the area a more advance database would be required. Ideally this would include digit strings along with all potential combinations of visemes/phonemes within a carrier phrase to give a more natural pronunciation as well as isolated digits and words. As the purpose of using thermal cameras would be for improving ASR in variable lighting condition, this should also include the collection of data under changing lighting conditions.

4.2 EQUIPMENT

Scanlon (2005) gives the minimum frame rate for lip-reading at 15fps. This was based experiments with human trained lip readers. Below15 fps the recognition rate drops substantially for human lip-readers. Equipment was selected assuming that this holds true for AVASR. The thermal camera that was used for data collection has a refresh rate of 50 Hz and a resolution of 160 × 120 pixels for the sensor. As the camera lacked the ability to record video to SD card, the videos were captured using PAL, reducing the frame to rate 25fps and stretching the image to 720 × 576 pixels. Standard video was filmed at 25fps using a full high definition camera at 1920 × 1080 resolution. However this was reduced to 720 × 576 pixels due to the larger amount disk space and time required to process Full HD files when editing. Despite the change in resolution, the standard video camera offers a much higher quality image than the thermal video camera. This highlights one of the key disadvantages thermal cameras have in ASR and it is expected that this will have had an effect on performance.

4.3 DATA COLLECTION

Participants were filmed from front on with both the thermal and standard cameras. The cameras were placed behind a LCD monitor where the words to be read were displayed. The positioning was to ensure that the participants were looking directly at the cameras during recording. The cameras were focused to on the participant to have their whole head in frame with some leeway for head movement.

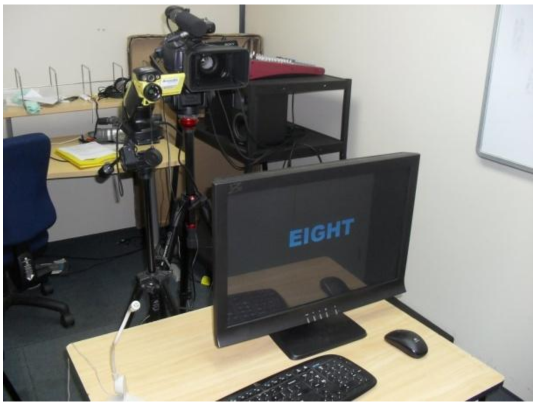


FIGURE 6 DATA COLLECTION VIEW TOWARDS CAMERAS

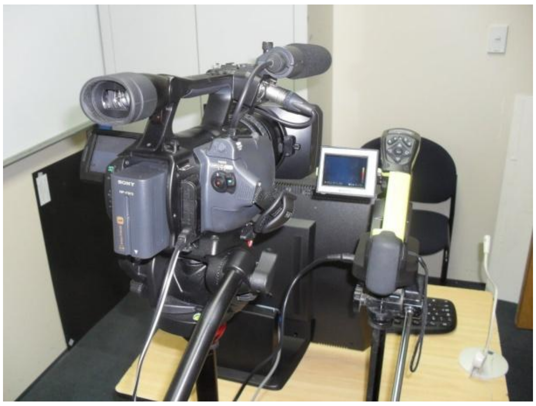


FIGURE 7 DATA COLLECTION VIEW TOWARDS PARTICIPANT

Participants read from a PowerPoint presentation of 50 words consisting of the numbers “zero” through “nine” repeated five times each in random order. The slides were timed to automatically change at an interval of approximately 3 seconds between each slide. The random order and timing of slides was to prevent counting and ensure that words were spaced apart to make the task of editing easier. The process was repeated three times for each participant, allowing for the participant to have a break in between. This gave a total of 150 words for each participant.

Eleven participants volunteered for the research. This resulted in a total of 1650 words. No screening was done on the participants. A wide variety of ethnicities took part resulting in a wide range of skin tones. Four participants were female and five were male. Two participants had facial hair (full beards) and the rest had none.

The recordings were edited into separate words with the thermal video, standard video, and audio split into separate files. Editing was completed using two computers side by side to try and ensure consistency between the two different datasets. Standard video and thermal video were edited to ensure the words were synchronized in timing and duration. The regions of interest were extracted manually from the standard video and thermal video. Audio from the standard video was stripped into separate files (at a sample rate of 44100 Hz) and used for audio recognition. Once the audio was extracted, separate audio files were generated with Gaussian white noise added at signal to noise ratios of 20, 10, 0, -10, and -20dB. This resulted in a total of 13200 files in the database: 1650 for thermal video, 1650 for standard video and 9900 for audio at different noise levels.

Signal to noise ratio was calculated using the following formula:

$$SNR = 10\log_{10}\left(\frac{Power_{Signal}}{Power_{Noise}}\right)$$

Where power is measured in root mean squared (RMS). The noise in the original signal is not included in this figure and are therefore the SNRs are not exact measurements.

A number of factors affected the quality of data collected. While the data collection was completed in a relatively quiet environment, the room was not soundproof and next to a busy office. As a result the sound of voices and phones along with doors opening and closing could be heard in the background of some samples. The sound of a noisy air conditioner could also be heard in the background and added noise to all samples. In addition, the thermal camera would auto adjust itself occasionally resulting in a clicking sound in the audio as well as affecting the image in the thermal video files. Although participants were asked to try and not move their head, many still moved from side to side as well as backwards and forwards as part of their natural speech which made consistent ROI selection difficult when editing. While these factors may have diminished overall results and editing them out may have led to higher recognition rates, they were left in the dataset so as to give a truer representation of the results obtained and in a way represent real-world situations.

4.4 Section Summary

A custom AV database was created for this research. This consisted of eleven participants filmed in thermal video and standard video reading the words “zero” through “nine” repeated fifteen times each. This resulted in 150 samples from each speaker for a total of 1650 utterances.

5. SYSTEM DESIGN

Separate systems were constructed for audio and visual speech recognition with HMM used for classification. These were combined using high level fusion to allow for separate testing of individual elements. MATLAB R2009b was used for programming and creating the recognition system along with the HMM MATLAB Toolkit (Murphy, 2005) and the VOICEBOX: Speech Processing Toolbox for MATLAB(Brookes, 1999) . This section is split into four parts describing the audio processing, the video processing, the HMM classification, and the testing process with results.

5.1 AUDIO PROCESSING

Audio recognition was an MFCC based on that described by Becchetti and Ricotti (1999). This section is split into two parts, the first describing the audio pre-processing, and the second describing the MFCC extraction process.

5.1.1 AUDIO PRE-PROCESSING

Five steps were taken in pre-processing the audio signal. The first step taken in pre-processing was silence removal to ensure that unneeded portions of the signal weren't processed. The second step was to run the data through a noise filter to reduce noise levels. The third step was to normalize the volumes of the signals so that. The fourth step was to apply a band pass filter to the signal so that only the areas of speech which contain the greatest information were speech processed. The fifth and final step was to apply a pre-emphasis filter to the signal to boost speech in the higher frequency ranges.

HMM based ASR can experience a significant reduction in performance if long silences are not removed from speech, therefore silence removal was performed using the functions provided for by Giannakopoulos (2010). The silence removal algorithm works by splitting the signal into non-overlapping 50ms frames. For each frame, the spectral energy and spectral centroid calculated and with the average spectral energy and average spectral centroid for the entire signal. If the frame has either the spectral energy or spectral centroid higher than that of the average for the entire signal it is classified as silence.

Noise reduction was then performed using the VOICEBOX function ssbusme (Brookes, 1999). This function enhances speech by first estimating the SNR then adaptively smoothing the signal based on that estimate.

Volume normalization was achieved by finding the peak amplitude of the signal and calculating the ratio to make the peak 90% of full volume. This was multiplied throughout the signal to ensure all words were at a similar sound level. The disadvantage of doing this is noise in the signal is also amplified for quiet speakers.

The signal was then passed through a band pass filter with the low cut-off frequency set at 300 Hz and the high cut off frequency at 7 kHz. The purpose of this was to further reduce noise in these frequency ranges. These values were chosen because according to Becchetti and Ricotti (1999) while the bandwidth human voice is approximately 16 kHz, most of speech energy is under 7 kHz and most noise occurs at low frequencies.

The final step taken in pre-processing was pre-emphasis. According to Becchetti and Ricotti (1999) because the high frequency speech formants have smaller amplitudes than low frequency speech formant, pre-emphasis is required to obtain similar amplitudes for all sounds. This is done via a first order Finite Impulse Response (FIR) filter with an impulse response in the z-domain of:

$$H(z) = 1 - a \cdot z^{-1} \quad 0 \leq \alpha \leq 1$$

This results in the time domain relationship between the pre-emphasized and input signal:

$$x'(n) = x(n) - \alpha x(n-1)$$

The value of alpha was set at 0.95 because Becchetti and Ricotti (1999) stated this was the value most often used.

5.1.2 AUDIO DATA EXTRACTION

To calculate the Mel-Frequency Cespral Coefficients (MFCCs), the melcepst VOICEBOX: Speech Processing Toolbox for MATLAB (Brookes, 1999) was used. The process of MFCC calculation is completed in several steps as shown in Figure 8.

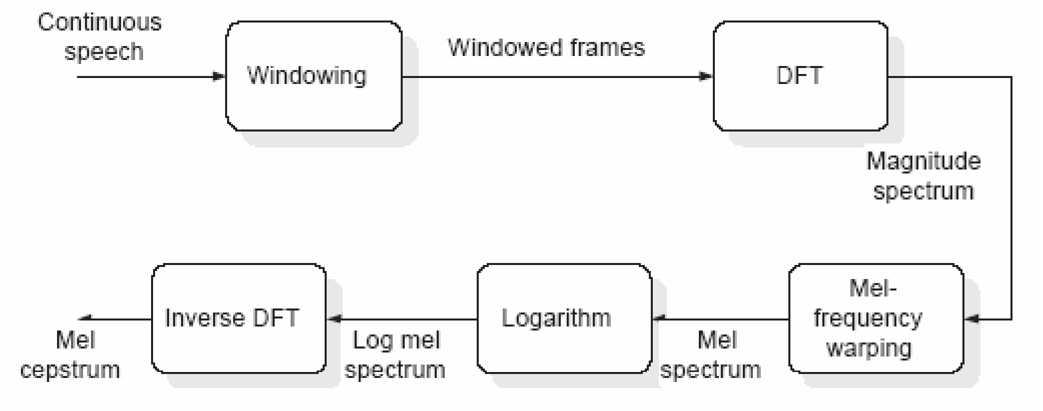


FIGURE 8 MFCC CALCULATION (SHANEH & TAHERI, 2009)

The first step taken in MFCC generation is to split the signal into short time intervals or windows. This is because spectral evaluation methods are reliable when the signal is unchanging. As voice signals are constantly changing this condition is only met for short periods of time. The simplest windows have a rectangular shape; however the presence of a window distorts the estimated spectrum. To minimise these effects the window can be shaped by multiplying the signal with a shaping function. In ASR the most common window shape, as was used in this research, is the Hamming window which has a raised cosine impulse response:

$$w(n) = \begin{cases} 0.54 - 0.46cos\left(\dfrac{2\pi n}{N-1}\right) & n = 0, \dots \dots, N-1 \\ 0 & otherwise \end{cases}$$

The size of a window is a trade off between the required time and the frequency resolution. Short windows (3-4ms) give a more accurate representation of the voice signal at a given time; however they give a lower resolution of frequencies. Longer windows (70ms and over) give a higher frequency resolution, however fast changes in the voice signal are not detected. As a compromise window size is usually selected to be 20-30ms in length. As such a window size of 30ms in length was selected with a step of 15ms between windows. (Becchetti & Ricotti, 1999).

In the second step, spectral analysis is performed on each window by taking the Discrete Fourier Transform (DFT) which moves the signal from the time domain into the frequency domain.

Thirdly, the results of the DFTs are then passed through a set of band pass filters to obtain the spectral feature of speech. The filters are set along the *Mel* scale. With the Mel scale the centre of each filter is placed at even intervals for frequencies below 1 kHz and a logarithmic scale for frequencies above 1 kHz. The number of filters was set at 24 because it simulates human ear processing. This results in an array of 24 coefficients.

The fourth step is to take of the logarithm of the filter bank output. The effect of the logarithm scale is that it reduces the component amplitudes at every level. In the final step, the MFCC coefficients are calculated by taking the inverse Fourier transform of the logarithms. (Becchetti & Ricotti, 1999).

5.2 VIDEO DATA EXTRACTION

The same method of data extraction was used for thermal video as for standard video. Video processing was a modified version of Motion Templates as proposed by Yau (2008). This method was chosen because of the advantages it offered over other methods which are: ease of calculation, high level of invariance to skin tone, easy to train for individual users and does not rely on artificial markers on the speakers' faces. As this method is a pixel based method it, it does not require detection of lips – which can be difficult in thermal images.

5.2.1 MOTION TEMPLATE GENERATION

Motion Templates use image subtraction based on intensity to give a single image to represent the series of lip movements within a sequence. This is done in three steps: first the ROIs are normalized in size, secondly the difference between consecutive frames is calculated, and thirdly the differences between frames are merged together. Once the MTs are generated, DCT coefficients are then extracted from the resulting image for HMM training.

The first step taken to produce a MT is to convert the ROIs to a constant size. This allows the ROIs to be subtracted from each other as well as providing a degree of normalization between ROIs, for example if the camera is at different focal lengths. A ROI size of 64×64 was chosen based on Scanlon's (2005) finding on ROI size which was that no loss of detail could be seen between 128×128 and 64×64, but a large loss of detail could be seen between 64×64, 32×32, and16×16.

The next step taken in producing a MT is to calculate the difference of frame (DOF) between adjacent frames in the video. The DOF is the result of image subtraction based on intensity and is calculated by first converting the images to greyscale so that the images a representative of intensity then finding the absolute value of the second frame subtracted from the first frame. This can be represented by the equation:

$$I(x,y)_{DOF} = |I_1(x,y) - I_2(x,y)|$$

Non-zero pixels in the DOF image represent differences in intensity between the two images which can be interpreted as movement (or changes in temperature for thermal video).

In the next step of MT creation, the DOF images are then converted to a binary image. Each non-zero value in the binary image represents movement (or change in temperature for thermal images). During conversion a threshold is applied on the images using a predefined value (α). This threshold value is required because not all the differences between the two images represent movement; some differences may represent noise and slight changes in illumination. The resulting binary image can be represented by the equation:

$$B_t(x,y) = \begin{cases} 1 & if\ DOF_t(x,y) \geq \alpha \\ 0 & otherwise \end{cases}$$

This represents the movement between two frames. However, as noted by Scanlon (2005) the amount of time between frames is small and does not provide significant temporal information. Therefore in the generation of Motion Templates multiple frames are combined, however adding the frames together does not contain temporal information about motion and direction of movement. To combine temporal information the successive DOFs are converted to from binary images to greyscale images and combined together with varying intensity related to time.(Yau, 2008).

Yau (2008) used a weighted binary method for combining DOFs. This was done by assigning a normalized value with respect to time for each frame and multiplying the DOFs by that value, making the most recent changed pixel the highest intensity. The values in the MT are then generated by selecting the highest values for that pixel out of the difference of frames. The intensity values of the MT in the t^th frame are given by the equation:

$$MT_t(x,y) = max \bigcup_{t=2}^{N} B_t(x,y).t$$

Where N is the total number of frames used to capture the mouth motion. A scalar valued greyscale image (MT) with pixel brightness representing the history of mouth motion is generated by computing the MT values for all pixel coordinates (Yau, 2008).

A different method for combining DOFs was used in this research. This is because weighted binary has the disadvantage that by selecting the highest value when combining DOFs only the most recent movements are included in the MT and the earlier movements are ignored. Therefore binary weighting does not give an accurate representation of the temporal movements in the recording of lip motion.

Instead an additive approach was used towards generating MTs. This was done by generating MTs on the fly by converting the first binary DOF to greyscale image with and halving the value to give the first MT. Successive MTs are generated the converting binary DOF to greyscale image and halving the value then adding half the value of the previous MT. This can be represented by the following equation:

$$MT_t(x,y) = \begin{cases} \frac{1}{2} B_t(x,y) & if\ t = 1 \\ \frac{1}{2} B_t(x,y) + \frac{1}{2} MT_{t-1}(x,y) & otherwise \end{cases}$$

The result is a MT with the value of each pixel representing both the most recent movements for that pixel combined with previous movements. More recent movements will contribute a higher value to the MT with earlier movements contributing a lower value to the MT. The MTs contributions are halved with additional DOF added meaning no series of movements will add up to the same as another series of movements as each contribution is effectively switching individual bits in a binary number.

The different technique used in generating MTs means that instead of a single MT being generated for an entire lip movement as done by Yau, lip movements need to be split into several MTs. While binary weighting can combine up to 255 DOFs (representing 256 video frames) using an eight bit greyscale image, the additive technique used in this research can only combine a maximum of eight DOFs (representing 8 video frames) into an eight bit greyscale image. This is because by halving the value each time the modifiers for each DOF are the same as each bit value in a binary number: 128, 64, 32, 16, 8, 4, 2, and 1. The effect of using more than 8 DOFs is that pixels representing movement further back will add 1 to the intensity level and result in confusion for lip movement history. It was chosen to use MTs of 6 DOFs with 3 frames between each MT as preliminary testing showed slightly higher recognition rates than using other values. As MTs uses image subtraction, the number of DOFs used to generate a MT of *N* length represents *N+1* video frames. Therefore each MT used for this research represents 7 frames of video, which equates 280ms in length with the 25 frames per second rate used.

MTs with little or no movement were rejected from the final data sequence. The detection of MTs with little or no movement was done by first calculating the MT for a sequence then applying simple thresh-holding based on the number of non-zero pixels within a MT. If the movement is below the threshold value, the MT is rejected. This is to minimise the effect of noise and small head movements not related to speech.

Figures 9, 10, 11, and 12 show video sequences and the MTs generated from the word zero.

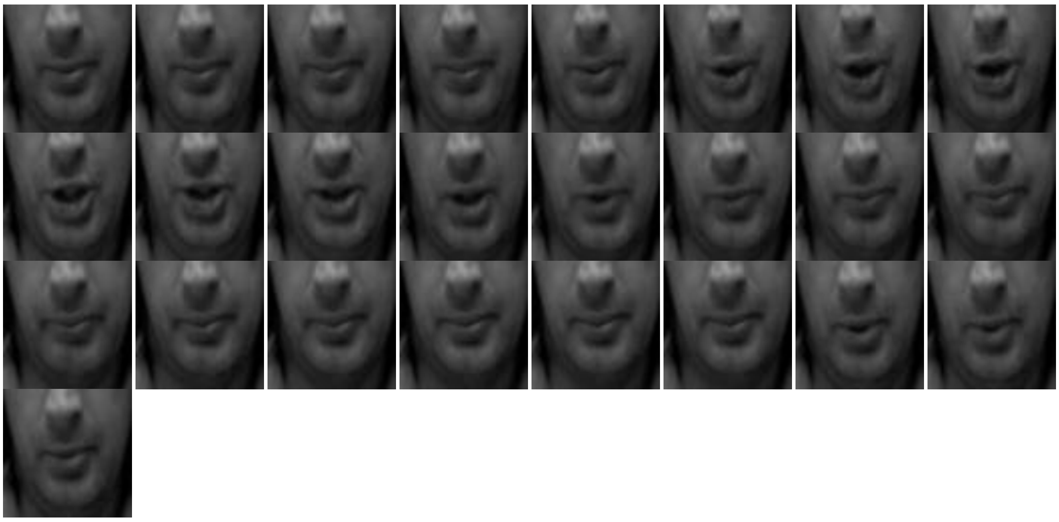


FIGURE 9 STANDARD VIDEO OF THE WORD “ZERO”

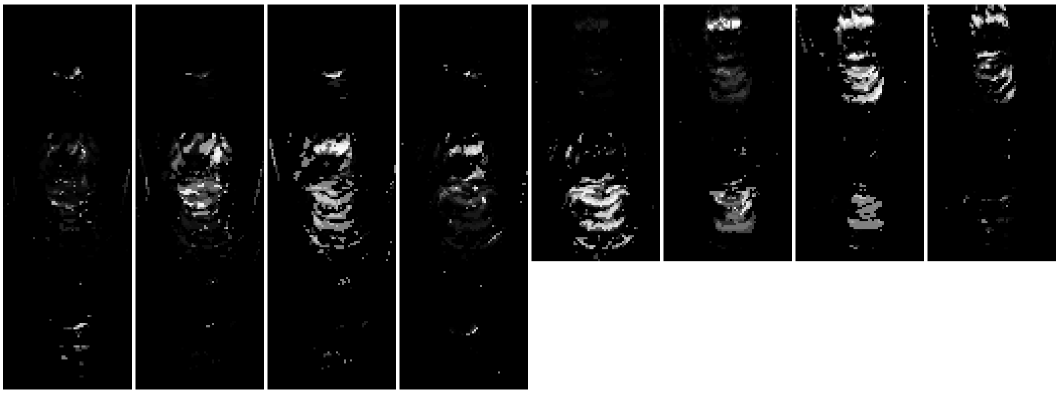


FIGURE 10 MOTION TEMPLATES GENERATED FROM STANDARD VIDEO FOR THE WORD "ZERO"

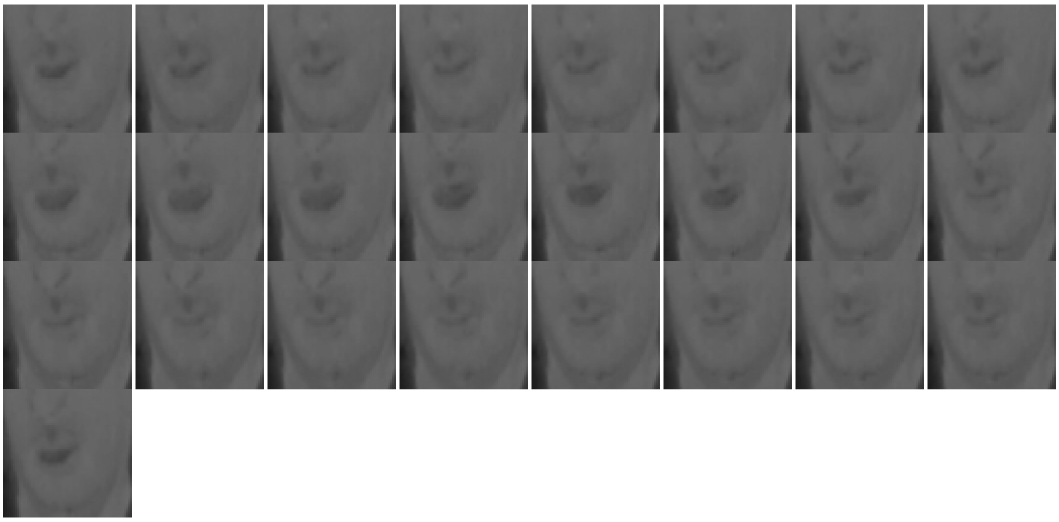


FIGURE 11 THERMAL VIDEO OF THE WORD "ZERO”

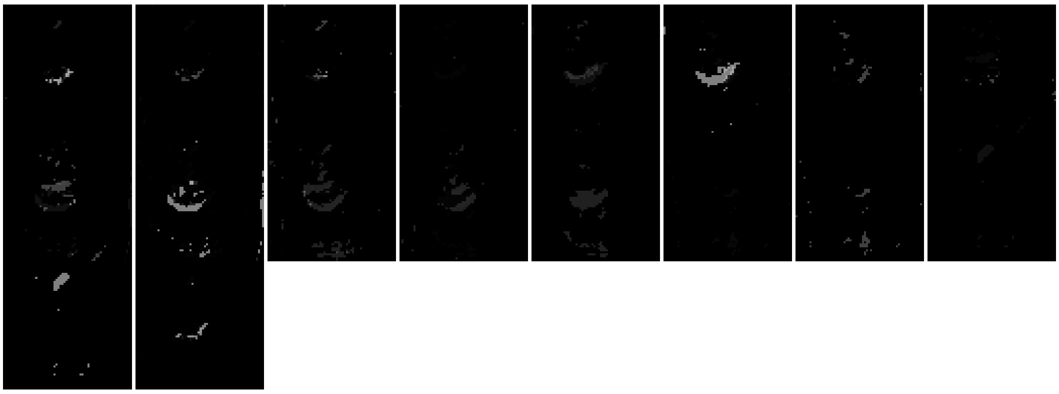


FIGURE 12 MOTION TEMPLATES GENERATED FROM THERMAL VIDEO OF THE WORD “ZERO”

As can be seen, a lot more information is obtained from the standard video than the thermal video. This may be due to the much lower resolution of the thermal camera’s sensor compared to the standard camera. Another factor is that this may be in part because the thresholding levels in either the conversion from greyscale to binary in the DOFs for MT were set too high or the difference in intensity levels was not high enough. These may be further improved by fine tuning of the values, but as they were already very low there is the danger that the influence of noise in the signal will increase.

5.2.2 DCT COEFFICIENT SELECTION

Once the MTs are generated, Discrete Cosine Transform coefficients are extracted for the HMM inputs. Numerous research papers have found that using DCT for extraction outperforms other methods of extraction for visual speech recognition.

DCTs are an image transform often used in image compression. The DCT shifts the image into the frequency domain. The output of the DCT is an $N \times N$ matrix the same size as the input image. When reverse DCT is performed on the matrix the image is reconstructed. The way image compression occurs is that the more DCT coefficients that are included in the reverse transform the closer in match the reconstruction is to the original. If fewer coefficients are used in the reverse DCT, the image will be reconstructed with less accuracy. This allows an image to be compressed and use less storage space. (Scanlon, 2005)

The ability DCTs have to reduce the amount of information required to represent an image reduces the size of the data being used by the HMM classifier in AVASR. The coefficients are selected from the top left hand corner of the matrix in a zigzag pattern as shown in Figure 13. This is because the coefficients for the waveforms at lower frequencies, which have a higher influence on the total image, emanate from the top left hand corner. The number of DCT coefficients can have a great impact on the results. If too few coefficients are selected not enough detail is provided for reliable recognition. If too many coefficients are selected more samples for training are required. (Scanlon, 2005).

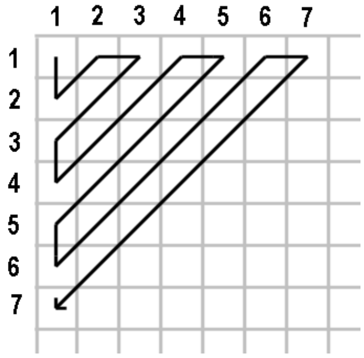


Figure 13 DCT coefficient selection order. Adapted from Yau (2008).

The number of DCT coefficients selected was 78 for both thermal video and standard video, this number was determined through pretesting and using the number of coefficients that gave the best accuracy.

5.3 HMM CLASSIFICATION

The HMM classification was implemented using the Hidden Markov Model (HMM) Toolbox for MATLAB (Murphy, 2005). The output of the HMMs is a vector containing the log likelihoods of the word with highest value being the most likely match.

The outputs of the HMMs can be combined using multiplication, addition, subtraction or division of the HMM output vectors (Scanlon, 2005). Multiplication was chosen because the magnitudes of the HMM outputs could be greatly different meaning addition had little effect on combining the HMM outputs. Because the HMM outputs are negative, and the selection was based on the output with the highest value which is the value closest to zero, the output selection was chosen to be the smallest value of the absolute value of the matrix multiplied.

Weighting for the three streams was manually calculated by running the algorithms separately and taking the amount of confusion for each word into consideration. This was done individually for each participant. For example, participant 1 the word "zero" had a 100% correct recognition rate for audio, however when the word "zero" was guessed, it was the correct word was "zero" 45.45% of the guesses. For the rest, the correct word was "nine" 27.27%, "eight" 3.03%, "seven" 9.091%, and "six" 15.15% of the time. The rest of the words were not guessed at all. Because HMM outputs are log likelihoods of the probability and are always negative, these values were subtracted from 1. This reduces the influence of the HMM outputs with a lower probability of being the correct answer and increases the influence of HMM outputs for results with a higher probability of being correct. As the weightings were calculates by hand, this was a very time consuming process. This process could be automated by incorporating a self testing loop in the training process for larger amounts of data.

Adding further adaptive weighting based on noise levels and therefore reliability in each stream could further improve results. By estimating the SNR ratio of an audio signal less weight could be given to the audio stream in a noisy environment and vice versa for video images. (Scanlon, 2005).

5.4 Testing and Results

The HMM training used in testing was speaker-dependent. For each participant, 14 samples of each word were used in training the HMM with the remaining sample used for testing. The training for video used clean data only whereas audio was tested with clean data and Gaussian white noise added at levels of 20, 10, 0, -10, and -20 decibel signal to noise ratio (SNR). As no noise was added to video, the testing was done on a perfect image with varying levels of noise in the audio signal only. The test was repeated five times with different samples selected for testing and the results were averaged to reduce sample bias. Figure 14 shows the average recognition rates achieved at the different noise levels.

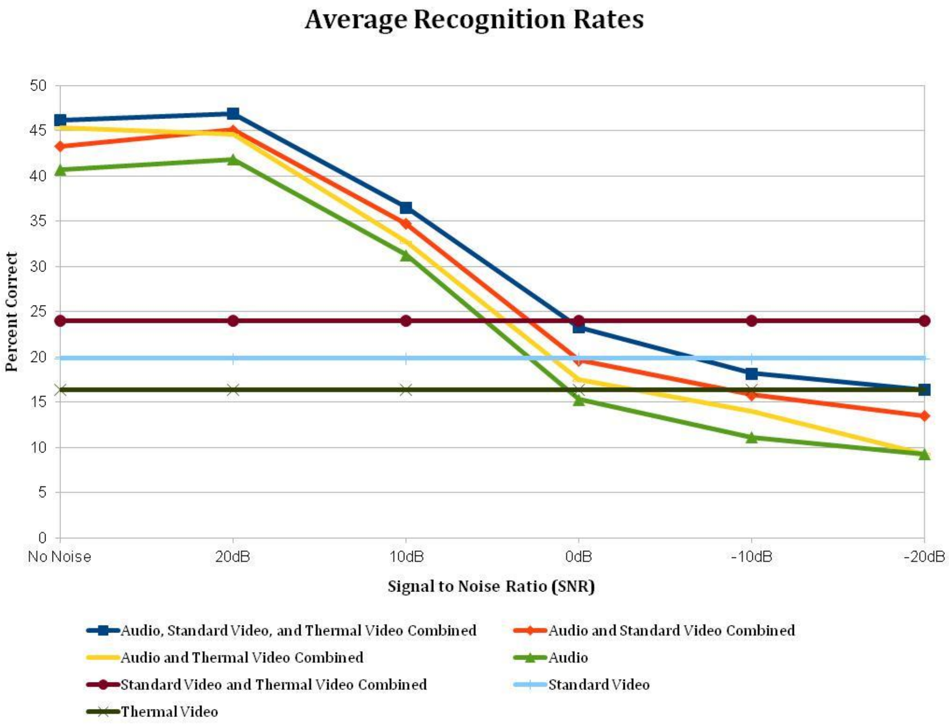


FIGURE 14 AVERAGE RECOGNITION RATES

The results were in line with expectations and showed an improvement of recognition rates for audio, standard video, and thermal video combined ASR above both combined standard video and audio ASR and audio only ASR across all noise levels. This was calculated to be a relative average improvement of 11.8% over combined audio and standard ASR and a relative average improvement of 39.2% on audio only ASR across all noise levels. Audio and standard video ASR combined gave a relative average improvement of 23.6% over audio only ASR, confirming existing theory with the visual modality benefiting speech recognition. Audio and thermal video ASR combined also improved over audio only ASR with a relative average improvement of 10.5% over audio only ASR.

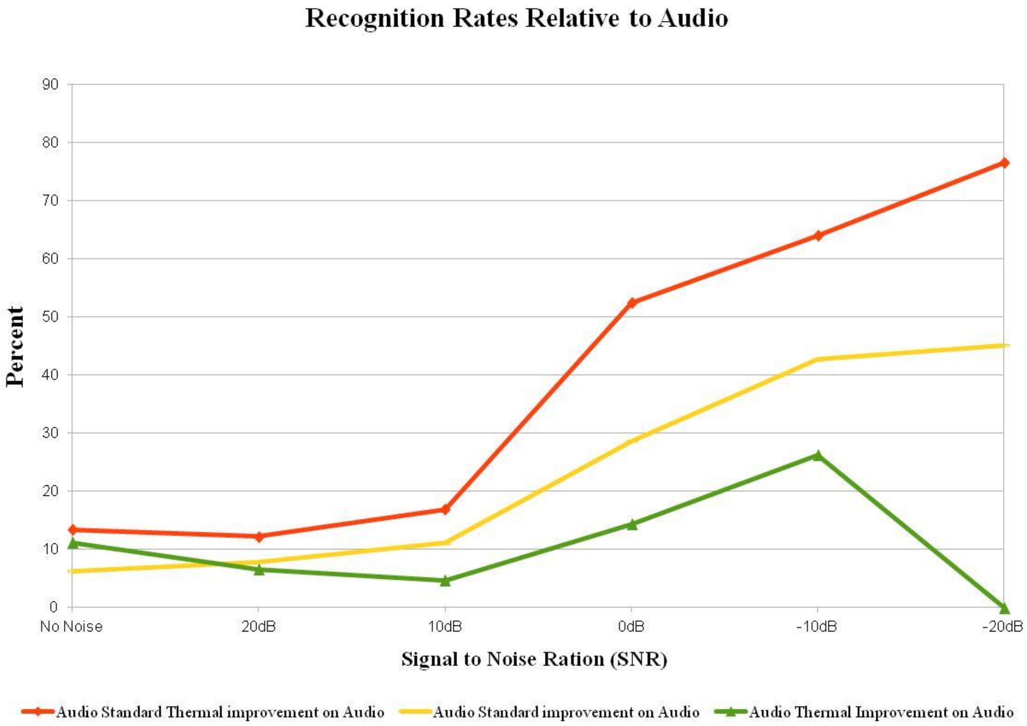


FIGURE 15 RECOGNITION RATES RELATIVE TO AUDIO

As can be seen from the chart above, the majority of improvement came at higher noise levels, where audio only ASR performs poorly with audio, thermal video, and standard video combined giving a relative improvement in recognition rates of 76.5% over audio only ASR. At these levels audio still carries a heavy influence on speech, resulting in poor recognition rates which are below that of the visual channels. These results may be further improved by adaptively adjusting the modifiers taking into consideration the noise in each modality as suggested by Scanlon (2005).

As expected, thermal video ASR did not perform as well as standard video ASR with standard video ASR returning a 19.8% recognition rate and thermal video ASR returning a 16.4% recognition rate. However, combined thermal video and standard video ASR returned an average recognition rate of 24% reflecting a relative improvement of 21% over standard video ASR.

While the results showed adding thermal video to ASR gave a clear improvement in recognition rates, the overall recognition rates were not as high as hoped. Part of the reason for this may have been due to the quality of the data used. Looking further into the results there was a large degree of variability between participants. This may be in part down the nature of the data – head movement, inconsistent ROI selection, voice level, and background audio noise. An example of this is Participant 2 – who while softly spoken had very little head movement and gave low results for audio only ASR and higher than average results for thermal video and standard video ASR, whereas Participant 10 returned high recognition rates for audio ASR, but had a large amount of head movement which resulted in a low recognition rates for both standard video and thermal video ASR.

As this data was tested using custom data, the recognition rates cannot readily be compared with other research in this area. Factors such as word selection, head movement, background noise levels, and consistent ROI selection can greatly affect results. The inability of researchers to compare results to find which techniques work better in AVASR has been lamented by Potamianos et al (2003), stating “Unfortunately, the diverse algorithms suggested in the literature are difficult to compare, as they are rarely tested on a common audiovisual database” (p1307).

Saitoh and Konoshi (2006) achieved 40% for thermal video and 76% for standard video. However, they used 5 words and random guessing alone would have resulted in a 20% recognition rate. Also, the visemes in each word differed greatly which may have had an influence on results.

Yau (2008) claimed recognition rates of 99% from MTs using HMM and DCT. The primary reason such high recognition rates were achieved was the use of a custom built headset for data collection. This allowed accurate selection of the ROI without the need for extraction. Further experiments by Yau (2008) showed that the recognition rates greatly decreased when rotating the MT (36.79% at 20°), changing MT size (7.14% at 50%), moving the MT (51.43% when moved horizontally and vertically by 5 pixels), and occluding the MT (69.29% when occluded by 20 pixels). This illustrates the need for accurate ROI selection, and many of these factors are expected to have influenced the results achieved in this research.

While the results are lower than those obtained elsewhere, the lack of a common database makes a true comparison impossible. However, the results are in line with expectations and existing theory. The addition of the visual modality to ASR by combining standard video ASR and audio ASR shows an improvement in recognition rates confirming existing theory. The addition of a thermal video ASR to audio ASR also improves recognition rates. This is further improved when thermal video ASR, standard video ASR, and audio ASR are combined to create a trimodal AVASR system.

5.5 SECTION SUMMARY

An AVASR system was created for testing the effect of adding a third modality to AVASR. Mel-frequency Cepstral Coefficient (MFCC) based speech recognition was used for audio speech recognition. For visual recognition, the standard video and thermal video were processed using a modified form of Motion Templates combined with DCT for feature extraction. HMM was used for whole word classification and high level fusion weighted to the reliability of each stream for each test subject was used for combination. The testing showed an improvement in AVASR recognition rates when thermal video was added, successfully confirming that the addition of thermal video to AVASR can improve ASR performance.

6. CONCLUSION

This research has successfully shown that the addition of thermal video to AVASR can improve ASR performance by increasing recognition rates a relative 11.8% over audio and standard video combined ASR and a relative 38.2% over audio only ASR when averaged over all noise levels. However, the system is very basic and a great deal of research would need to be completed before thermal cameras could be used as part of an AVASR system in a real world environment.

Also proposed in this research was an alternate method of producing Motion Templates to improve their ability to track motion over time. While the results were not as great as hoped, a number of factors may have affected performance: the quality of data, the inconsistent ROI selection, and the choice of HMM and DCT over other methods to name a few. There are also many limitations to this research and a great deal of further research would be required before a trimodal AVASR system could be implemented in a real world environment.

Limitations and suggestions for further areas of research are as follows:

- As constructive research is interpretivist in nature, several different solutions to the problem can be arrived at and is iterative in nature. This means the solution arrived at was not expected to be a perfect solution, and there are many improvements to the algorithm that can be made including further pre-processing and post-processing of the images.

- Instead of complete word recognition, splitting the signal into visemes and phonemes may yield better results and would be recommended for large vocabulary speech recognition.
- While Gaussian white noise was added to the audio signal, this is not a true representation of the type of noise an ASR system would encounter in the real world. White Gaussian noise affects all frequency ranges whereas noise from natural sources will only affect a few frequencies. This means adding white noise greater impacted audio recognition than would happen naturally and the performance gains may not be different to those encountered in the real world.
- Although noise was added to the audio signal, no testing was done with the effect of changing lighting and temperature conditions. As the purpose of using thermal imaging cameras for automatic speech recognition is due to their invariance to illumination, further research in this area should be done.
- Due to the system being tested on a custom database that was created for the sole purpose of the research, the results cannot be readily compared with other research as different data was used. Further research could involve the development of an AV database which would enable researchers to compare results.
- While the use of DCT coefficients for data extraction and HMM for classification have been shown to be more efficient than other techniques in previous research, this research was based on different methods of pre-processing and does not necessarily hold true for use with MTs. Possible

further research would involve the exploration of different data extraction and classification methods .

- Further possible research could include the investigation of low level fusion for the signals. This would provide a greater integration of the data before entering the classifiers and may lead to improved results.
- The low resolution of the thermal camera may have impacted on results. The use of a thermal camera with a higher resolution sensor may further improve results. However these are vastly more expensive and could not be used here due to budget limitations.

A large hindrance to the further research and use of thermal cameras in AVASR is not only the cost, but also time. The process of collecting and editing data was very time consuming and took up a large portion of this project. To encourage further research, I recommend the development of an audio visual database which incorporates thermal video to enable researchers to readily develop alternate solutions and compare results.

While there are many limitations and a large amount of further research is required, the choice of constructive research means that multiple solutions may be arrived at, and is iterative in nature therefore it was not expected to arrive at the perfect solution. However, the solution arrived at has certainly shown that the addition of thermal video to create a trimodal AVASR system can lead to improved ASR performance.

REFERENCES

Appenrodt, Jorg, Al-Hamadi, Ayoub, Elmezain, Mahmoud, & Michaelis, Bernd. (2009). *Data gathering for gesture recognition systems based on mono color-, stereo color- and thermal cameras*. Paper presented at the 1st International Conference on Future Generation Information Technology, Jeju Island, Korea.

Baum, L.E., & Petrie, T. (1966). Statistical inference for probabilistic functions of finite state markov chains. *Annuals of Math. Statist., 37*(6), 1554-1563.

Becchetti, Claudio, & Ricotti, Lucio Prina. (1999). *Speech recognition: theory and C++ implementation* Chichester, New York Wiley.

Brookes, Mike. (1999). Voicebox: speech processing toolbox for MATLAB Retrieved May 25, 2012, from http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

Davis, K.H., Biddulph, R., & Balashek. (1950). Automatic recognition of spoken digits. *Acoustical Society of America*(24), 627-642.

de Villiers, M. R. . (2005). *Three approaches as pillars for interpretive information systems research: development research, action research and grounded theory.* Paper presented at the 2005 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists on IT research in developing countries, South Africa.

The Electromagnetic Spectrum. (2008). Retrieved May 15, 2012, from http://www.ccrs.nrcan.gc.ca/resource/tutor/fundam/chapter1/03_e.php

Giannakopoulos, Theodoros. (2010, September 24). A method for silence removal and segmentation of speech signals, implemented in Matlab Retrieved May 25, 2012, from http://www.mathworks.com/matlabcentral/fileexchange/28826-silence-removal-in-speech-signals

Goecke, Roland, & Millar, Bruce. (2004, October). *The Audio-Video Australian English Speech Data Corpus AVOZES.* Paper presented at the INTERSPEECH 2004 - ICSLP, Jeju Island, Korea.

Hermosilla, Gabriel, Ruiz-del-Solar, Javier, Verschae, Rodrigo, & Correa, Mauricio. (2006, October). *Face recognition using thermal infrared images for human-robot interaction applications: a comparative study.* Paper presented at the 2009 6th Latin American Robotics Symposium (LARS), Valparaiso, Chile.

Huang, Jing, Potamianos, Gerasimos, & Neti, Chalapathy. (2004). Audio-visual speech recognition using an infrared headset *Speech Communication, 44*, 83–96.

Juang, B.H., & Rabiner, Lawrence R. (2004). Automatic speech recognition – a brief history of the technology development. Retrieved from http://www.ece.ucsb.edu/Faculty/Rabiner/ece259/Reprints/354_LALI-ASRHistory-final-10-8.pdf

Kasanen, Eero, Lukka, Kari, & Siitonen, Arto. (1993). The constructive approach in management accounting research. *Journal of Management Accounting Research, 5*, 243-264.

Kong, S, Heo, J, Abidi, B, Paik, J, & Abidi, M. (2005). Recent advances in visual and infrared face recognition - a review. *The Journal of Computer Vision and Image Understanding, 97*(1), 103-135.

Lee, Bowon, Hasegawa-Johnson, Mark, Goudeseune, Camille, Kamdar, Suketu, Borys, Sarah, Liu, Ming, & Huang, Thomas. (2004, October). *AVICAR: Audio-Visual Speech Corpus in a Car Environment.* Paper presented at the INTERSPEECH 2004 - ICSLP, Jeju Island, Korea.

Lucey, P, Dean, D, & Sridharan, S. (2005). *Problems associated with current area-based visual speech feature extraction techniques*. Paper presented at the International Conference on Auditory-Visual Speech Processing (AVSP), Vancouver Island, British Columia, Canada.

Lucey, P., Martin, T., & Sridharan, S. (2004). *Confusability of phonemes grouped according to their viseme classes in noisy environments*. Paper presented at the 10th Australian International Conference on Speech Science & Technology, Sydney, Australia.

Martinez, B., Binefa, X., & Pantic, M. (2010, June). *Facial component detection in thermal imagery.* Paper presented at the Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on.

Millar, J Bruce, Wagner, Michael, & Goecke, Roland. (2004, October). *Aspects of speaking-face data corpus design methodology.* Paper presented at the INTERSPEECH 2004, Jeju, Korea.

Morris, Nigel J W, Avidan, Shai, Matusik, Wojciech, & Pfister, Hanspeter. (2007). *Statistics of infrared images* Paper presented at the IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR '07. , Minneapolis, Minnesota, USA.

Murphy, Kevin. (2005, June 8). Hidden Markov Model (HMM) toolbox for Matlab Retrieved May 25 2012, from http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html

Patterson, E. K., Gurbuz, S., Tufekci, Z., & Gowdy, J. N. (2002, May). *CUAVE: A new audio-visual database for multimodal human-computer interface research.* Paper presented at the Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on.

Petajan, E. D. (1984). *Automatic lipreading to enhance speech recognition.* Paper presented at the Global Telecomm. Conf., Atlanta, GA.

Potamianos, Gerasimos, Graf, Hans Peter, & Cosatto, Eric. (1998). *An image transform approach for HMM based automatic lipreading*. Paper presented at the International Conference on Image Processing, Chicago, IL.

Potamianos, Gerasimos, Neti, C., Gravier, G., Garg, A., & Senior, A. W. (2003). Recent advances in the automatic recognition of audiovisual speech. *Proceedings of the IEEE, 91*(9), 1306-1326. doi: 10.1109/jproc.2003.817150

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77*(2), 257-286. doi: 10.1109/5.18626

Romanos, Amelia. (2012, February 15). Greens call for faster solution for deaf MP, *The New Zealand Herald*. Retrieved from http://www.nzherald.co.nz/nz/news/article.cfm?c_id=1&objectid=10785641

Rothkrantz, L.J.M., Wojdel, Jacek, & Wiggers, Pascal. (2006). *Comparison between different feature extraction techniques in lipreading applications*, St Petersburg, Russia.

Saitoh, Takeshi, & Konishi, Ryosuke. (2006). *Lip reading using video and thermal images*. Paper presented at the SICE-ICASE International Joint Conference 2006 (SICE-ICCAS2006), Busan, Korea.

Scanlon, P. (2005). *Audio and visual analysis for speech recognition.* Doctor of Philosophy, University College Dublin.

Scanlon, P., & Reilly, R. (2001, 2001). *Feature analysis for automatic speechreading.* Paper presented at the Multimedia Signal Processing, 2001 IEEE Fourth Workshop on.

Shaneh, Mahdi, & Taheri, Azizollah. (2009). Voice command recognition system based on MFCC and VQ algorithms. *World Academy of Science, Engineering and Technology*(33), 534-538.

Shaw, M. (2001). *The coming-of-age of software architecture research.* Paper presented at the ICSE-2001., Los Alamitos, CA.

Socolinsky, D. A., & Selinger, A. (2004, June-July). *Thermal face recognition in an operational scenario.* Paper presented at the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. .

Sumby, W.H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Acoustical Society of America,* (26), 212-215.

Yau, Wai Chee. (2008). *Video analysis of mouth movement using motion templates for computer-based lip-reading.* Doctor of Philosophy, RMIT University.

APPENDIX A: INDIVIDUAL RESULTS

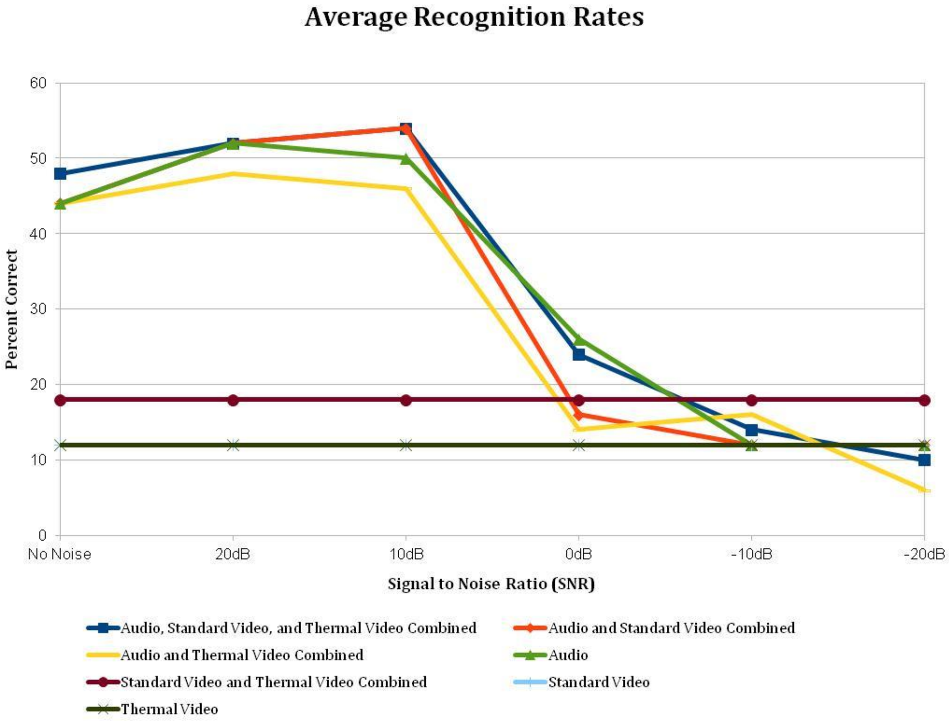


FIGURE 16 PARTICIPANT 1 RESULTS

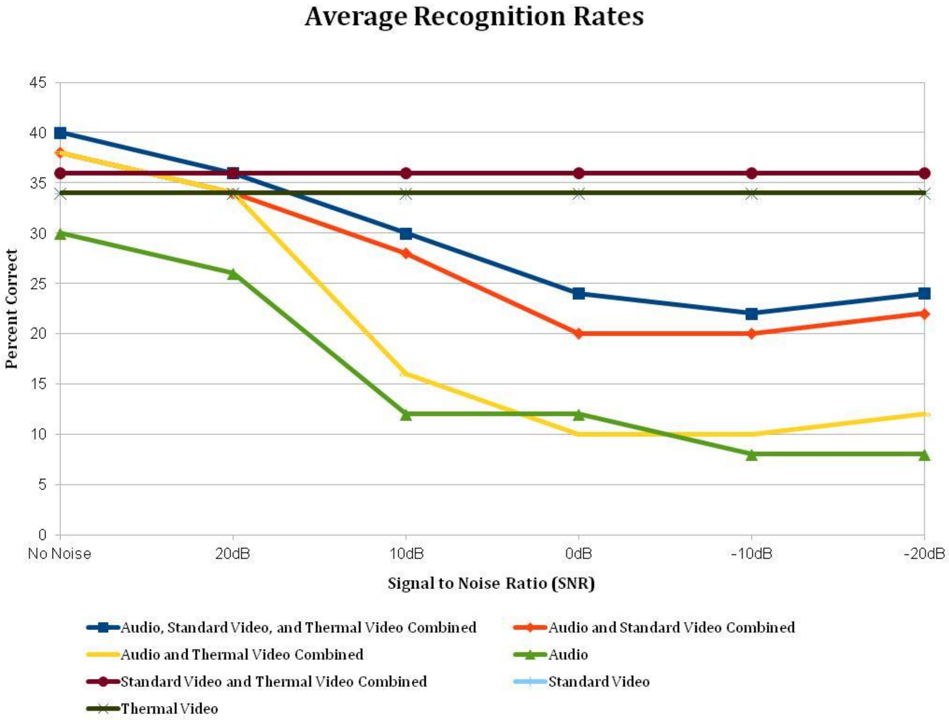


FIGURE 17 PARTICIPANT 2 RESULTS

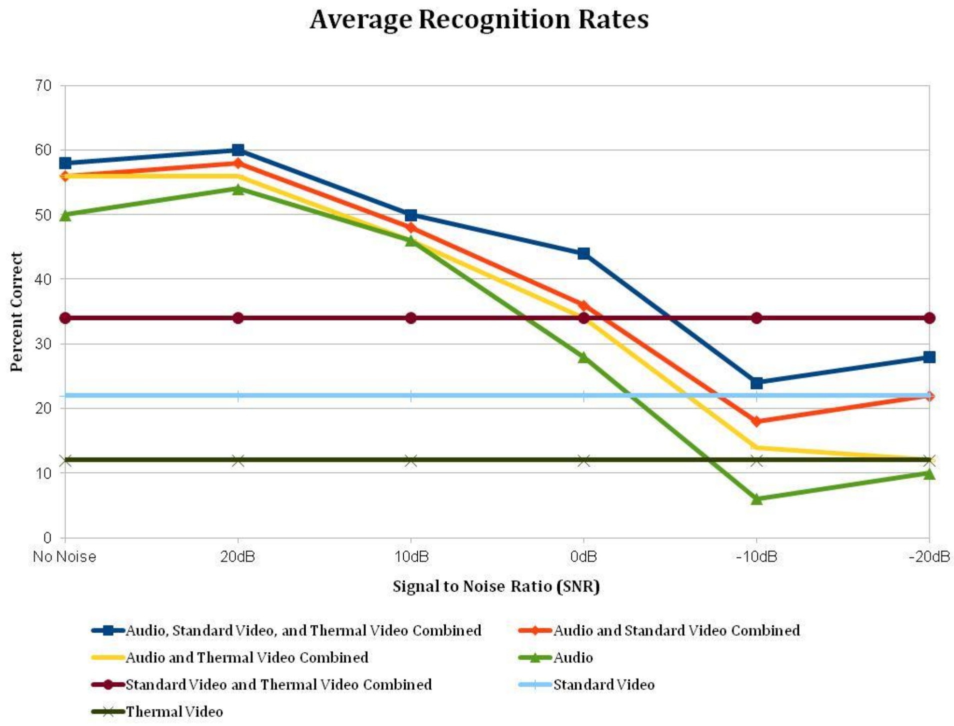


FIGURE 18 PARTICIPANT 3 RESULTS

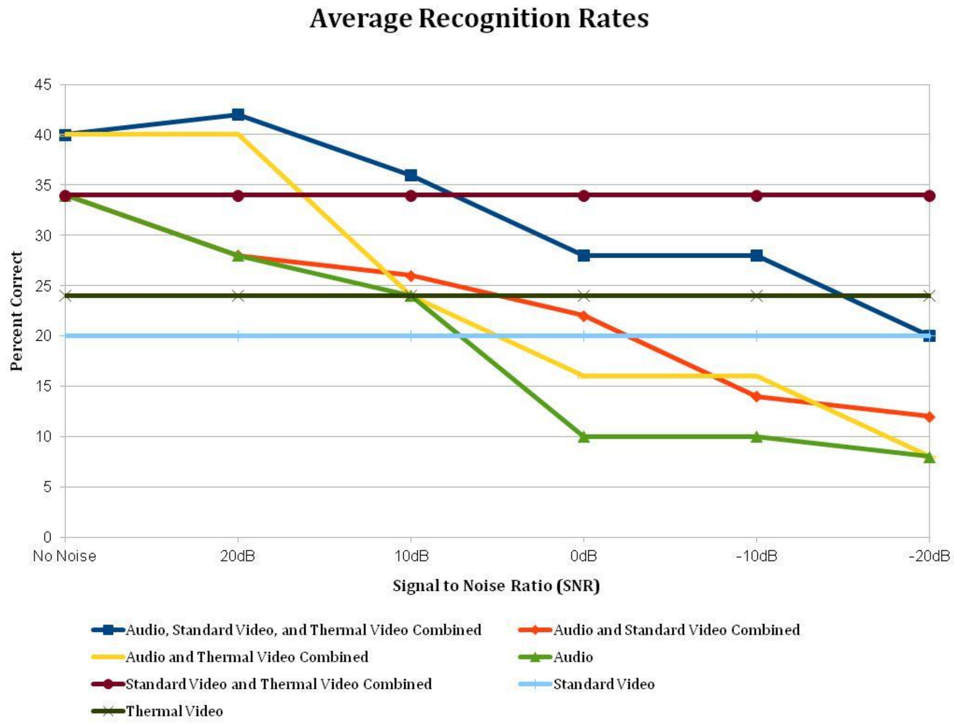


FIGURE 19 PARTICIPANT 4 RESULTS

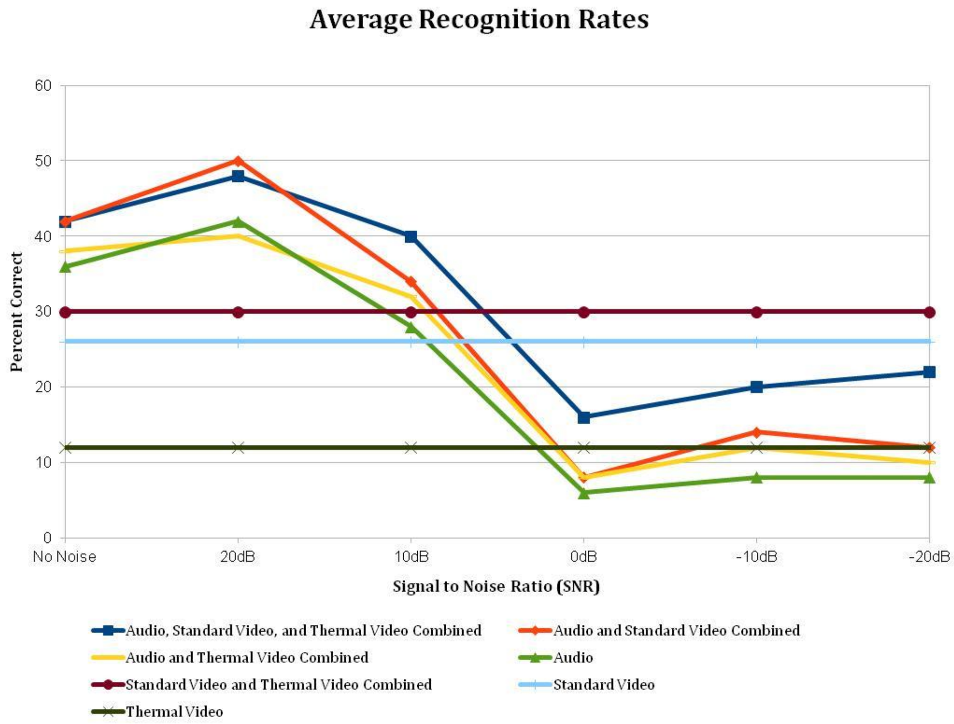


FIGURE 20 PARTICIPANT 5 RESULTS

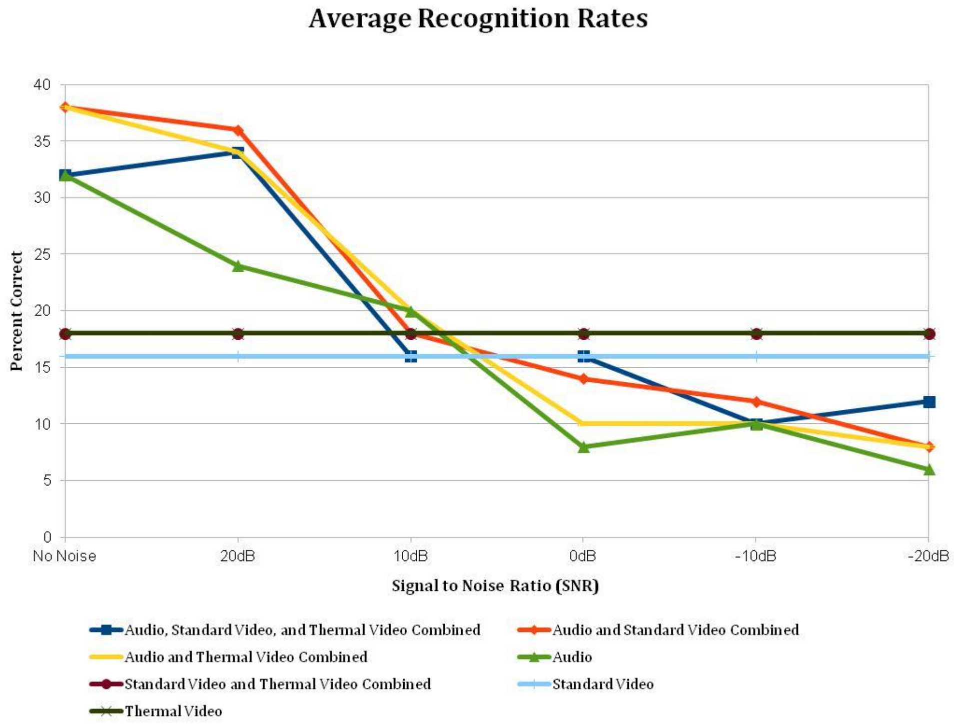


FIGURE 21 PARTICIPANT 6 RESULTS

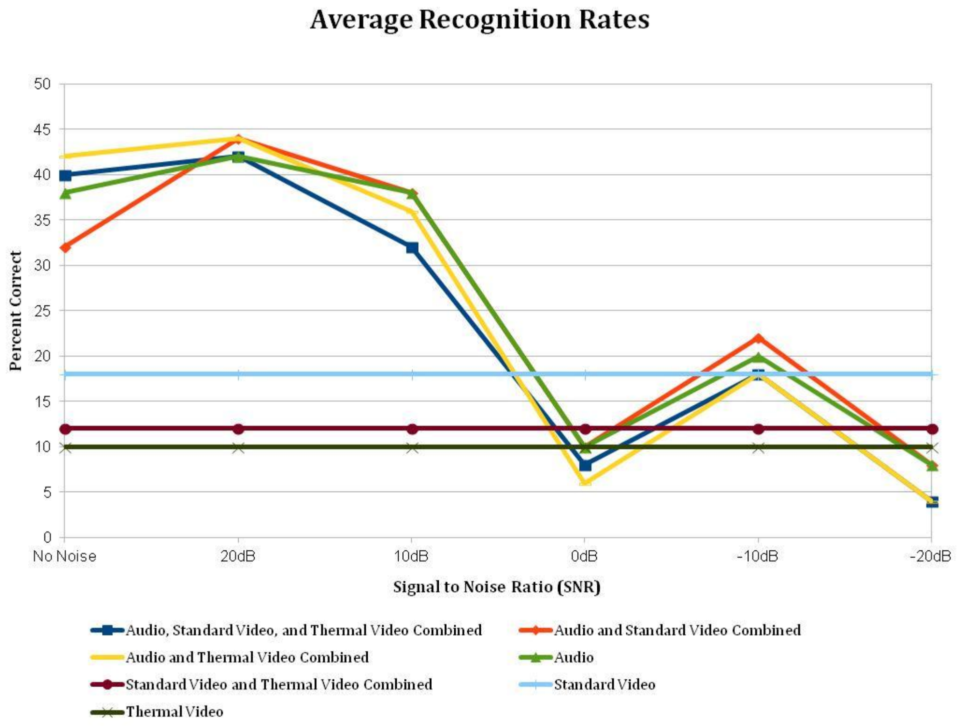


Figure 22 Participant 7 results

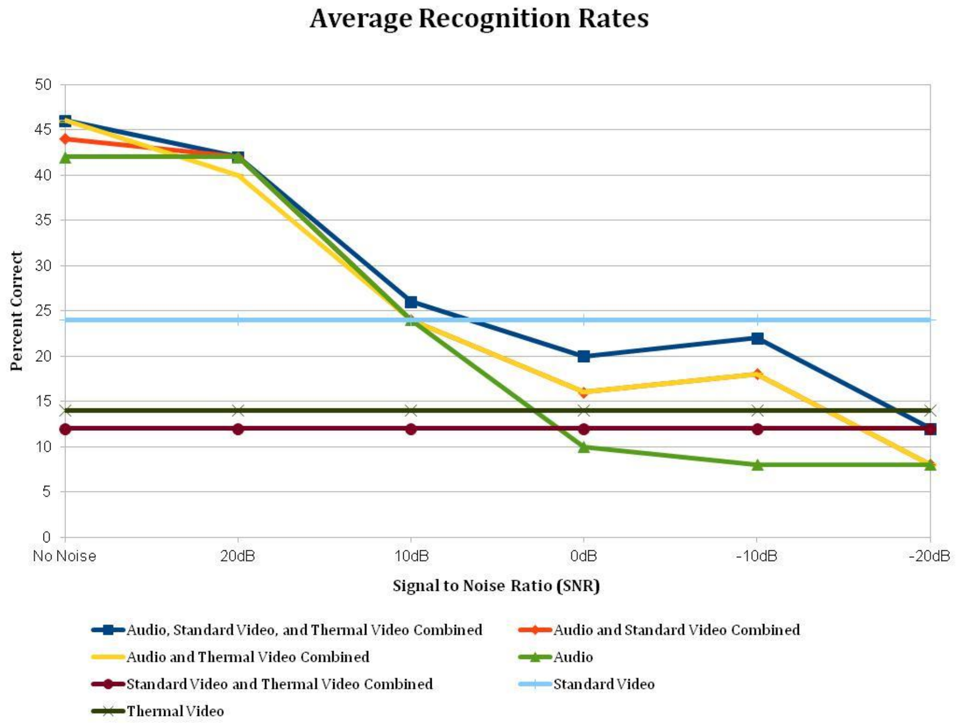


Figure 23 Participant 8 results

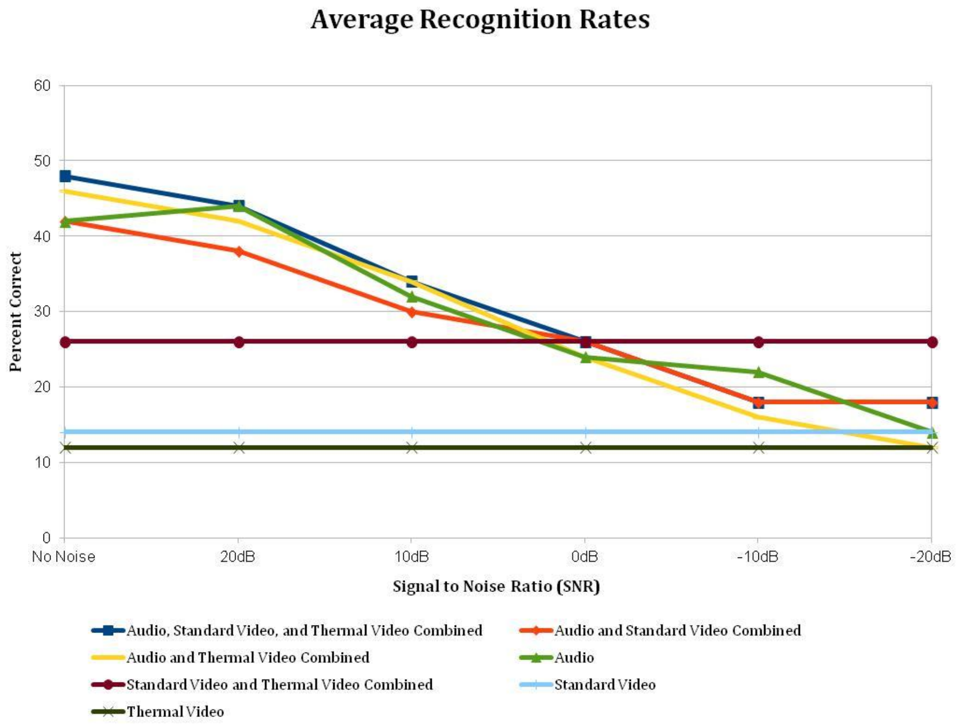


FIGURE 24 PARTICIPANT 9 RESULTS

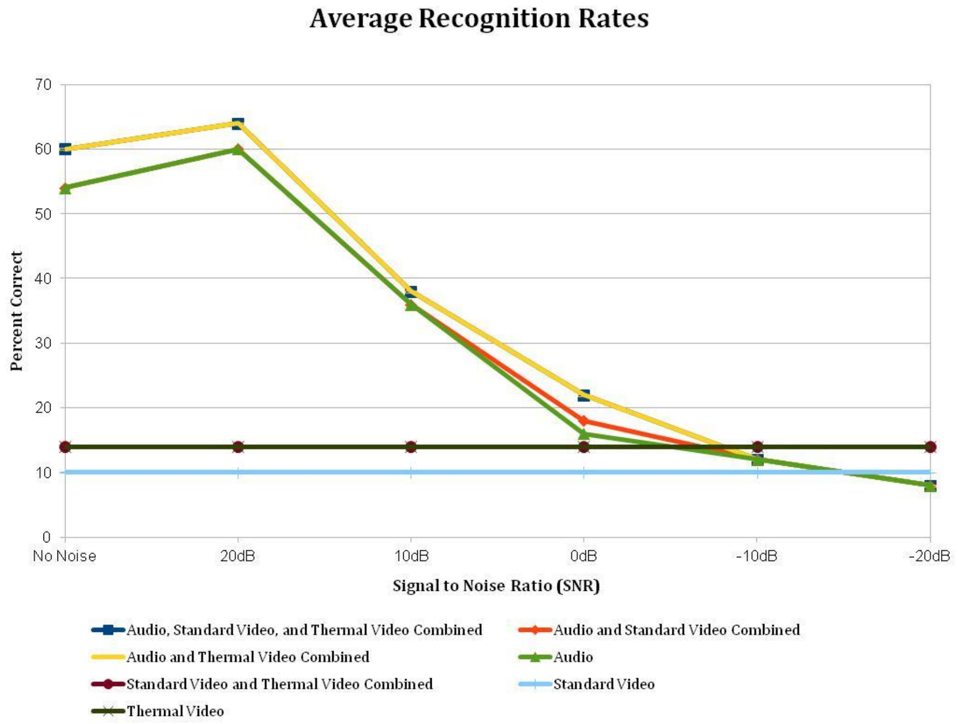


FIGURE 25 PARTICIPANT 10 RESULTS

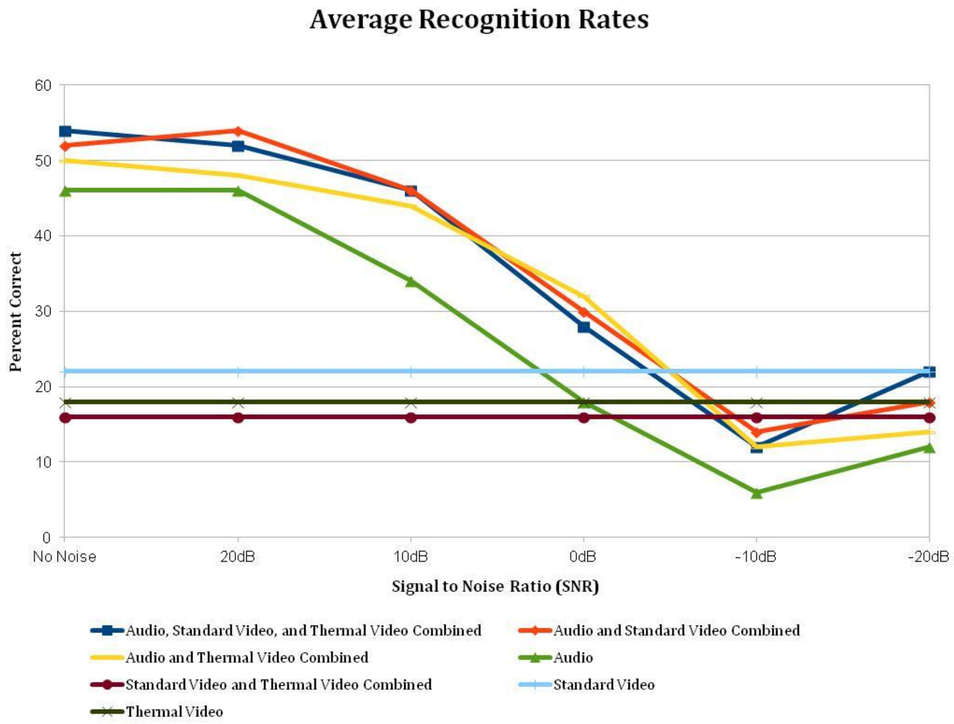


FIGURE 26 PARTICIPANT 11 RESULTS