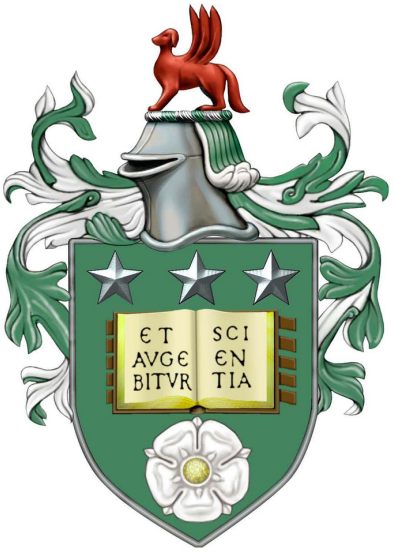
Audio-Visual Speech Processing for Multimedia Localisation

by

Matthew Aaron Benatan

Submitted in accordance with the requirements for the degree of Doctor of Philosophy



The University of Leeds
School of Computing
September 2016

Declarations

The candidate confirms that the work submitted is his/her own, except where work which has formed part of a jointly authored publication has been included. The contrinbution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

Chapter three extends work from the following publications:

Matt Benatan and Kia Ng. Cross-Covariance-Based Features for Speech Classification in Film Audio. *Journal of Visual Languages and Computing, volume 31, Part B, 215-221. 2015.*
Matt Benatan and Kia Ng. Cross-Covariance-Based Features for Speech Classification in Film Audio. *Proceedings of the 21st International Conference on Distributed Multimedia Systems, 72-77. 2015.*

The following publications contain early versions of concepts discussed in chapter five:

Matt Benatan and Kia Ng. Feature Matching of Simultaneous Signals for Multimodal Synchronization. *Proceedings of the 2nd International Conference on Information Technologies for Performing Arts, Media Access, and Entertainment, volume 7990 of Lecture Notes in Computer Science, 266-275. 2013.*
Matt Benatan and Kia Ng. Multimodal Feature Matching for Event Synchronization. *In Proceedings of the 19th International Conference on Distributed Multimedia Systems, 9-13. 2013.*

The candidate confirms that the above jointly authored publications are primarily the work of the first author. The role of the second author was editorial and supervisory.

Abstract

For many years, film and television have dominated the entertainment industry. Recently, with the introduction of a range of digital formats and mobile devices, multimedia's ubiquity as the dominant form of entertainment has increased dramatically. This, in turn, has increased demand on the entertainment industry, with production companies looking to increase their revenue by providing entertainment media to a growing international market. This brings with it challenges in the form of multimedia localisation - the process of preparing content for international distribution. The industry is now looking to modernise production processes - moving what were once wholly manual practices to semi-automated workflows.

A key aspect of the localisation process is the alignment of content, such as subtitles or audio, when adapting content from one region to another. One method of automating this is through using audio content as a guide, providing a solution via audio-to-text alignment. While many approaches for audio-to-text alignment currently exist, these all require language models - meaning that dozens of languages models would be required for these approaches to be reliably implemented in large production companies. To address this, this thesis explores the development of audio-to-text alignment procedures which do not rely on language models, instead providing a language independent method for aligning multimedia content. To achieve this, the project explores both audio and visual speech processing, with a focus on voice activity detection, as a means for segmenting and aligning audio and text data.

The thesis first presents a novel method for detecting speech activity in entertainment media. This method is compared with current state of the art, and demonstrates significant improvement over baseline methods. Secondly, the thesis explores a novel set of features for detecting voice activity in visual speech data. Here, we show that the combination of landmark and appearance-based features outperforms recent methods for visual voice activity detection, and specifically that the incorporation of landmark features is particularly crucial when presented with challenging natural speech data. Lastly, a speech activity-based alignment framework is presented which demonstrates encouraging results. Here, we show

that Dynamic Time Warping (DTW) can be used for segment matching and alignment of audio and subtitle data, and we also present a novel method for aligning scene-level content which outperforms DTW for sequence alignment of finer-level data. To conclude, we demonstrate that combining global and local alignment approaches achieves strong alignment estimates, but that the resulting output is not sufficient for wholly automated subtitle alignment. We therefore propose that this be used as a platform for the development of lexical-discovery based alignment techniques, as the general alignment provided by our system would improve symbolic sequence discovery for sparse dictionary-based systems.

Acknowledgements

Firstly I would like to thank Dr. Kia Ng - my supervisor and my friend. Working with you over the past seven years has been enormously inspiring. Thank you for your mentorship, encouragement, and for teaching me the value of pursuing crazy ideas (and for the wisdom to know when the ideas are perhaps a little too crazy).

Thank you also to Andy Bulpitt for taking on supervision during the concluding months of this PhD. You have done a superb job of picking up the project at the last minute, and I'm very grateful for the guidance you've provided through this crucial time. I would also like to thank my co-supervisors Derek Magee and Katja Markert. Our conversations were always very valuable, and helped me to expand my understanding of, and interest in, the field I have chosen to pursue.

To my friends at The University of Leeds - Sam, Alicja, Luke, Bernhard, Leroy, Christian, Dan, Olly, Jen, Aryana and all in the School of Computing. You each contributed to making my time at Leeds thoroughly enjoyable. Thank you for the fascinating discussions, the evenings spent at the pub, and for helping to keep me sane throughout this PhD.

I would also like to thank the EPSRC and ZOO Digital - the collaboration between academia and industry has been hugely valuable, and I am very grateful to both for providing the financial support required for this work. Thank you also to Stuart Green for your commitment, support and guidance throughout this project.

Thank you also to Duke University and the University of Leeds for granting permission to use their audiovisual content within this work.

To my close friends Rob and Chris - from all those years playing in bands, to making wacky videos, to simply geeking out over lengthy sessions of tech-talk or boardgames. You have each been instrumental in the development of my interests - both creatively and technologically - and I'm hugely thankful for your support, guidance and continuing friendship. Thank you also to my long standing friend, Dr. Thomas Hazlehurst, for the lunch dates, the LaTeX tuition and for your incredible beard.

While rewarding, the past few years have also been incredibly demanding, both intellectually and emotionally. This experience would have been far more difficult without the tremendous dedication and support of my best friend and partner, Rebecca, who has kept me on a steady path throughout the ups and downs of PhD life. I am deeply grateful to you for your continuing love, support and patience.

Lastly, thank you to my parents, Dan and Debby. Your tremendous support and encouragement throughout the years has motivated me to continue challenging myself to embark on new and interesting pursuits. Without your support, I certainly wouldn't have engaged in such a rewarding and valuable journey.

Contents

List of Figures

List of Tables

List of Abbreviations

AAM - Active Appearance Model

ACM - Active Contour Model

ADR - Automatic Dialogue Replacement

ALED - Adaptive Linear Energy-Based Detection

ASM - Active Shape Model

ASR - Automatic Speech Recognition

AUC - Area Under the Curve

EER - Equal Error Rate

CD-DNN - Context Dependent Deep Neural Network

CLM - Constrained Local Model

CNN - Convolutional Neural Network

CWT - Continuous Wavelet Transform

DCT - Discrete Cosine Transform

2D-DCT - Two Dimensional Discrete Cosine Transform

DFT - Discrete Fourier Transform

DNN - Deep Neural Network

DP - Dynamic Programming

DTW - Dynamic Time Warping

DWT - Discrete Wavelet Transform

EM - Expectation Maximisation

FST - Finite State Transducer

FN - False Negative

FP - False Positive

GMM - Gaussian Mixture Model

VBGMM - Variational Bayes Gaussian Mixture Model

HTK - Hidden Markov Model Toolkit

HLBP - Haar Local Binary Patterns

HMM - Hidden Markov Model

HOG - Histogram of Oriented Gradients

LDA - Linear Discriminant Analysis

LBP - Local Binary Patterns

LP - Linear Predictor

LPC - Linear Predictive Coefficient

LOWESS - Locally Weighted Scatter-Plot Smoothing

LTSD - Long Term Spectral Distance

MCMC - Markov Chain Monte Carlo

MCMCDA - Markov Chain Monte Carlo Data Association

MFCC - Mel Frequency Cepstral Coefficient

MFCC CC - Mel Frequency Cepstral Coefficient Cross Covariance

MLLT - Maximum Likelihood Linear Transform

NN - Neural Network

NTSC - National Television System Committee

PAL - Phase Alternating Line

PCA - Principal Component Analysis

RASTA-PLP - Relative Spectral Transform Perceptual Linear Prediction

RBF - Radial Basis Function

ROC - Receiver Operating Characteristics

ROI - Region of Interest

RNN - Recurrent Neural Network

SECAM - Séquential Couleur à Mémoire

SFD - Spectral Flatness Detection

SNR - Signal to Noise Ratio

SOLA - Synchronised Overlap Add

STFT - Short Time Fourier Transform

SVM - Support Vector Machine

TN - True Negative

TP - True Positive

VAD - Voice Activity Detection

V-VAD - Visual Voice Activity Detection

VoIP - Voice Over IP

WSOLA -Waveform Similarity Synchronised Overlap Add

Chapter 1

Introduction

For many years, film and television have been the primary forms of entertainment multimedia. These industries have continued to grow with the advent of the internet and the ever increasing number of digital multimedia formats. The ubiquity of digital devices across the globe means that the international market is now larger than ever before - presenting a significant opportunity to multimedia production companies. As such, interest in tapping into these markets has grown, with more and more companies competing to provide content localisation services to adapt content for international distribution. With this competition comes the desire to complete projects more efficiently, as time requirements form a key factor in the selection of localisation services.

One of the bottlenecks in the localisation process is the reliance on manual processes for translation and adaptation. As such, this work looks to develop methods to automate parts of the localisation workflow, in order to reduce cost and time requirements for multimedia localisation. To do so, the project investigates both audio and visual voice activity detection, before exploring language-independent methods for multimedia content alignment. First, we introduce some concepts in film post production workflows and discuss how they may benefit from automation.

1.1 Overview of Film Post Production Workflows

This section gives an overview of the film post-production workflows relevant to this work. These are Automatic Dialogue Replacement (ADR) and subtitle localisation. Both are part of the localisation process. Multimedia localisation is defined as the process of preparing content for international distribution. This includes the process of subtitling and dialogue adaptation as well as the adaptation of promotional materials such as posters, trailers, etc.

Format	Frame Rate
PAL	25 fps
NTSC	24 fps
SECAM	25 fps

Table 1.1: Common visual multimedia formats and corresponding framerates.

1.1.1 Automatic Dialogue Replacement

ADR is a crucial component for both multimedia localisation and film post-production in general. In the case of general film post-production, the original dialogue is first recorded on set (on a soundstage). This initial recording is often noisy and of insufficient quality for the final product [67]. As such, a separate stage of recording - ADR - takes place, in which the actors rerecord their lines in a sound treated studio to obtain high quality recordings. Despite its name, ADR has not traditionally incorporated automatic processes. As such, it typically relies heavily on manual editing - with an audio engineer being responsible for overseeing the recording process, and for aligning the new dialogue recordings to the source video.

Over the past decade manufacturers of post-production software have begun to introduce automated tools for the ADR process. These involve automatic speech alignment tools, such as those developed by Adobe [137], which automatically align new dialogue recordings to the reference audio from the original recording. While this has proven to be successful, it typically only works for brief, clean pieces of dialogue. This is as noise can have a detrimental effect on the alignment, and in certain cases the original signal is too noisy to achieve any alignment whatsoever [67]. As such, a noise-robust solution would be attractive. One such approach would be audio-to-video alignment, which would use the video content as a reference for the subsequent audio alignment. In this way, the issue of noise in the source audio is overcome by using visual information.

1.1.2 Format Conversion

As well as modifying the content itself for subtitling or adaptation, localisation also involves converting multimedia formats to comply with region-specific standards. Three common formats used for entertainment content are PAL, NTSC and SECAM. Each format has different specifications for frame rate and picture resolution. For the purpose of multimedia alignment, framerate is the most critical of these features. Table 1.1 gives an overview of common formats and their respective framerates.

While NTSC is technically a 30 fps format, it has an effective frame rate of 24 fps due

to a processing technique termed 3:2 pulldown [54]. This is used to correct the 30 fps framerate, which is used to comply with 60 Hz vertical scanning frequency, by manipulating the way frames are distributed into video fields. 3:2 pulldown works by transmitting the first frame for 3 video fields, and the subsequent frame for 2 video fields. As such, 2 frames are transmitted for every five video fields, resulting in an average of 2.5 video fields per frames. Given a 60 Hz refresh rate ,this results in a framerate of $60 \div 2.5 = 24$ fps.

While Table 1.1 shows the most popular current framerates, an increase in types of multimedia content and a growing variety of digital platforms means the number of formats is increasing. Currently, there are over 10 different formats being used for visual media content [37]. This means that localisation could involve re-scaling from any one to any other of these formats. As such, automatic re-scaling of content would be advantageous, as it would not require prior information, and could be easily applied to uncommon or previously unencountered formats.

1.1.3 Subtitle Localisation and Dialogue Adaptation

Subtitling is the process of adding closed-captions to a piece of multimedia in preparation for either domestic or international distribution. In the case of domestic distribution, the subtitles contain an approximation of the original dialogue, modified for readability. This is done as often the dialogue sections will be fairly extensive, thus a shortened approximation of the dialogue is provided to ensure that the resulting subtitles can be read easily. In the case of international distribution, a similar process is followed (ensuring that the subtitles are easy to read), but the dialogue also undergoes a separate process - adaptation.

Dialogue adaptation is the process of adapting the source dialogue to a destination language, and is used in both subtitle and audio localisation (e.g. dubbing). Part of this process is simply the translation of content from one language to another, however straightforward translation is not sufficient for high quality content. This is partially due to phonological composition - for example, a phrase with identical meaning in English and German would have a very different phonetic composition, and thus the German dialogue would not fit with the English mouth movements on screen. Adaptation is also responsible for preserving cultural values in the content, such as by adapting the content of jokes or other material to identify more closely with the target audience.

The process of subtitle localisation often requires significant time investment by a language expert, who is not only capable of translating, but of skilfully adapting the content to suit its target audience. One of the time consuming tasks during adaptation is the process of watching a film and marking up areas containing speech. This process could

thus be improved through the use of automatic speech detection, which could save time by informing the translator as to where they should focus their attention. This would reduce time requirements by automatically doing a 'first pass' over the material, which in turn would reduce the overall cost of localisation. This could be further enhanced by automatically identifying associations between segments of subtitles and segments of film. This would allow the translator to easily move between semantically-related excerpts. It would also mean that misaligned subtitles could be used, as associations could be found automatically, thus this would not rely on the original resources being aligned. This would be particularly useful given the increasing variety of formats [37].

Subtitle localisation also involves modifying subtitle timestamps to correspond to the destination format, which often has a different timebase. As such, an automated method for rescaling timestamps would be beneficial. This would involve audio-to-text alignment, for which there are already a number of methods, such as those developed by Katsmanis *et al.* [65] and Goldman *et al.* [41]. The difference here is twofold:

1. As localisation involves working with many different languages, a language independent approach would be advantageous, as it would not require a multitude of language models.

2. The existing methods have been developed for use on clean audio, and do not tackle the challenges presented by entertainment media, thus an approach robust to entertainment media content would be beneficial.

1.2 Motivation and Contributions

The aim of this work is to develop a system for subtitle and audio track alignment which does not require a language model. This is attractive for post-production processes as, currently, a majority of alignment tasks are undertaken manually. Furthermore, existing approaches for audio-to-text alignment, such as those developed by Katsmanis *et al.* [65] and Goldman *et al.* [41], require language models. This greatly limits their potential applications, however it has been shown that some language models can be successfully applied for cross-lingual alignment tasks [77]. These approaches are still fairly limited, as they rely on languages sharing a significant degree of phonological content. As such, current methods require annotated training data provided by experts, which can be costly and time consuming to obtain (particularly when considering that these are required for many languages). Furthermore, speech recognition has demonstrated highly variable performance on noisy audio, such as the audio present within film and television. Given that

multimedia production companies deal with a broad range of both content and languages (particularly where localisation is concerned), it would be beneficial to have a one-fits-all method for content alignment. This alignment would ideally be both language independent and robust to the noise conditions present in entertainment media. Such a solution would be desirable as either a fully automated or semi-automated solution for entertainment media post-production workflows.

This work looks to realise such a solution through the combination of audio and visual voice activity detection, and speech-to-text alignment methods. The fundamental concept here is that, through leveraging broad features from the content - such as patterns of speech activity - a language independent method for multimedia alignment can be developed. This also addresses the problem faced by speech recognition given noisy media, as this work focuses on detecting speech activity, rather than on extracting finer semantic content. This reduces the complexity of the classification problem - a logical step given the non-trivial task of speech/non-speech discrimination in complex mixed audio signals.

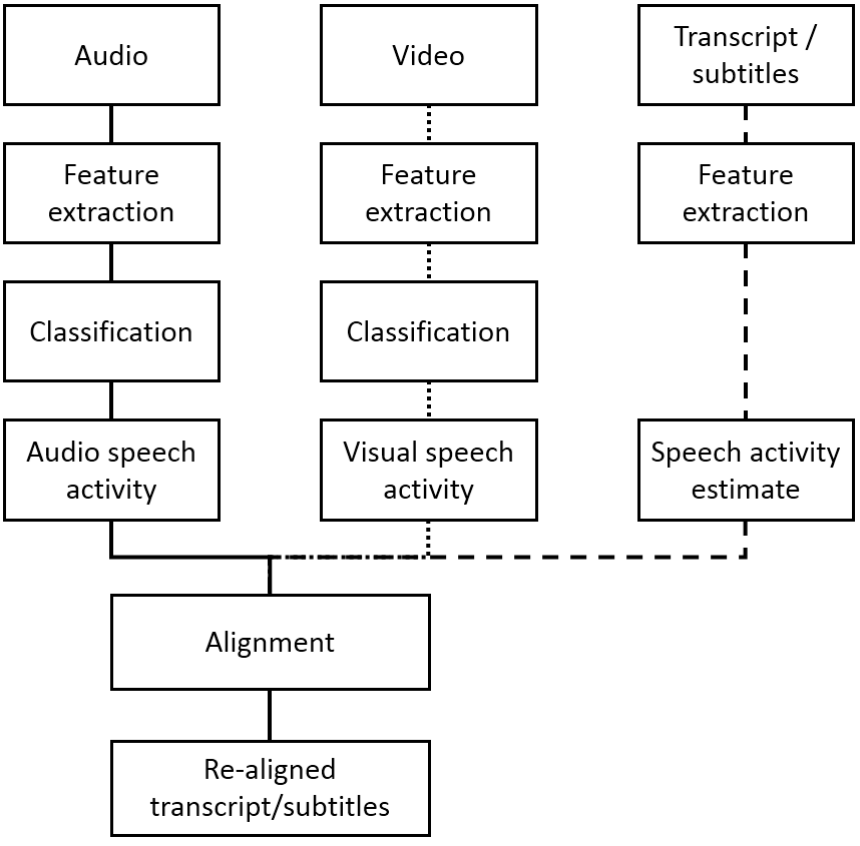


Figure 1.1: Diagram of speech activity-based text-to-audio alignment.

The alignment process presented in Figure 1.1 focuses on aligning speech activity patterns with subtitle data. This is achieved through a novel alignment method comprising a combination of general signal alignment and incremental segment-wise alignment. This approach is made possible by a number of segmentation and alignment strategies developed through this work. These are used to automatically identify and associate segmentation points (anchors) within the data, which can then be used to estimate a scaling coefficient. This is then applied to re-scale the time base of the source material to match that of the destination material.

The thesis also introduces a novel method for detecting visual speech activity in challenging natural speaker data. This is achieved through applying state-of-the-art landmark localisation methods to extract facial landmarks. These are then combined with appearance-based features, and used to train a binary speech/non-speech classifier. The visual speech activity information is later explored for use within the alignment process as a means of confidence scoring, and is also investigated for audio-to-visual speech alignment.

To summarise, three core contributions are presented in this thesis:

- **Audio Voice Activity Detection in Entertainment Audio** This work presents a novel method for audio voice activity detection within film and television media. The method is based on the creation of Mel Frequency Cepstral Coeffient (MFCC) Cross-Covariance features (MFCC-CC), and demonstrates competitive results, outperforming state-of-the-art approaches for audio voice activity detection in entertainment media.

- **Visual Voice Activity Detection for Natural Speech Conditions** An approach for detecting voice activity using visual speech features is presented. The approach combines both landmark and appearance-based features, and demonstrates strong performance on natural speech data, outperforming contemporary methods. Furthermore, this work provides a detailed evaluation of how appearance and landmark-based features perform individually, and shows that significant performance gain can be achieved through utilising landmarks obtained via state-of-the-art landmark localisation methods. This is particularly the case when presented with challenging natural speech data, such as data containing dynamic speaker movements, variable illumination and partial occlusions.

- **Language Independent Subtitle Alignment** A method for utilising audio speech activity and subtitle data for content alignment is presented. This is achieved through leveraging speech activity patterns in the audio, video and textual content. These patterns are used to segment the content, after which DTW and other dynamic

programming-based algorithms are used to estimate the scaling coefficient required to align the data.

This work also produced two datasets to facilitate the development of voice activity detection methods. These are:

- **Feature-Film Annotations Dataset** This dataset comprises annotations of six feature-films - The Bourne Identity, Kill Bill Volume 1, I Am Legend, Saving Private Ryan, Disney's Hercules and The Fellowship of the Ring. Each film has been carefully segmented into speech and non speech content. This process involved two passes by two separate annotators to ensure optimal labelling.
- **Natural Speech Dataset** This dataset comprises 7 videos of 7 different speakers in natural speech conditions, totalling 105 minutes of speaker data. The dataset contains numerous examples of challenging visual speech conditions, including partial occlusions, variable illumination and dynamic speaker movements. This has been carefully annotated into speech and non-speech components.

1.3 Thesis Overview

This thesis explores audio and visual methods for voice activity detection, and applies them within an alignment framework specifically designed not to rely on a language model. The thesis begins with an overview of crucial related technologies in Chapter 2, where the underlying processes for audio and visual feature extraction are introduced. A range of recent methods for multimedia synchronisation are also presented, in which several sequence alignment strategies are presented.

Chapter 3 discusses the development of a novel set of features for audio voice activity detection in entertainment media. These features are tested using two machine learning algorithms: support vector machines and random forests. We show that competitive performance can be achieved using these features, outperforming state-of-the-art voice activity detection algorithms on a dataset of feature film material.

In Chapter 4, we explore the use of a combined feature set for visual voice activity detection. These features combined state-of-the-art landmark localisation methods with popular appearance-based features to create a featureset robust to challenging speaker data. This chapter demonstrates that, while appearance based features are useful, landmark features are particularly crucial for achieving good performance on challenging data. Furthermore, we demonstrate that combining both landmark and appearance-based features achieves the best performance overall.

A strategy for language independent alignment is presented in Chapter 5, where speech activity patterns in audio and text information are used for associating and aligning audio and text data. Visual voice activity is also explored for audio-to-visual alignment, and as a means of improving audio-to-text alignment. We demonstrate that, while it is possible to improve on the original misaligned data, inaccuracies in the data stand in the way of wholly accurate alignment.

The thesis concludes with an overview of the work and a discussion of potential directions for future development.

Chapter 2

Background

This chapter reviews literature concerning three core project themes: audio speech processing, visual speech processing, and feature matching and alignment. For each theme, an introduction to key developments is presented, followed by discussion on how the technologies explored relate to the key goals of this work, and how the techniques discussed influenced project development.

2.1 Audio Speech Processing

Here, a number of approaches for processing speech audio are explored, focusing largely on developments in audio speech feature extraction and Voice Activity Detection (VAD).

2.1.1 Audio Speech Feature Extraction

Speech feature recognition and classification has been of scientific interest since the 1970s [7][58]. A crucial component in all speech classification systems developed since is the extraction of audio features. The choice of features can have a significant impact on the overall system, with certain features enhancing the discrimination capability of the classification method. The features discussed here were selected due to their prevalence in speech processing literature.

Linear Predictive Coding

Linear predictive coding (LPC) is based on work in communications signals carried out in the mid 20th century [114][135][34] that was further developed and applied within speech compression in the 1960-70s [5][84][85]. The primary concept behind LPC is that a sample of a discrete-time signal can be predicted as a linear combination of past signal

values [30]. The model for LPC is based on an approximation of the vocal tract whereby speech production is described as the product of a tube of varying diameter. Variation in the diameter of the tube results in differing resonant frequencies which are analogous to voiced speech formants. The model also incorporates white noise components which represent sibilants and plosives.

Through estimating the formants and applying inverse filtering [30], the speech signal can be analysed to produce coefficients which describe the amplitude, frequency components, formants and the residue signal (the remnant of the signal following inverse filtering). This data can then be synthesized using LPC by reversing the process; applying the residue signal to a filter determined by the formant information, resulting in an approximation of the original signal.

While LPC was initially developed for signal compression, it has since been used for a variety of speech processing tasks including ASR [57], speech activity detection [93] and other voicing recognition tasks [19].

RASTA-PLP

RASTA-PLP, or Relative Spectral Transform - Perceptual Linear Prediction, is a method of speech feature extraction introduced in work by Hermansky *et al.* [53]. The underlying principle of this method is to minimize the difference between speakers while preserving important information, such as phonetic content, which can be used for speech processing tasks. In this way, the method is able to improve speaker independent application of speech processing techniques such as speech recognition [70] and speech detection [36].

To obtain RASTA-PLP features, the signal is first analysed to obtain the critical-band power spectrum, before the spectral amplitude is transformed via a compressing static nonlinear transformation. The time trajectory of each transformed spectral component is then filtered and transformed through an expanding static nonlinear transformation. The equal loudness curve is then multiplied and raised to simulate the response of the human auditory system. An all-pole model of the resulting spectrum is then computed following the conventional PLP technique [52].

In Hermansky *et al.*'s RASTA-PLP paper, the RASTA-PLP approach is evaluated against the original PLP approach described in their earlier work [52] and exhibits a considerable advantage for speaker independent continuous speech tasks. Since their development, RASTA-PLP features have gone on to be used in a number of speech processing tasks. From the perspective of this work, the most crucial application of RASTA-PLP is in voice activity detection, for which they have achieved encouraging results when combined with neural network-based classifiers [36].

Short Time Fourier Transforms

Short time Fourier transforms (STFTs) are crucial for many audio processing tasks. These are obtained by applying a sliding window to an audio source and obtaining the Fourier transform of the segment within the window. This is expressed as:

$$X_m(\omega) = \sum_{n=-\infty}^{\infty} x(n)w(n - mR)e^{j\omega n} \tag{2.1}$$

where $X_m(\omega)$ is the Discrete Time Fourier Transform (DTFT) for a given input signal $x(n)$ with window function $w(n)$ and hop size R (in samples) between DTFTs. The principal here is that, while audio spectra vary considerably over time, variance is minimal within sufficiently small segments. As such, obtaining the STFT at regular intervals over time provides a good evaluation of the audio content at discrete intervals, while also allowing changes over time to be analysed by observing all of the segments in the sequence. STFTs are used as the basis for many audio processing tasks as these provide the basis for the extraction of other features, such as Mel Frequency Cepstral Coefficients (MFCCs). These are therefore used in a broad variety of speech processing tasks, including VAD [119] and audio source separation [8].

Mel Frequency Cepstral Coefficients

MFCCs were initially developed for use within ASR systems [87], and have since been used widely within speech processing applications such as ASR, VAD, and other audio classification tasks. MFCCs are obtained by first taking the Fourier transform of an audio frame and mapping the resulting power spectrum onto the Mel scale - a non-linear scale based on the human auditory response curve [87][78].

Once the power spectrum has been mapped to the Mel scale, the log power at each of the Mel frequency filterbanks are obtained. A Discrete Cosine Transform (DCT) is then applied to the Mel frequency filterbanks, and the MFCCs are the coefficients resulting from the DCT. Typically, the first 13 DCT coefficients are used for speech processing tasks.

Since their development, MFCCs have become increasingly popular for speech processing applications, often chosen over the LPC features used in earlier speech processing work. This has resulted in a broad range of audio perception applications using MFCCs, including music information retrieval [126], music modeling [35], ASR [87] and VAD [69].

2.1.2 Audio Voice Activity Detection

This section explores a number of methods for voice activity detection, covering both early approaches such as statistical VAD methods, as well as recent approaches which use more sophisticated machine learning-based methods. A range of different VAD approaches for entertainment multimedia are explored, including approaches which have been applied to feature film content.

Early Approaches for Voice Activity Detection

Earlier approaches for voice activity detection were developed for communications technologies such as VoIP [100]. In many of these cases, detecting speech activity was a relatively straightforward task, involving the separation of speech from background noise. As such, these earlier systems primarily involved statistical operations for evaluating changes in the signal. One such method, described in Sakhnov *et al.*'s work [109], involves evaluating the current frame energy with respect to the energy of silence frames, as given by:

$$\begin{aligned} if \quad (E_i > kE_{silence}) \quad where \quad k > 1 : Frame \quad is \quad ACTIVE \\ else \quad Frame \quad is \quad INACTIVE \end{aligned} \tag{2.2}$$

where $E_{silence}$ is the mean background noise, E_i is the energy of the frame at point i and k is user definable, allowing a safe band for the adaptation of the threshold to account for varying signal to noise ratio (SNR) of the signal. Due to the static threshold, this approach proved to be insensitive to varying speaker dynamics.

To improve upon this, a dynamic approach, adaptive linear-energy based detection (ALED), was proposed [110]. This involved varying the threshold according to statistical information of speech dynamics; using the fall time of syllables to dynamically alter the threshold value. This improved upon the previous approach through the reduction of phoneme clipping, proving to be more sensitive to varying speaker dynamics.

The use of spectral content to improve voice activity detection has also been proposed [100]. The method discussed by Prasad et al. is spectral flatness detection (SFD) the process of classifying speech or noise segments according to the variation of their spectral content. The spectrum is first computed using the DCT, after which the spectral variance (the energy variance within the frame across frequency) of each DCT frame is calculated. The frames are then classified via:

$$\begin{aligned} if \quad (\sigma_i > \sigma_{th}) : Frame \quad is \quad ACTIVE \\ else \quad Frame \quad is \quad INACTIVE \end{aligned} \tag{2.3}$$

Where σ_{th} is the variance threshold, determined by the spectral variance of the noise content, and σ_i is the spectral variance of the current frame. To account for varying speaker dynamics, this is also dynamically varied according to speech dynamics data, by:

$$\sigma_{th\_new} = (1 - p)\sigma_{th\_old} + p\sigma_i \tag{2.4}$$

where σ_{th_new} is the new threshold, σ_{th_old} is the previous threshold and p is used to scale the threshold according to speech dynamics. This approach proved to be more successful in low signal to noise ratio conditions due to its use of spectral information when compared to the energy-based algorithms described above [100].

Prasad *et al.* [100] go on to discuss comprehensive VAD (CVAD) a VAD algorithm that uses decision rules to select from the algorithms discussed above. This proved highly successful in subjective quality, though was also the most computationally expensive.

Of the VAD algorithms discussed here, ALED, SFD and CVAD demonstrated the best performance, with SFD and CVAD demonstrating the highest subjective quality ($> 75\%$). Of the least computationally expensive algorithms, ALED demonstrated more consistent accuracy in objective tasks with fewer misdetections overall and misdetections below 10% in discontinuous monologue and rapidly spoken monologue tasks [100]. While these results are reasonable, all of these algorithms have designed for fairly straightforward speech detection applications, such as VoIP. Detecting speech in complex mixed audio, such as in entertainment media, is not as simple a task, and requires more sophisticated VAD approaches. As such, the following section explores a number of more recent developments in VAD specifically designed for speech detection in complex mixed audio signals.

Recent Approaches for Voice Activity Detection

This section surveys a number of leading approaches for VAD in challenging mixed audio data, including methods used applied for VAD in entertainment media such as radio broadcasts, television and film.

The approach proposed by Sonnleitner *et al.* [119], for classifying speech/non-speech in radio broadcasts, exploits spectro-temporal variations of speech signals via Short Time Fourier Transforms (STFTs) to discriminate between speech and non-speech signals. The

approach is based on the principle that music contains clear sustained harmonic patterns whereas speech contains more variable patterns with less consistent harmonic trajectories (as demonstrated in Figure 2.1). The approach analyses STFTs across adjacent frames and computes the inter-frame lag (frequency shift) using cross-correlation. This produces a simple feature vector which represents the inter-frame frequency shifts for a given audio signal. The feature is used to train a random forest classifier to classify content as speech or non-speech. This approach was tested on a dataset comprising music and speech content from radio broadcasts, and demonstrated very strong performance, achieving accuracies of $> 97\%$. This is encouraging as the approach was developed specifically for speech detection in complex multimedia content, making it an attractive approach to explore for speech classification in film audio. One criticism of this approach is its use of a median filter of approximately 10 seconds duration. Given that this work looks to develop an accurate method of aligning speech content, this filter would not be feasible as it would place too great a restriction on processing resolution. As such, this work will investigate the use of the approach discussed in Sonnleitner's paper [119] for speech detection in film audio, but will implement a version without the median filter in order to facilitate greater speech detection resolution.

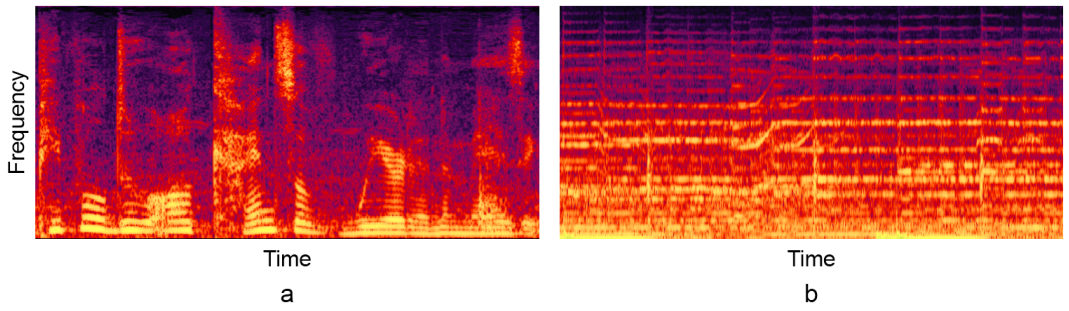


Figure 2.1: Spectrograms of speech (a) and music (b) content illustrating the difference in harmonic patterns.

Another recent approach presented in the paper by Eyben *et al.* [36] uses a voice activity detector based on Long Short-Term Memory Recurrent Neural Networks (LSTM-RNN). The approach was designed around LSTM-RNN's ability to model dependencies over time by incorporating information over a number of time steps. This is achieved via the LSTM-RNN's memory cells - memory which can be written to, read from or deleted according to feature context and previous outputs. This is done via multiplicative input, output and forget units whose weights are computed during training. In this way, the cells learn when to access the relevant parts of past context. The features used in this approach

are RASTA-PLP features, introduced in Hermansky *et al.*'s work [53]. These features are based on the short-term spectrum of speech, and use the Relative Spectral Transform (RASTA) method to make the PLP features more robust to linear spectral distortions. The approach uses cepstral coefficients 1-18, as well as their delta coefficients, resulting in a 36 dimensional feature vector. The VAD demonstrates good performance on a synthetic test validation set, with an average equal error rate (EER) of 10.4%, outperforming the state-of-the-art algorithm presented by Sohn *et al.* [118]. However, it is less effective on film audio, with an average EER of 33.2%.

A recent film-centered approach proposed by Tsiartas *et al.* [125] utilizes bilingual audio streams for speech detection. This identifies speech segments through correlating spectral coefficients between two different language tracks using a Long Term Spectral Distance (LTSD) feature. This feature is obtained by comparing the MFCC features across two audio streams of different languages. The LTSD values are low in non-speech regions and high in speech regions, as these are the regions in which the audio tracks differ substantially. The approach proved to be fairly successful, demonstrating an accuracy of between 84% and 87% in classifying clean and noisy speech in film audio. While this approach demonstrates good performance on film data, it requires bilingual audio tracks to perform classification. As such, this is not suitable for the target application as it would not work with a single audio track.

Another recent approach uses a dataset comprised entirely of television material (thus similar to film) and looks to differentiate between speech and music data [103]. This method first computes a Discrete Wavelet Transform (DWT) from each frame of the audio data. The paper compares the DWT method using a range of three wavelets - Haar, Symlets2 and Daubechies8 [103]. The resulting DWT features are then used to train a supervised classifier - the paper compared both SVM and GMM (Gaussian Mixture Model)-based classification algorithms. The DWT method proved to be fairly successful, achieving an accuracy of 95.4% when classifying speech and music content. This was achieved using the GMM-based classifier with the Daubechies8 wavelet. These results were obtained on similar data to that used by Sonnleitner *et al.* [119], in that the dataset comprised only music and speech. This makes the DWT approach attractive for the target application, however perhaps not as attractive as approaches which have achieved better results [119], or those which have demonstrated good performance on the target media [36].

While not applied to entertainment media, the method proposed by Kinnunen *et al.* [69] demonstrates that MFCC features can be used for effective discrimination of speech and non-speech content. Here, MFCC features are extracted from audio information, and

the MFCC, MFCCΔ and MFCC$\Delta\Delta$ features are used to train an SVM. The approach was tested using a subsection of the NIST2005 dataset [44] and two custom datasets, totalling 171 minutes of training data and 325 minutes of test data. Both linear and Radial Basis Function (RBF) kernels are explored for SVM classification, with the RBF kernel achieving the best results with an EER of 8%.

Another method, proposed by Chin *et al.* [20], combines MFCC features with a RBF Neural Network (RBF-NN) and continuous wavelet transform (CWT) for speech/non-speech discrimination. This was tested on the CUAVE database [98] with simulated noise content to evaluate its performance over a range of signal to noise ratios (SNR). The approach achieves strong results, outperforming a number of competing methods for VAD tasks [20]. One drawback of this method's evaluation is its exclusive use of the CUAVE database - which consists of 36 individuals reciting digits 0 to 9. Thus, a more comprehensive evaluation using a larger more varied dataset would have been helpful to support the method's efficacy.

Several other approaches in the literature have demonstrated an accuracy of $> 90\%$, however, these either have limited data, such as in the work by Pinquier *et al.* [99], which has only 9 main speakers in its dataset, or make use of non-film audio, such as Lu *et al.*'s work [80], whose data includes radio and news broadcasts (which typically do not have the same sonic variance as film data).

Summary

This section has explored a number of recent approaches for speech detection in complex mixed audio signals, covering a range of audio features and classification strategies. A key recurring concept present in all of the approaches investigated is that temporal information is crucial for audio discrimination tasks. This is unsurprising, as the temporal components of speech have already proven to be central to many existing ASR techniques (partly as this is crucial to the development of language models). Another key concept appearing in several approaches is the use of perceptually motivated features [36][125]. This is also logical, as human speech and hearing each played a crucial role in the other's evolutionary development. Thus, features designed according to our auditory perception are likely to be advantageous for speech processing.

Only two of the approaches found in the literature were applied to film audio - Eyben *et al.* [36] and Tsiartas *et al.*'s work [125] - with the rest being applied to other types of entertainment media content such as radio broadcasts [119] and television programs [103]. Of the two film-centric approaches, only Eyben *et al*'s could be applied to a single audio track, as the other required bilingual audio data. Given this, Eyben *et al.*'s approach would

serve a good baseline for VAD development.

Of all entertainment media-based approaches investigated, Sonnleitner *et al.*'s approach [119] achieved the strongest performance, with accuracies exceeding 97%. As such, the approach will be evaluated using feature-film audio data in order to determine whether it is capable of achieving similar levels of performance in the target domain.

The literature demonstrates that one of the key factors to be considered in the development of VAD approaches is the application context. This is as different types of audio content will contain different variability in speech and non-speech data. For example, in a straightforward VoIP setting, the variability in both speech and non speech content is fairly low. On the contrary, in the tasks considered by Sonnleitner and Eyben's work, the variability can be fairly high: with there being a great degree of variability in the negative classes. In the case of Eyben's work, this comprises a range of atmospheric sounds, e.g. traffic and crowd noise, which complicates the task of non-speech rejection. In the application context of Sonnleitner's work, this comprises varying types of music - much of which is vocal music, which blurs the line between speech and non-speech content. In the context of feature-film audio, an additional challenge is presented: as well as a significant variety of non-speech content in the form of sound effects and music, the speech content also contains much greater variability. This is due to highly emotive speech with vastly varying dynamics, including shouted speech, whispered speech and varying rates of dialogue. Thus, when developing VAD for feature film content, a significant variation in background noise must be considered, along with significant variability in the speech content itself.

A number of methods for speech feature extraction have been discussed in this section. Of these, MFCCs are of particular interest. While they have demonstrated encouraging performance for some VAD applications [69], as well as robust performance under challenging SNR conditions [20], the literature does not contain examples of MFCC VAD applied to entertainment media. This is surprising given that MFCCs have a number of properties which make them ideal for this application context. Firstly, their use of perceptually-scaled features is advantageous given the strong link between the human auditory system and speech production [11]. Secondly, cepstra are sensitive to periodicity in the frequency domain - this makes them particularly attractive given that previous work has highlighted the importance of spectro-temporal patterns in VAD applications [119]. As such, this work uses MFCCs as the foundation of the audio VAD approach proposed in section 3.

While audio VAD will form the basis of the approach in this work, it would benefit from other modalities in order to improve its robustness to noise. This could be achieved with the use of visual VAD, whereby the video information could be utilised to enhance speech detection in the presence of audio noise. This has proven to be useful in previous

work [2], and as such this work will explore the use of visual speech features to enhance speech detection in entertainment media.

2.2 Computer Vision Approaches for Speech Processing

A number of techniques have been developed for extracting and processing visual speech information. These include automatic lip reading, visual voice activity detection and methods for automatic speaker identification. These technologies typically rely on underlying computer vision processes for face detection and feature extraction. This section reviews a number of computer vision technologies used within facial feature extraction before exploring existing developments in visual speech processing.

2.2.1 Face Detection

Face detection is crucial in most visual speech processing tasks, as it is necessary for initialising other information extraction processes, such as landmark localisation. A number of approaches have been developed for face detection, from Viola and Jones' work on Haar-like features [132], to more recent approaches such as the Histograms of Oriented Gradients (HOG) approach presented by Dalal *et al.* [29], developed to provide more robust detection performance in variable conditions.

Haar-Like Features

Viola and Jones' Haar-like features approach [132] is a popular method for face detection within still images and video content. The approach is based on Haar wavelets [46], and achieves computational efficiency through the use of summed area tables and adaptive boosting (AdaBoost). This allows the sum of rectangular areas in the image to be computed using a finite number of lookups.

The features are defined as the intensity difference between two to four rectangles, as demonstrated in Figure 2.2. For example, in feature a the feature value is the difference in the average pixel value in the grey and white rectangles. The Haar detector exploits three different types of rectangular features - edge features (a and b), line features (c and d) and four-rectangle features (e and f) for object detection.

AdaBoost is then used to train a cascade of weak detectors according to a decision threshold. For each Haar-like feature in the pool, AdaBoost finds the optimal threshold and confidence scores. This information is then used to select the best feature with the

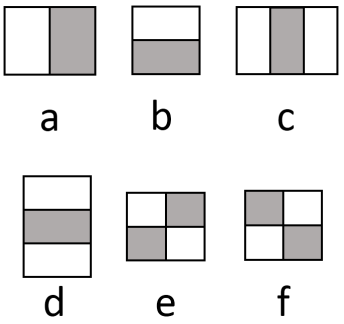


Figure 2.2: Illustration of Harr-like rectangular features.

minimum Z score. This is repeated until all of the weak classifiers are trained, resulting in a Haar cascade which can be used for image classification/object detection.

In the case of face detection, the process begins by evaluating whether regions in an image can be represented by Haar-like features associated with general facial features. If this is true for a given region, the process moves to the next phase in the cascade, evaluating against Haar-like features which correspond to more specific facial characteristics. If this is satisfied for all stages in the cascade, the region is classified as containing a face.

One of the limitations of the earlier Haar-like features implementation is the poor performance on non-frontal faces. To improve on this, Lienhart and Maydt [76] extended the work to allow rectangular features to be combined in a greater variety of ways to improve object detection. This was achieved through the introduction of 45 degree rotated rectangular features by using rotated integral images.

Rowley *et al.*'s work introduced a three step approach for non-upright face detection [106]. This uses two neural network classifiers. The first step estimates the pose of the face in the detection window prior to adjusting the image and applying a standard face detector. The three key steps for detection are therefore: 1) estimate the pose of the face, 2) use the pose estimation to de-rotate the image window, 3) apply the face detector to the de-rotated window. Investigations into the performance of this approach demonstrated a detection rate of up to 96.0% on rotated face data. Two drawbacks of this approach are: 1) as the classifiers primarily work independently, the resulting detection rate is generally the product of the correct classification rates of the two classifiers; and 2) image de-rotation is computationally expensive.

To improve on this, Viola and Jones proposed a technique for multi-view face detection based on their earlier work [62]. This uses diagonal filters to focus on diagonal structures within the image window. These work in the same way as they previous filters, and can be

computed via 16 lookups; looking at the 16 corner pixels.

This approach yielded similar accuracy to that proposed by Rowley et al., achieving a detection rate of up to 95.0% [62]. While this produced similar accuracy results, the Viola-Jones approach proved to be advantageous as it did not require image de-rotation, thus being more computationally efficient and providing a faster method for face detection.

Local Binary Patterns

Another popular method for face detection is based on Local Binary Patterns (LBP) [1]. LBP works by creating feature vectors comprising binary information from pixels within a given window. The first step in computing the LBP feature is to divide the window into cells of $n * n$ pixels. Next, each pixel, p_c, in the cell is compared to its neighbouring pixels, p_n. If $p_n >= p_c$, a value of 0 is recorded. If $p_n < p_c$, a value of 1 is recorded. As illustrated in Figure 2.3, this results in an 8 digit binary number which represents p_c in relation to its neighbouring pixels. Once this has been done for all pixels in the cell, a histogram is computed over all cell values, giving the frequency of each value in the cell. This histogram forms the basis of the LBP feature vector. The final feature vector is obtained by concatenating the feature vectors for all cells within the window.

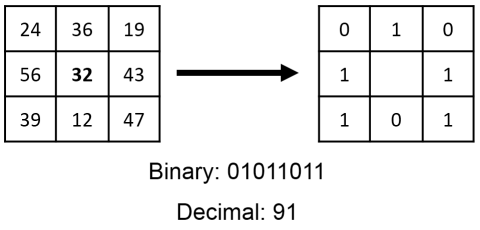


Figure 2.3: Illustration of LBP feature computation.

The resulting feature is then used to train a machine learning algorithm, such as support vector machine or random forest. LBP features have demonstrated strong performance in a range of computer vision tasks, including face detection [1], texture analysis [95] and other object recognition tasks.

LBP has been successfully implemented in a number of face detection frameworks, and has exhibited strong performance. In [47], LBP features are used with a SVM for face detection, achieving a 97.8% detection rate on the MIT-CMU dataset [105]. Similarly, in [138], a modified LBP approach is proposed which utilises RGB and YUV colour space data. LBP histograms are extracted from these colour spaces, after which histogram

matching and SVM methods are used for face classification. The approach achieved a 93.8% detection rate on a bespoke colour database comprising 356 frontal faces. A modified version of LBP, termed improved LBP (ILBP), was proposed by Jin *et al.* [60]. As with standard LBP, this creates a histogram based on pixel value relationships within a cell. In this case, this is achieved by comparing all pixels in a cell with the mean intensity of all pixels in the cell. In later work by Jin *et al.* [61] these are used to train a cascade AdaBoost detector, which achieves a 93% detection rate on the MIT-CMU database.

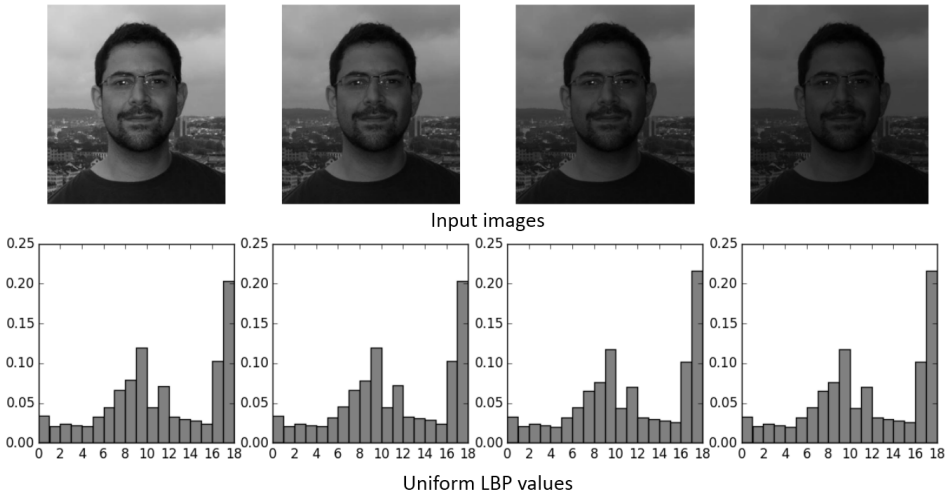


Figure 2.4: Illustration of LBP invariance to illumination conditions.

While the detection rate of LBP is comparable to Haar-like features, it has the principle advantage of being more robust to variable illumination conditions, as illustrated in Figure 2.4. The uniform LBP values in Figure 2.4 are obtained by aggregating the 8 digit cell representations into k bins, where k is the number of unique 8 digit representations obtained from the image (in this case $k = 18$). The histogram data here demonstrates LBP's invariance over a range of lighting conditions. This illumination invariance has been exploited in work by Roy *et al.* [107], which combines both Haar wavelet and LBP-based features to create Haar Local Binary Pattern (HLBP) features. Their work demonstrates that this achieves improved performance on data containing variable lighting conditions, with the HLBP approach outperforming the standard Haar-like features approach and a competing LBP-based approach on 3 out of 4 datasets.

Histogram of Oriented Gradients

Another method that has been successfully applied for face detection [31] [48] and other human detection tasks [29] is the Histogram of Oriented Gradients (HOG) feature descriptor. This was proposed by Dalal *et al.* in [29], and was initially developed for pedestrian detection in static images. The approach is based on the principle of evaluating the frequency of gradient orientation in localised image segments to create histograms of gradient orientations for a number of connected segments. In a similar fashion to LBP, these histograms are concatenated to form the final HoG feature vector.

Figure 2.5: Example of HOG features. Left: input image. Right: HOG features.

To obtain the HOG feature, the image window is first divided into a number of cells. The gradient values are then computed by applying a 1-D centred point discrete derivative mask to the image data. For colour images, the gradients are computed for each channel, and the channel with the largest norm is used as the gradient vector for the pixel. This is followed by orientation binning, in which the gradient magnitude of each pixel within the cell is used to cast a weighted vote on the cell orientation. This information is used to quantise the cell orientation into one of 9 possible bin values. To facilitate better invariance to illumination, cells are grouped into larger overlapping blocks and contrast normalisation is applied separately to each block. The final HOG feature is the vector of all normalized cell responses from the blocks in the detection window. This can be used to train machine learning algorithms for object detection tasks. Figure 2.5 illustrates the transformation from original image to HOG features.

In their paper, Dalal *et al.* demonstrate that the approach can be used to train an SVM for human detection tasks. Crucially, they show that their method significantly reduces

false positive rates when compared to leading Haar wavelet-based approaches for human detection. A key factor in the approach's performance is its invariance to geometric and photometric transformations [29], making it particularly attractive for object detection tasks in challenging lighting conditions.

Summary

Several methods for face detection in image and video content have been explored. Of the methods explored, HOG features are perhaps the most attractive given their robust performance in variable illumination conditions. This is advantageous given that the work looks to develop technologies for use in entertainment media - which contains video which is highly dynamic both in terms of variable object orientation and variable lighting conditions. While LBP-based detectors have exhibited some improvement on Haar wavelet-based detectors, the Haar-like features approach is particularly attractive given its consistent track record for face detection tasks and the fact that readily available implementations are incorporated into prevalent computer vision libraries [96]. As such, both Haar wavelet and HOG-based approaches will be explored for face detection within this work.

Once the face region has been detected, landmark features need to be identified in order to locate the mouth region and extract visual speech information. The next section therefore goes on to investigate a range of landmark localisation approaches which can be used for the localisation of the mouth region and extraction of visual information for use in visual speech detection.

2.2.2 Landmark Localisation

Landmark localisation concerns the detection of facial landmarks, and follows the face detection step in the visual speech information extraction process. A broad variety of landmark localisation techniques have been developed over recent years, including shape and appearance-based approaches - such as Active Shape Models (ASM) [24] and AAMs [23] proposed by Cootes *et al.* - as well as pictorial structure-based approaches, as proposed by Kazemi *et al.* [66].

Active Appearance Models

Earlier approaches for landmark localisation include ASMs, which were first proposed by Cootes *et al.* [24]. ASMs are based on the concept of modelling an object as a deformable statistical model with a mean shape. This concept comes from the idea of Active Contour Models (ACM) [63], but unlike ACMs, ASMs limit the possible deformable

shapes according to information from the training set. Since their development, ASMs have proven to be a popular method for landmark localisation [89][104].

Following the development of the ASM, Cootes *et al.* went on to develop the Active Appearance Model [23]. This extends their earlier work on ASMs through incorporating grey-level appearance information. The approach generates models by combining a model of shape variations with a model of appearance variations in a shape-normalised frame. This is done using a training set of images for which key landmarks are marked. All sets are aligned to a common co-ordinate frame and represented by a vector $\mathbf{x}$. Principal Component Analysis (PCA) is then applied to the data, allowing any feature to be approximated simply by:

$$\mathbf{x} = \bar{\mathbf{x}} + \mathbf{P}^s\mathbf{b}^s \tag{2.5}$$

where $\bar{\mathbf{x}}$ is the mean shape, $\mathbf{P}^s$ is a set of orthogonal modes of variation and $\mathbf{b}^s$ is a set of shape parameters.

A statistical model of the grey-level appearance is obtained by warping each example image so that its control points match the mean shape (via a triangulation algorithm). The grey level information is then sampled from the shape normalised image over the region covered by the mean shape. The impact illumination variance is minimised by normalising the example samples. Once normalised, the linear model of the grey level appearance can be obtained via PCA, similarly to the shape model:

$$\mathbf{g} = \bar{\mathbf{g}} + \mathbf{P}^g + \mathbf{b}^g \tag{2.6}$$

where $\bar{\mathbf{g}}$ is the mean normalised grey-level vector, $\mathbf{P}^g$ is a set of orthogonal modes of variation and $\mathbf{b}^g$ is a set of grey-level parameters. In this way, the shape and appearance of any example can be described by vectors $\mathbf{b}^s$ and $\mathbf{b}^g$, corresponding to the shape and grey-level content respectively. The shape parameter is then scaled to ensure that the units of $\mathbf{b}^s$ are equivalent to $\mathbf{b}^g$, before the vectors are concatenated and PCA is applied to the concatenated feature vector $\mathbf{b}$:

$$\mathbf{b} = \mathbf{Qc} \tag{2.7}$$

where $\mathbf{Q}$ are the eigenvectors and $\mathbf{c}$ is a vector of appearance parameters controlling the shape and grey-level model parameters.

The approach has been used for face landmark localisation using a training set of 400 images, with each image containing 122 labels corresponding to the main features. This was used to generate an active appearance model capable of explaining 98% of the observed

variation using 80 parameters.

Given this AAM and an image to be interpreted, the model parameters need to be found which optimise the match between the target image and the synthesised example. This is achieved through the use of an iterative algorithm described by Cootes *et al.* [23] which minimises the magnitude of the difference vector, $\Delta = |\delta\boldsymbol{\phi}|^2$, for

$$\delta\boldsymbol{\phi} = \boldsymbol{\phi}_i - \boldsymbol{\phi}_m \tag{2.8}$$

where $\boldsymbol{\phi}_i$ is the vector of grey level values from the image and $\boldsymbol{\phi}_m$ is the vector of grey-level values for the current model parameters.

Since their development, AAMs have go on to demonstrate strong performance in facial landmark localisation tasks [23], [2], [75], [42].

Recent Developments in Landmark Localisation

More recent developments in landmark localisation have seen significant improvements in performance, particularly with regard to landmark localisation and tracking in challenging data. Specifically, approaches such as those proposed by Kazemi *et al.* [66], Zhu *et al.* [139] and Saragih *et al.* [113] have demonstrated strong performance in data incorporating variable illumination and occlusions. This is particularly valuable given that this work is interested in feature film data. As such, the data will contain natural speaker behaviours, which will include gestures that are likely to cause partial or whole obstructions. The data is also likely to contain a significant degree of movement through scenes with variable lighting, thus making robust performance in these conditions particularly critical.

Following the AAM work, Cristinacce and Cootes developed their Constrained Local Model (CLM) approach for landmark localisation [28]. This approach has demonstrated improved performance on challenging face data incorporating variable illumination conditions and occlusions.

The approach uses the same method as the AAM to build a combined shape and texture model, however this work modifies the texture sampling method. Here, for each feature, a training sample is obtained and normalised, producing pixel values with zero mean and unit variance. These texture patches are then concatenated, resulting in a single grey value vector for the training image. The resulting vectors and normalised shape co-ordinates are then used to construct linear models according to the method in Cootes *et al.*'s work [23], shown in equations 2.4 and 2.5. As with the AAM work, these are combined using PCA to form a joint model.

The next step in the CLM approach is to perform a shape constrained local model search.

Given this joint shape and texture model, a set of grey value regions can be generated for a set of features. This is achieved by applying the templates to a search image and computing the response images, where (x_i, y_i) is the position of feature point i and $\phi_i(x_i, y_i)$ is the response of the i^{th} feature template. These positions can be concatenated into a vector as:

$$\mathbf{x} = (x_1, ...x_n, y_1, ...y_n)^T \tag{2.9}$$

Here, **x** is computed as:

$$\mathbf{x} \approx T_{\mathbf{t}}(\bar{\mathbf{x}} + \mathbf{P}^s\mathbf{b}^s) \tag{2.10}$$

where $\mathbf{b}^s$ are the shape parameters and $T_{\mathbf{t}}$ is a similarity transform from the shape model frame to the response image frame. These can be concatenated into $\mathbf{p} = (\mathbf{t}^T|(\mathbf{b}^s)^T)^T$, allowing **x** to be represented as a function of **p**. Given an initial value for **p**, the search process optimises the function $f(\mathbf{p})$ according to the image response surfaces $\boldsymbol{\phi}_i$ and the statistical shape model from the training set. The function is given as:

$$f(\mathbf{p}) = \sum_{i=1}^{n} \boldsymbol{\phi}_i(x_i, y_i) + K \sum_{j=1}^{s} \frac{-b_j^2}{\lambda_j} \tag{2.11}$$

where the second term is an estimate of the log-likelihood of the shape given parameters b_j and eigenvalues λ_j, and K is a weight determining the relative importance of good shape and strong feature responses. The function $f(\mathbf{p})$ is optimised using the Nelder-Meade simplex algorithm [92].

The CLM approach uses a straightforward algorithm centred around three core processing steps, as illustrated in Figure 2.6: i) initial feature point input, following by ii) fitting the joint shape and texture model to the feature points, and iii) using the shape constrained search method to predict a new set of features. Steps ii and iii are repeated until the model converges. When applied to video content, the initialisation points are obtained from the previous frame if possible, and are otherwise generated via global search.

Performance evaluation using the BIOID [59] and XM2VTS [88] datasets demonstrated the CLM method outperformed a number of leading approaches for landmark localisation and tracking [28], including AAM, TST [27] and SOS [26] approaches.

Further approaches have since been developed based on the initial CLM method. In Saragih *et al.*'s work [113], they propose a method termed Regularised Landmark Mean-Shift (RLMS). This method has two key advantages over the original CLM approach. The first is an improved procedure for model fitting, which is achieved via a new method for approximating feature likelihood maps using nonparametric representations.

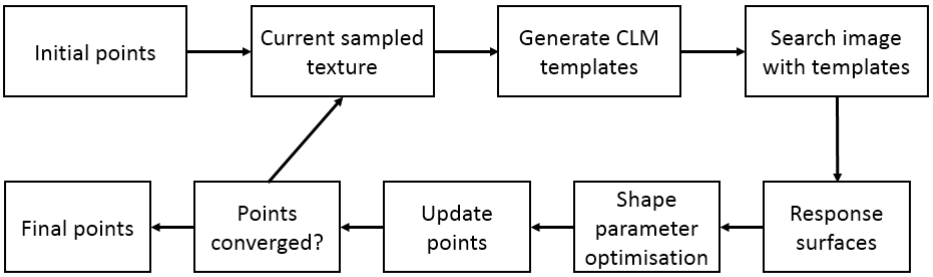


Figure 2.6: Flow diagram of CLM search algorithm.

The second key advantage of this approach is its method for handling partial occlusions, as demonstrated in Figure 2.7. The problem addressed here is that, in a parts-based solution such as CLM or RLMS, the likelihood of individual features is determined according to the training set. Thus, if a face is partially occluded, the model will not consider the occluded feature to be a likely candidate unless a similar example is present in the training set. In Saragih *et al.*'s work, performance on data containing partial occlusions is enhanced by modifying the candidate likelihood function to account for potential outliers. In doing so, the model is able to handle partial occlusions, assuming that the majority of candidates fit the model [113].

This method was evaluated using the MultiPIE [43] and XM2VTS [88] datasets for performance on still images and the FGNet [101] dataset for performance on sequences. In both cases, results demonstrated enhanced performance when compared to a number of existing approaches, with improved fitting and tracking metrics when compared to ASM [24], convex quadratic fitting [133] and GMM-based approaches [45].

Figure 2.7: Example of approach from [113]'s performance on partially occluded data.

Another recent landmark localisation approach was proposed by Kazemi *et al.* [66]. This approach uses a cascade of regression trees to estimate and refine the position of facial landmarks. Here, the x, y coordinate of the ith landmark in a given image, $\mathbf{I}$, is given as $x_i \in \mathbb{R}^2$. Given this, the vector $\mathbf{s} = (x_1^T, x_2^T, ...x_p^T)^T \in \mathbb{R}^{2p}$ denotes all p facial landmark

coordinates in $\mathbf{I}$, corresponding to the face shape. $\hat{\mathbf{s}}^{(t)}$ therefore denotes the current shape estimate. Each regressor in the cascade is represented by $r_t()$, and predicts an update vector given the image and an estimate $\hat{\mathbf{s}}^{(t)}$. This prediction is then added to the current shape estimate to improve the estimate:

$$\hat{\mathbf{s}}^{(t+1)} = \hat{\mathbf{s}}^{(t)} + r_t(\mathbf{I}, \hat{\mathbf{s}}^{(t)}) \tag{2.12}$$

A crucial factor in using the cascade-based approach is that the regressor's predictions are made based on features computed from the image, $\mathbf{I}$, and indexed according to the current shape estimate $\hat{\mathbf{s}}^{(t)}$. As such, the confidence of each feature increases with each iteration through the cascade, and the approach is more robust to geometric variance as it is able to correct itself with respect to current estimates.

The regressors are trained using the gradient tree boosting algorithm described by Hastie *et al.* [50] using a sum of square error loss function. During training, each regressor is trained iteratively, using the parameters generated by the previous regressor. In this way, a cascade of T regressors is generated as:

$$r_0, r_1, r_2, ...r_{T-1} \tag{2.13}$$

where each r_t comprises 500 weak regressors. The implementation in Kazemi *et al.*'s paper [66] uses $T = 10$ of these ensemble regressors for the cascade.

The approach was trained and tested using the HELEN [74] dataset, from which 2000 images were used for training and 330 were used for testing. Performance was evaluated with respect to two leading ASM-based approaches, STASM [89] and CompASM [74]. Results demonstrate that Kazemi *et al.*'s approach achieves significantly improved performance on the HELEN dataset, with an average error of 0.049, compared with 0.111 and 0.091 obtained by the STASM and CompASM approaches respectively. Further to this, the paper evaluates the impact of using a cascade approach, with the error improving from 0.085 to 0.049 when compared with a single level ensemble.

As with Saragih's approach [113], this method also produces robust performance on partially occluded data, as demonstrated in Figure 2.8. This is due to the fact that, as the cascade progresses, the model is evaluated and updated with regard to the overall shape estimate, $\hat{S}^{(t)}$. As such, occluded regions will have limited impact as long as a majority shape components agree with the model, thus resulting in a robust overall estimate despite occluded data at specific landmarks.

Figure 2.8: Example of approach from [66]'s performance on partially occluded data.

Summary

This section has discussed a number of key approaches for facial landmark localisation, covering earlier developments such as AAMs [23] as well as more recent developments based on CLM [28] and cascade ensemble [66] approaches. As the application focus of this work is entertainment media, a high degree of variance in pose, lighting and speaker gestures can be assumed. Thus, an approach which demonstrates strong model fitting on datasets comprising faces in a natural setting, and robust performance on occluded data is desirable. Given this, the approaches proposed by Kazemi *et al.* [66] and Saragih *et al.* [113] are the most suitable methods investigated, as both demonstrate significant performance advantages over state of the art on 'faces in the wild' data, and both perform well on occluded data.

2.2.3 Visual Speech Processing

Processing visual speech information has been an active area of research for a number of years, producing numerous developments in areas including visual voice activity detection (V-VAD) [75] and automatic lip reading [49]. These approaches make use of a number of methods for extracting visual speech features, and these typically fall into two categories: appearance-based features and landmark-based features. Appearance-based features typically use grey-scale information from the image, such as two dimensional cosine transform (2D-DCT) and LBP. Landmark-based features, on the other hand, use the information obtained from shape estimation approaches, such as AAM and CLM, rather than directly using the grey-scale information. This section explores a number of existing approaches for processing visual speech information in order to determine which features and feature combinations are likely to be useful for modelling visual speech information in entertainment media.

Visual Automatic Speech Recognition and Automatic Lip Reading

The approach for automatic lip reading described by Hassanat *et al.* [49] combines geometric and appearance-based features for automatic visual speech recognition. This method first employs face detection and lip region detection before extracting mouth-based features. While the method does not use facial landmarks, it does make use of mouth height and width parameters to give an impression of mouth shape. These are combined with appearance-based features to give an impression of the presence of the teeth or tongue. The method was evaluated using a bespoke dataset comprising 26 speakers, and achieved a word recognition rate of 76.4% for speaker dependent tests, and 52.8% for speaker independent tests. While this falls short of more recent approaches such as Thangthai *et al.*'s work [123], this work highlights the importance of appearance-based features for modelling visual speech. This is as the presence or absence of the tongue and/or teeth provides crucial information as to the speech components most likely to correspond with the visual speech gestures [49].

In a paper by Lan *et al.* [73], AAM-based features are used to extract visual speech information. These features are obtained using a Linear Predictor (LP) based tracker, which was shown to improve landmark estimates across frames when compared with an AAM-based tracking approach on the chosen dataset. The work also investigates a number of other features, including 2D-DCT and eigen-lip features. These are evaluated to determine their performance on visual speech recognition tasks on the Grid dataset [22]. The results in Lan *et al.*'s paper demonstrate that a feature comprising both AAM shape and appearance features achieved the strongest performance, with a word accuracy of 59%. Crucially, this work demonstrates that the combined approach outperforms the individual shape and appearance features. This is further supported by Hassanat *et al.*'s work [49], which goes on to use combined AAM shape and appearance features for viseme classification.

In Thangthai *et al.*'s work [123], Deep Neural Networks (DNNs) are investigated for use in visual speech recognition. Here, combined AAM shape and appearance components are again used, with the method first extracting these features from the lip region in a given video frame. Hierarchical Linear Discriminant Analysis (HiLDA) features are then obtained by applying Linear Discriminant Analysis (LDA) to a set of high dimensional features constructed from the first, middle and last frame in a 15 frame window. A Maximum Likelihood Linear Transform (MLLT) is then applied to the features in order to minimize intra-class distance while maximising inter-class distance. These features are used as the input for a context dependent DNN-HMM model (CD-DNN), which is trained using stochastic gradient descent for four hidden layers with a hyperbolic tangent activation

function. The paper evaluates the performance of the CD-DNN using both simple AAM features and HiLDA with respect to GMM and conventional HMM-based approaches. Results demonstrate that the HiLDA method achieves the strongest results, with 84.7% word accuracy on the speaker dependent RM-3000 dataset. This is with respect to the simple AAM-based approach, which achieved a word accuracy of 77.49%, and the HTK baseline approach described in the paper [123], which achieved a word accuracy of only 47.5%.

Visual Voice Activity Detection

While the automatic lip reading approaches give some impression of which techniques may be useful, one of the aims of this work is to develop an alignment strategy that does not rely on a language model. Given that automatic lip reading is a speech recognition task, all successful automatic lip reading approaches require a language model. Here, a broader technique for visual speech processing is explored - visual voice activity detection (V-VAD). The idea here is that it may be possible to leverage audio-visual speech activity patterns for sequence association and alignment.

A number of V-VAD approaches have been proposed over recent years. These include simple GMM-based approaches [2], as well as more comprehensive Neural Network (NN) oriented approaches [75] and unsupervised methods [121]. These have been developed to enhance existing VAD systems through incorporating visual information, as this has proven to improve performance in the presence of acoustic noise [2] and on data containing variable audio speech dynamics [121].

Several recent developments in V-VAD use DCT and GMM-based approaches, such as work by Almajai *et al.* [2] and Navarathna *et al.* [91]. In Almajai *et al.*'s work, the mouth is detected using an AAM, and 2D-DCT information is extracted from a ROI centered around the detected mouth area. Expectation Maximization (EM) clustering is then used to create two GMMs - one for speech, and one for non-speech. The visual speech is then classified using the GMM to determine whether a given sample contains speech activity. This method was evaluated on a subset of the Grid corpus, and demonstrated reasonable results with V-VAD accuracy of 72%. Crucially, this work demonstrated the advantage of combining audio and visual speech features, with the combined audio-visual VAD achieving $\approx 90\%$ accuracy in noisy data, where the audio-only VAD achieved $\approx 50\%$ accuracy.

In Navarathna *et al.*'s work [91], DCT information from the mouth ROI is concatenated over 7 frames prior to dimensionality reduction via Linear Discriminant Analysis (LDA). GMM's are then used to model the data and classify an input vector as either speech or non-speech, similarly to Almajai *et al.*'s approach [2]. The method is evaluated using the

CUAVE database [98], and produces reasonable results, with a false positive (FP) rate of 24.2% and missed detection (MD) rate of 27.6% on frontal face data. This yields a half total error rate (HTER = $\frac{FP+MD}{2}$) of 25.9%. This error rate increases significantly for profile views, for which it achieves a HTER of $\approx 35\%$. These results support the intuition that frontal face information is far more informative when modelling visual speech activity. That said, this is likely to be more critical for appearance-based methods such as the DCT used here, as they have no mechanism of estimating features which correspond to occluded areas of the face, whereas this may be achievable with landmark-based methods.

More recently, appearance-based approaches for V-VAD have been used in more sophisticated classification frameworks. In Tao *et al.*'s paper [121], optical flow is combined with mouth height and width information to create a base feature vector. The temporal characteristics are then modelled using short-time zero crossing rate and short-time variance information, after which the resulting features are combined prior to dimensionality reduction using principal component analysis (PCA). The EM algorithm is then used to automatically cluster the speech and non-speech components, modelling them as a GMM with two univariate mixtures. The speech or non-speech classification is then determined according to the value of the GMM cluster means, with a higher mean indicating that the cluster represents speech information, and a lower mean indicating that the cluster represents non-speech information. This method is evaluated using the MSP-AVW Corpus [124], and achieves an accuracy of 79.1% using the unsupervised model, and an accuracy of 86.9% using a supervised variation of the model. A key focus of this investigation is V-VAD's performance on variable speech dynamics. This work demonstrates that, while audio-VAD achieves better results for normal speech, V-VAD achieves better results on whispered speech, with consistent results obtained regardless of speech dynamics.

Recent work by Le Cornu *et al.* [75] uses two different neural network-based methods. The first of these is a backpropagation neural network trained on DCT coefficients extracted from the mouth region. Their paper details two versions of the backpropagation network: first using only the DCT coefficients, and secondly using the DCT coefficients and their first order temporal derivatives. These are described as "NN DCT" and "NN DCT Δ" respectively.

Their second neural network approach explores the use of CNN's for speech classification - using raw pixel intensities as the input information and using the CNN for feature discovery and classification. As with their first neural network approach, there are two variants of this method - both without and with temporal information, described in the paper as "CNN Static" and "CNN Stack 3". For the CNN Static approach, the CNN architecture comprises two sets of standard CNN layers. These each contain a convolutional

stage, a max-pooling stage and a dropout stage. The first convolutional stage comprises 32 filters of size 3×3, and the second comprises 64 filters of the same size. Each of the convolutional stages are followed by non-overlapping max-pooling with max filters of size 2×2, after which dropout is applied to each max-pooled layer with probability $p = 0.2$. These layers are followed by a single fully-connected layer comprising 512 units with dropout of probability $p = 0.5$, and an output layer which uses a softmax activation function. The CNN Stack 3 approach uses the same architecture, but incorporates temporal information via the early-fusion technique. This method stacks three video frames at the input stage, enabling the first convolutional stage to convolve across consecutive frames to build a representation of local motion direction.

These approaches are investigated on a subset of the Grid corpus, and compared with the method from Almajai *et al.* [2]. The results demonstrate that the NN-DCT approach achieves the strongest results in speaker independent tests, with a VAD accuracy of 78.7%, improving on the CNN and GMM [2] approaches which achieve a VAD accuracy of 74.7% and 70.5% respectively. This work supports earlier findings by Almajai *et al.* [2] and Navarathna *et al.* [91], confirming that 2D-DCTs are strong appearance-based features for detecting speech activity, and goes on to demonstrate that VAD performance can be further enhanced with the use of more sophisticated machine learning algorithms such as Neural Networks (NNs).

While not as prevalent in the literature as appearance-based methods, a number of landmark-based V-VAD approaches have been explored. One such approach proposed by Liu *et al.* [77] performs lip landmark extraction using active contour models and rotational template matching. Once the lip landmarks are extracted, a vector, $\mathbf{v}(t)$, containing both static and dynamic features is constructed to model the movement of the lip region over time. AdaBoosting is then used for speech classification, with an ensemble of weak classifiers predicting on elements within the feature vector $\mathbf{v}(t)$. The content is classified as speech if the output from the ensemble satisfies a voting threshold, and non-speech otherwise. While this approach produces reasonable results, achieving an equal error rate (EER) of < 0.15, the work uses a very limited dataset, consisting of only 150 seconds of data, of which only 30 seconds was used for testing. Furthermore, the work provides no comparison to other approaches or baseline methods.

Another approach proposed by Aubrey *et al.* [6] utilises both shape and appearance-based features from an AAM, using a Hidden Markov Model (HMM) to model lip-based shape and appearance characteristics to classify visual speech activity. The approach uses eigendecomposition to obtain the first 10 eigenvectors, which contain 75% of the original appearance energy. These are then used to train a HMM using 600 frames of visual speech

data. This approach is evaluated by calculating the number of correct silence detections, rather than speech detections. While results are encouraging, achieving up to 90% for correct silence classification, the dataset used here also contains only 150 seconds of data.

Both Liu *et al.* [77] and Aubrey *et al.* [6] use considerably smaller datasets than the other V-VAD approaches discussed here, which use at least ≈60 minutes of data [75], with some using considerably more [121]. Hence, it was felt that neither of these landmark-based V-VAD investigations provide a sufficiently comprehensive evaluation from which the efficacy of landmark features for V-VAD can be determined. As such, this work will look to carry out a more comprehensive evaluation of landmarks for V-VAD tasks, specifically with regard to more challenging speaker conditions such as those likely to be encountered in entertainment media.

Summary

This section has explored a number of methods for visual speech processing, with a focus on visual speech recognition and visual voice activity detection. Existing work in visual automatic speech recognition (visual ASR) has demonstrated the combining both shape and appearance features can yield performance improvements when compared to individual features, as demonstrated in work by Thangthai *et al.* [123] and Lan *et al.* [73]. Further to this, work on visual ASR has highlighted the importance of appearance features, in that they indicate the presence/absence of the tongue and teeth - crucial features for differentiating between different types of speech phenomena.

While visual-ASR approaches clearly demonstrated enhanced performance when combining shape and appearance features, the V-VAD approaches found in the literature were largely appearance based. While these have demonstrated strong results using primarily appearance-based features, they have all been evaluated using fairly ideal datasets with static speakers and consistent illumination. This raises the question of whether these approaches would work as well given more variable data, and whether the inclusion of robust landmark localisation approaches, such as those discussed in Section 2.2.2.2, would enhance performance under these conditions.

As such, this work will look to further explore the use of combining shape and appearance features for visual speech processing, with a specific focus on utilising robust landmark localisation methods. This work will also explore the use of 2D-DCTs for providing appearance information, as these have demonstrated strong performance in several V-VAD approaches [2][75][91].

2.2.4 Computer Vision Approaches for Speech Processing - Summary

This section has evaluated a number of existing computer vision approaches for visual speech feature extraction as well as approaches for processing visual speech information, such as visual-ASR and V-VAD.

The review of the literature demonstrates that there are a number of promising methods for face detection and landmark localisation, both crucial processes for obtaining features associated with visual speech. Of these methods, those demonstrating robust performance on data containing variable illumination and partial occlusions are most attractive, as these are challenges present within entertainment media. This makes the more recent developments in landmark localisation, such as proposed by Saragih *et al.* [113] and Kazemi *et al.* [66], particularly attractive.

Existing work on visual speech processing demonstrates a strong bias towards appearance-based features, particularly for V-VAD. Despite this, work on visual-ASR has demonstrated that combining appearance and shape-based features yields some improvement [73]. However, investigations into combined features for V-VAD use limited datasets, and are thus not sufficiently conclusive as to whether this is also the case for detecting visual speech activity. This work will therefore look to conduct a more comprehensive study on the use of shape and appearance-based features for V-VAD applications, with a specific focus on applying state of the art methods in landmark localisation.

2.3 Feature Matching and Sequence Alignment

A key aim of this work is to develop a method for multimedia alignment using speech information. As such, feature matching and sequence alignment methods are crucial. This section explores a number of methods for matching features and aligning sequences using speech information.

2.3.1 Speech to Text Alignment

A number of approaches exist for audio to text alignment, including work by Katsmanis *et al.* [65], Goldman *et al.* [41] and Braunschweiler *et al.* [16]. The method proposed by Katsamanis *et al.* provides an adaptive, iterative speech recognition and text alignment solution capable of aligning long, noisy audio content, and also allows for transcription errors. The system first segments an audio stream into chunks of 10-15 s duration. Speech recognition is then applied to the individual segments to provide an estimate of the word sequences. This is aligned with a reference transcript, and the alignment is evaluated

using a minimum-number-of-words criterion. If the number of aligned words exceeds this criteria, the sequence is considered to be aligned and is removed from subsequent cycles. If the criteria is not met, the audio is repartitioned and the cycle is repeated for the unaligned segments. For improved performance, the acoustic models used for speech recognition are adapted at each iteration using information extracted from the reliably aligned regions. This is achieved by applying Maximum Likelihood Linear Regression in two stages. The first stage involves training a global transformation, after which a class-based transformation is built for the phonemes corresponding to the reliably aligned content. Thus, the three core phases of the algorithm are: recognition, alignment, adaptation. These are repeated five times, with each iteration improving both the overall alignment and the acoustic model. This approach has demonstrated strong results, outperforming Viterbi-based approaches and achieving $> 70\%$ correctly aligned words on noisy audio and corrupted transcription data.

Another approach, proposed by Anguera *et al.* [4], is perhaps more applicable to this work. This is as it is specifically designed for applications with very limited resources, as is the case with audio to text alignment for a single feature film. In this work, the audio is first time stretched using the synchronous overlap-add algorithm [136] in order to reduce the rate of speech, as this improves alignment. The audio is then segmented into segments of up to 30 s in length. These audio segments are then decoded using a phoneme recogniser trained using a Hungarian language model. The segment outputs are then converted back to their original time-base before being concatenated into a single sequence. For the text processing, the textual input is first normalised to produced a set of individual grapheme-like symbols. The resulting phoneme and grapheme sequences are then aligned using dynamic programming to find the optimal global alignment between both strings. This achieved by first constructing a global distance matrix of cost functions between phoneme/grapheme data, and then tracking back through the matrix using dynamic programming to find the optimal alignment. The use of dynamic programming is particularly useful here as it is able to account for inaccuracies in the matrix (e.g. incorrect transcriptions), providing a correct global alignment even in the case of noisy data. Evaluation of this approach demonstrated strong results when applied to Catalan and Spanish data, achieving errors of $< 5\%$ and $< 10\%$ on pooled utterances of Spanish and Catalan respectively. As this approach relies on a language model, it is not wholly appropriate for our target application, as this work looks to develop an approach which does not rely on a language model. Nevertheless, this demonstrates that the process of tokenization followed by dynamic programming achieves strong alignment results. This is relevant as this work could adopt a similar approach, albeit without the use of a language model for the tokenization step.

In Hazen *et al.*'s work [51], a method for aligning transcripts and long speech recordings is proposed. As with Anguera *et al.*'s work [4], this approach first applies ASR to the audio, this time via the SUMMIT ASR system [40][97]. For improved performance on the target data, the transcript is used to heavily bias the recogniser's language model. The ASR produces word-level tokenisation of the audio data, which is then used with the transcript information to find corresponding points between the speech and text data. These points are then used as anchors for the alignment stage. The alignment process, termed pseudo-forced alignment, aims to find an optimal alignment for the speech and text data while allowing for transcriptions errors. As such, this process allows for word insertion or deletion as well as substitution of words which exist in the transcript. This is achieved using a phonetic-based out-of-vocabulary word filler model [9] and a finite state transducer (FST). The FST model allows the process of data manipulation to be controlled via penalty weights in order to ensure that correct words are rarely modified. After this phase, the transcript and speech data are aligned, and segments containing substitutions, insertions or deletions are marked. The ASR is then re-applied to these segments to provide estimates of the substituted or inserted content using the ASR's language model. This process also uses a FST network, this time using the FST to moderate insertions and substitutions by allowing marked segments to be modified while preserving segments in which no insertions or substitutions have been made. This approach has demonstrated strong performance, dramatically reducing the ASR's word error rate from 24.3% to just 8.8%. As with the other audio-to-text alignment methods explored here, this relies on a language model, making it unsuitable for the target application of this work. However, this highlights a number of useful mechanisms for alignment. The first of these is the use of anchor points between the data - points of higher confidence which can be used to align information across different media. Secondly, this work introduces the concept of iteratively correcting the alignment based on previous alignments steps. This incremental approach to alignment would likely be useful as it allows for higher confidence segments to be aligned first, thereby improving the efficacy of alignment when subsequently aligning lower confidence regions as a general alignment has already been achieved.

Work by Lyu *et al.* [83] proposes a method for aligning speech and audio from separate languages is proposed. This has been developed for the alignment of Taiwanese and Mandarin content. The approach first segments the audio content into a number of semantically significant segments - in this case, the segments correspond to individual news stories. A speech recogniser is then applied to transcribe the speech information into Taiwanese spoken syllables. Machine translation is then used to provide word-by-word translation of the corresponding Mandarin document into a sequence on Taiwanese

tonal syllables, after which the resulting word sequences are aligned using Dynamic Time Warping (DTW) [12]. The advantage of using DTW is that it is capable of length-invariant sequence alignment, making it ideal for applications involving data with different time bases. This method for cross-lingual alignment proved to be fairly successful, achieving an alignment accuracy of 82.5% (33/40 words correctly aligned). This was achieved despite poor performance of the ASR, which exhibited an accuracy of only 57.7%, demonstrating that DTW was able to achieve strong general results despite noisy data. While this method is interesting in that it tackles the issue of cross-lingual alignment, as with the other methods reviewed, it again relies on a language model. Despite this, the work clearly highlights the capabilities of DTW, with it achieving strong alignment results despite noisy data. Given this, DTW would likely be an ideal method for cross-media alignment of audio and video or subtitle information, particularly given a system which aims to achieve alignment by language-independent means.

2.3.2 Automatic Speech Alignment

For current multimedia workflows, automatic speech alignment via audio-to-audio alignment is often used during the ADR process. This facilitates quick substitution of soundstage recordings with the ADR recordings by automatically aligning audio by means of spectral content. This has been a key interest within ADR for many years, with early attempts to automate the process being made in the 1980's [15].

More recently, systems such as WordFit [86] and VocAlign PRO [79] proved that it was possible to use dynamic programming to automatically align similar sections of audio, though the systems only worked successfully under ideal circumstances [116]; i.e. between two signals of similar spacings under low noise conditions. Several approaches were explored to improve upon the performance of automatic speech alignment [18][102][32] through the use of dynamic time warping (DTW) - an algorithm designed to find an optimal alignment between two sequences.

DTW works by finding the optimal path through a cost matrix, as demonstrated in Figure 2.9. The cost matrix is obtained by computing a distance value (or cost) for each feature in a query sequence, a, to each feature in a reference sequence, b. For audio-to-audio alignment, these features are typically spectro-temporal features [116][137][117]. Once the cost matrix has been constructed, dynamic programming is used to find the optimal path through the matrix from $(0, 0)$ to $(n_1 - 1, n_2 - 1)$, for signals of length n_1 and n_2. This produces a warping path - a path mapping the query signal to the reference signal - which can be used to align the signals. In the case of automatic speech alignment, this warping

path is used to inform how the audio query signal is time-stretched or time-compressed to fit the reference audio.

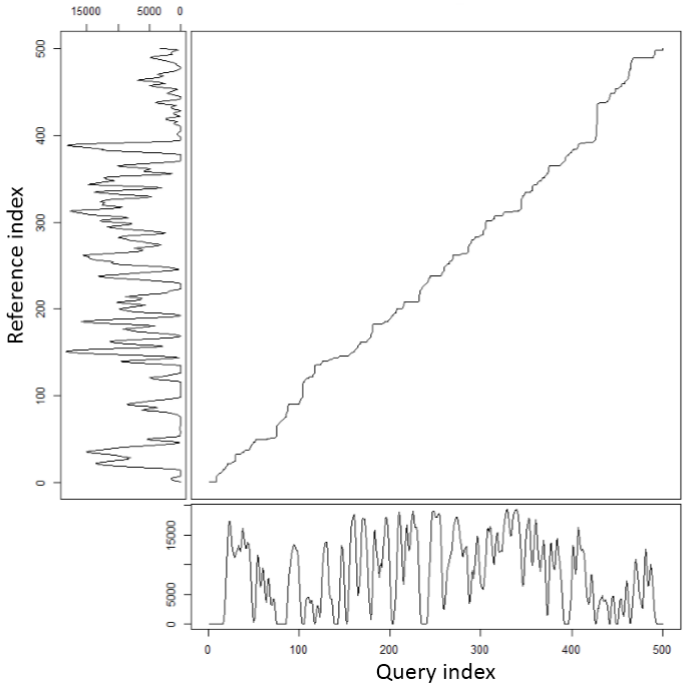


Figure 2.9: Diagram path through cost matrix mapping query signal to reference signal produced by DTW.

While applications of DTW in Soens *et al.*'s work [116], WordFit [86] and VocAlign PRO [79] were successful in identifying similarities for audio-to-audio alignment tasks, they did not address the issue of spacing, and thus was not consistently accurate, particularly when timing differences between speech sequences were large.

Further work explored the incorporation of prosodic features within speech alignment [128], however the performance of these approaches was also found to be inconsistent, often resulting in distortions relating to acoustic differences. Later developments involved the use of linear predictive coding (LPC) and standard DTW to achieve automatic post-synchronisation of speech signals through detecting and aligning spectral content using LPC cepstral vectors [127]. The signal could then be time-scaled through the use of the Waveform Similarity Synchronized Overlap Add algorithm (WSOLA) [129]. This system proved to be more effective when presented with timing differences between utterances, however time-stretched results often produced notable distortions. Further work saw this approach adopted and expanded upon through applying the DP principals to classify the signals into segments of speech and non-speech. While this was more successful in certain

scenarios, it also resulted in significant misalignments due to the alignment getting stuck in local minima.

These algorithms all contend with drawbacks which produce unnatural distortions in the resulting signal. This stems from the requirement of the warping function curvature: this should allow very steep or flat gradients to account for temporal differences in speech signals, but it should also be smooth enough to avoid unnatural sounding artefacts in the time-aligned results. This was addressed in Soens *et al.*'s work on the development of split dynamic time warping [116][117]. The algorithm first segments the two waveforms into speech and non-speech segments. The reference speech segments are then delimited by time markers (α_r, β_r), with $1 \leq r \leq R$. For each of these, there must correspond a replacement segment $(\lambda_{r-1}, \lambda_r)$. Corresponding pairs can be found by splitting the replacement speech waveform in which all non-speech segments have been removed in a pre-processing step at time instants:

$$\lambda_r = \frac{\int_{\beta_r}^{\alpha_{r+1}} g(x)\tau(x)dx}{\int_{\beta_r}^{\alpha_{r+1}} g(x)dx} \tag{2.14}$$

for $r = 1, 2, 3, ...R-1$, where $\lambda_0 = 0$ and λ_R is the duration of the reduced replacement. In this expression, $\tau(x)$ represents the linearly interpolated DTW path between the reference (along the x-axis) and the reduced replacement using the symmetric Sakoe-Chiba local constraint [111]. $g(x)$ is a Gaussian weighting function that is used to bias the split towards the speech segment boundaries.

The second step comprises of the recalculation of the timing relationships for each pair of matching speech segments. This uses the same DTW algorithm as step 1 using a Sakoe-Chiba band to speed up re-computation [111]. To reduce distortions in the resulting signal, the warping path is smoothed using locally weighted regression (LOWESS) [21]. The effects of smoothing on DTW alignment are illustrated in Figure 2.10.

LOWESS smoothing proved to be both highly effective and computationally efficient. Despite this, the process still produced occasional artefacts, as such a correction procedure was also implemented to improve the consistency of natural sounding results by correcting for unnatural deceleration and acceleration generated by the smoothing algorithm.

In performance evaluations the system demonstrated significant advantages over the industry benchmark, VocAlign [117], with performance improvements of 44.8% in lip-sync accuracy and 51.9% in speech quality over the baseline system [116]. Improvements were primarily found when using the system on speech signals with large structural timing differences.

The approaches discussed here all tackle the problem of audio-to-audio speech align-

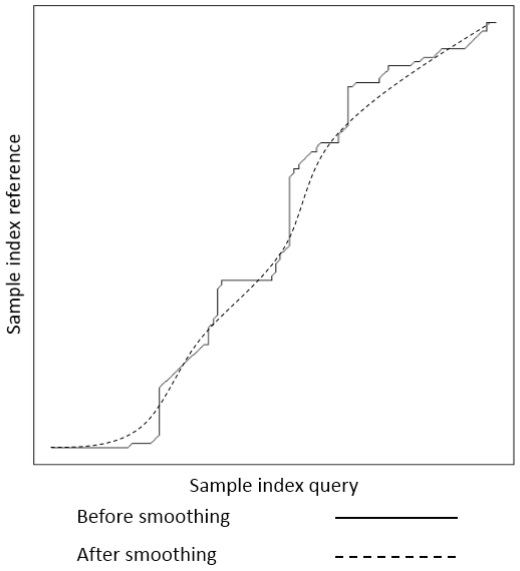


Figure 2.10: Illustration of smoothing process from [116].

ment through finding associations between spectro-temporal features. While this does not relate directly to the core aims of the project, this review has been crucial in demonstrating the capabilities of DTW - an algorithm which continues to be a crucial component within audio alignment systems. This further supports the notion that DTW and related dynamic programming-based methods are robust for multimedia alignment tasks, making them strong candidates for use within this work.

2.3.3 Summary

This section has explored a number of methods used for alignment, with a focus on audio feature alignment and audio-to-text alignment. While some of the audio-to-text approaches required relatively few resources, all of the audio-to-text alignment methods relied on a language model for audio tokenisation. This makes them unsuitable for solving audio-to-text alignment by language independent means. However, these approaches have introduced a number of concepts that will be useful for developing an alignment solution. One key method used in alignment in both audio-to-audio and audio-to-text alignment is dynamic programming sequence alignment. This has demonstrated strong performance for aligning tokenised sequences in Anguera *et al.*'s work [4], and is a crucial component in DTW, which itself is central to all audio-to-audio alignment methods explored. Another useful method is presented by Hazen *et al.* [51], which uses an incremental framework

to iteratively correct sequence alignment - first starting with a general estimate, before improving alignment by re-analysing specific regions. Lastly, DTW has proven to be a robust approach for sequence alignment, being applied for text-to-speech alignment in work by Lyu *et al.* [83], as well is in all of the automatic speech alignment processes explored.

In conclusion, while none of the literature explored tackled the exact problem of language-independent alignment, the approaches reviewed have all been valuable in highlighting useful methods for sequence alignment. As such, this work will look to incorporate these approaches within a solution developed for the language-independent alignment task.

2.4 Conclusion

This chapter has covered three key areas of interest to the project: audio speech processing, computing vision techniques for speech processing, and methods feature matching and sequence alignment. The literature reviewed here has guided each of the core sections of this project, providing both conceptual guidance to inform the development of solutions, and baseline systems against which these solutions have been evaluated. Each section here has also been summarised, providing an overview of how the literature influenced project development. The following chapters will go on to present the development of methods in audio speech processing, visual speech processing and sequence alignment which form the core contributions of this thesis.

Chapter 3

Detecting Speech in Entertainment Audio

3.1 Introduction

In order to utilise speech within an audio-based alignment system, it is first necessary to identify speech segments within the audio signal. As discussed in Chapter 2, numerous methods exist for speech detection, however many of these are incapable of robust speech detection within entertainment media. This is largely due to the variety of audio phenomena present in entertainment media, including atmospheric noise and sound effects [55], and audio effects applied to speech during post-production, such as distortion and pitch shifting [131].

This section discusses the development of a Voice Activity Detection (VAD) approach designed to detect speech in complex mixed audio signals, with a specific focus on solving the problem of speech classification in entertainment media. Furthermore, this approach looks to solve the problem of speech classification without the use of a language model, to allow the work to be incorporated into a language-independent approach for audio-visual matching and synchronisation.

The chapter begins with an introduction of the datasets used for the VAD development, before describing the concepts behind the key competing approach and a novel speech classification method developed through this work. The chapter goes on to present the results of classifier tuning, after which results are presented in the form of an initial investigation and a comprehensive evaluation which compares the proposed approach with current state of the art. The chapter concludes with an investigation into how the amount of data impacts classifier performance, and an examination of the method's performance on

Film	Genre	Dataset
Constantine	Drama/Fantasy/Horror	Partial Film Dataset
Shrek 3	Animation/Adventure/Comedy	Partial Film Dataset
Knocked Up	Comedy/Drama/Romance	Partial Film Dataset
Blood Diamond	Adventure/Drama/Thriller	Partial Film Dataset
I Am Legend	Drama/Sci-Fi/Thriller	Whole Film Dataset 1 & 2
The Bourne Identity	Action/Mystery/Thriller	Whole Film Dataset 1 & 2
Kill Bill Vol. 1	Action	Whole Film Dataset 1 & 2
Saving Private Ryan	Action/Drama/War	Whole Film Dataset 1 & 2
Disney's Hercules	Animation/Adventure/Comedy	Whole Film Dataset 2
The Fellowship of the Ring	Adventure/Drama/Fantasy	Whole Film Dataset 2

Table 3.1: Dataset content and film genres. Genre labels according to The Internet Movie Database (IMDb) [56]

non-English speech data.

3.2 Datasets

Three dataset configurations have been used for audio voice activity detection investigations. Each comprises a variety of genres to give a comprehensive impression of voice activity detection performance on a range of content. Each dataset has been manually annotated to provide a human-defined ground-truth against which the predicted output of VAD approaches can be evaluated. The datasets are as follows:

- Partial Film Dataset
- Whole Film Dataset 1
- Whole Film Dataset 2

The composition of each dataset and the genre labels of each film are given in Table 3.1.

The Partial Film Dataset was created to facilitate fast prototyping and classifier tuning on a modest but diverse dataset. The films were specifically chosen to cover a range of genres in order to incorporate a variety of sound design approaches, sound effects and voice types. For each film in the dataset, 10 minutes of speech data has been obtained through manual annotation of the film audio, and 20 minutes of non-speech data has been obtained by the same means. This was done by going through the film content from the beginning until 10 and 20 minutes of speech and non-speech data were obtained respectively. The

speech and non-speech segments were then concatenated to form two blocks of data per film: one 10 minute block of speech, and one 20 minute block of non-speech.

The ratio of speech to non-speech was determined following investigations into the speech/non-speech ratio within films using Whole Film Dataset 2. As demonstrated in Table 3.2, speech typically comprises between 20-30% of the total audio track for the films investigated, with some films exhibiting more dialogue at up to 45% speech content. In order to ensure an accurate impression of classifier performance, the upper-bound of 45% was used as a guideline, and thus the Partial Film Dataset was constructed using a 2:1 ratio of non-speech to speech. This was done to ensure a significant degree of positive classifications existed within the dataset, thus helping to ensure a more accurate impression of classifier performance.

Whole Film Dataset 1 was constructed according to the films used in Eyben *et al.*'s work [36]. This allowed for direct comparison between the VAD approaches explored here and the approaches investigated in their paper. Whole Film Dataset 2 is an extension of the first, adding films of the Fantasy and Animation genres to further diversify the content. In total, Whole Film Dataset 2 comprises approximately 13 hours of audio data.

To ensure optimal accuracy when annotating the data, both audio and visual speech information was used to inform the annotation process (by listening to the speech, and watching visual speech cues when these were available). While much of the data could be clearly differentiated into one of the two classes, the annotation process introduced a number of challenges. Firstly, judgement had to be made as to whether background speech should be classified as speech, as this may not be present in the dialogue. The decision was made that, if the speech was intelligible, it should be classified as speech, however if it was unintelligible (e.g. crowd noise), it would not be classified as speech. This is as classifying intelligible background speech as non-speech could impact the classifier's performance on content with quieter speech segments, or where the speech is purposefully engineered to sound distant. Secondly, a decision had to be made as to whether brief vocalisations (such as grunts) should be classified as speech. The decision here was to classify foreground vocalisations (i.e. close-miked) as speech, as often short periods of dialogue, e.g. one or two words, are likely to have similar spectro-temporal qualities to these kinds of vocalisations. As such, these were classified as speech to try and prevent the rejection of potentially useful speech information. Lastly, there was the issue of where to cut the audio: if cut at non-zero points, this would introduce time-domain artefacts which could negatively impact feature extraction. As such, the audio was cut on zero-crossing points to minimise the introduction of distortions from data segmentation.

Film	Speech
I Am Legend	18.26%
The Bourne Identity	22.87%
Kill Bill	23.46%
Saving Private Ryan	30.74%
Disney's Hercules	45.22%
The Fellowship of the Ring	27.96%

Table 3.2: Speech percentage per film for Whole Film Dataset 2.

3.3 Machine Learning Techniques for Voice Activity Detection

Due to the challenging nature of complex mixed audio signals, such as described in work by Sonnleitner *et al.* [119] and Eyben *et al.* [36], machine learning approaches for VAD have been explored. The techniques discussed here use signal processing methods to extract audio features prior to training binary classifiers to discriminate between speech and non-speech information. The first technique explored was developed specifically for speech discrimination in multimedia content, and is described by Sonnleitner *et al.* [119]. In Section 3.3.2, a novel VAD method is introduced which uses a correlation matrix-based approach for feature selection and dimensionality reduction. While the former approach demonstrates reasonable performance, the latter achieves particularly encouraging results, outperforming contemporary and state of the art approaches on a range of feature films.

3.3.1 Sonnleitner et al.

One of the most successful methods for speech classification found in the literature is the approach described by Sonnleitner *et al.*, which achieves accuracies of $> 97\%$ on speech classification tasks. The approach was developed to classify radio broadcasts as either speech or music - the two predominant types of content in radio material. The approach uses Short Time Fourier Transforms (STFTs) to exploit spectro-temporal variations of speech signals in order to discriminate between speech and non-speech content. More specifically, the approach exploits the sustained harmonic information in music, and the lack of sustained harmonic information in speech, to discriminate between the two content classes.

Audio Processing and Spectral Features

The approach compares adjacent STFT audio frames to determine the amount of spectral fluctuation, with low fluctuation indicating music, and high fluctuation indicating speech. The STFT here uses a log frequency axis in order to improve sensitivity to musical structures. This is achieved by computing the cross-correlation between time frames $\mathbf{x}_t$ and $\mathbf{x}_{t+offset}$, whereby the cross-correlation is used to estimate the degree of correlation between shifted versions of the vectors for a range of lags l, where, for two vectors **x** and **y** of length N, the cross-correlation for all lags $l \in [-N, N]$, gives a cross-correlation series of length $2N + 1$. The cross correlation is given as:

$$R_{\mathbf{xy}}(l) = \sum_i x_i y_{i+l} \tag{3.1}$$

Input vectors **x** and **y** correspond to time frames, and the lag corresponds to a shift in frequency content. r_{xcorr} is defined as the maximum cross-correlation over a range of lags, as given by:

$$r_{xcorr}(\mathbf{x}_t, \mathbf{x}_{t+offset}) = \max_1 R_{\mathbf{x}_t, \mathbf{x}_{t+offset}}(l) \tag{3.2}$$

where the lag is denoted by $l \in [-l_{\max}, l_{\max}]$, which corresponds to the frequency shift between bins. r is defined as a special case of the cross-correlation, where lag $l = 0$, thus zero-lag cross-correlation is defined as correlation, i.e.:

$$r(\mathbf{x}_t, \mathbf{x}_{t+offset}) = R_{\mathbf{x}_t, \mathbf{x}_{t+offset}}(0) \tag{3.3}$$

Correlation gain, defined as $r_{xcorr} - r$, is used to indicate the prevalence of speech within the signal. This provides a measure of spectral fluctuation, whereby cross-correlation is maximal at frequency lag $l = 0$, thus the gain $r_{xcorr} - r = R(0) - r = r - r = 0$. Hence, for signals with greater spectral fluctuation (such as speech), $R(l)$ will be maximal for some value of $l \neq 0$, and the gain $r_{xcorr} - r$ will be positive. Likewise, the gain will be lower for audio in which music is dominant, as the cross-correlation between frames is greater. As harmonic content is a key factor in discriminating between speech and music, a log scale has been used to ensure that harmonic relationships are represented as constant offsets to the fundamental frequency, thus allowing continuous frequency changes and harmonic relationships to be picked up by cross-correlation between frames. The paper describes using audio data sampled at 22.05 kHz, to which an STFT is applied via a Kaiser window of size 4096 samples. The magnitude spectrum $|Z(f)|$ is then computed, and the STFT magnitude spectrum is mapped to the logarithmic cent scale. The features considered

are the first 150 frequency bins of the spectrum, corresponding to 0 Hz to approximately 800 Hz. The paper also describes the use of a maximum lag parameter, $l_{\max}$, which has a value of 3 in order to account for chromatic sequences within musical content [119]. The final implementation evaluates audio at 200 ms intervals using a frame of 50 STFT blocks around each observation point. For each STFT block, two feature vectors are computed: the cross-correlation, xc, and the correlation, c. The difference of the vectors provides the feature vector $r = xc - c$, after which r is smoothed using a rectangular window of width 5. The index of the dominant frequency bin within the observation window is then appended to the feature vector, thus resulting in a vector of 48 feature values per observation. The features were used to train a random forest classifier with 200 estimators (trees) and 10 features per tree. This was trained on an annotated set of 21 hours of randomly selected audio from radio stations.

Modifications to Approach

The approach described in Sonnleitner *et al.*'s paper [119] uses a sliding median window with a duration of approximately 10 s. While this yields strong results, entertainment media contains brief segments of speech, with many utterances of sub-10 s duration [38]. As such, adjusting the output using a 10 s long median window does not provide the desired resolution. For the implementation used here, the median window has been removed in order to provide the classification resolution necessary for identifying shorter utterances.

3.3.2 MFCC Cross-Covariance Features

While successful in its application context, the approach proposed by Sonnleiter *et al.* [119] is not as well tailored to film audio. This is as feature film audio is far more variable than radio audio, as it is comprised of a more diverse mixture of audio sources. This includes music, speech and sound effects, whereas radio broadcasts typically only contain either speech or music. The complexity of the problem is further compounded by the fact that different audio sources can occur concurrently in feature films, making it desirable to identify dialogue amongst other audio phenomena.

To address this, a novel approach for speech detection in complex mixed audio signals is proposed. The approach uses cross-correlation of MFCC features applied to an annotated dataset to extract feature pairs with the greatest inter-set cross-correlation difference between speech and non-speech data.

The motivation behind these features is drawn from the principles underlying Sonnleitner's work [119], principally the observation that speech has distinct spectro-temporal

characteristics when compared to other types of audio information (this is illustrated in Chapter 2, Figure 2.1). It is clear from other speech processing techniques, such as the method presented by Kinnunen *et al.* [69], that this quality can be exploited by utilising speech activity in different critical bands. This is as speech affects certain critical bands more than others, meaning that it should be possible to determine the presence of speech by examining the correlation between a critical band which is known to be more sensitive to speech with one that is known to be less sensitive to speech. Thus, the principle here is to find pairs of critical band coefficients (in this case MFCCs) which have the greatest difference in correlation between speech and non-speech data. In doing so, it is possible to pick out the coefficient pairs whose correlation is most significantly affected by the presence or absence of speech, and thus the coefficient pairs which are likely to be most informative for the speech classification task.

Once the inter-set MFCC correlation differences have been obtained, the most significant *n* features are processed to produce MFCC Cross-Covariance (MFCC CC) features. These features are then used to train a binary classifier. Two machine learning approaches have been investigated for binary classification: support vector machines (SVM) and random forests. Investigations demonstrate that random forests outperform SVMs on speech classification within entertainment media.

Audio Processing and Spectral Features

A number of spectral features were considered for audio feature extraction, including MFCCs, Linear Predictive Coefficients (LPCs), and STFTs. MFCCs were chosen as the feature for a number of reasons:

1. Their prevalence in the literature for speech processing tasks [35][87][69].

2. The use of perceptual scaling (the Mel scale).

3. The fact that they are cepstral features, rather than spectral features.

The use of Mel scale features has several key advantages. The first is that there is a strong link between the human voice and the human auditory system [11]. Secondly, entertainment audio is intentionally mixed by humans, to be experienced by humans - this implies a strong relationship between the data and human auditory perception. Given these two factors, it is sensible to assume that perceptually scaled features would be optimal for speech detection tasks in this context. Lastly, the fact that they are cepstral features means that they are sensitive to patterns in the frequency domain, and previous work has

demonstrated that frequency-domain patterns are useful for discriminating between speech and non-speech information (Section 2.1.2).

For this work, stereo audio signals recorded at 44.1 kHz at a bit depth of 16 bits were used. The MFCC features are extracted by first applying a sliding window to the data with frame size 25 ms (1102 samples), with a step of 10 ms (441 samples) between frames. The Discrete Fourier Transform (DFT) for each frame is obtained via:

$$S(k) = \sum_{n=1}^{N} s(n)h(n)e^{j2\pi kn/N} 1 \leq k \leq K \tag{3.4}$$

where $h(n)$ is a sample window of length N, K is the DFT length, and k is an integer corresponding to the DFT bin number [115]. The power spectrum for the frame $s(n)$ is obtained by:

$$P(k) = \frac{1}{N}|S(k)|^2 \tag{3.5}$$

A set of triangular filters are then applied to the power spectrum signal to obtain the Mel-spaced filterbank. This is done by first converting the frequency range, f, to the Mel scale by:

$$M(f) = 1125ln(1 + f/700) \tag{3.6}$$

A 26-vector Mel-spaced filterbank is constructed according to the linearly spaced Mel frequencies via:

$$H_m(k) = \begin{cases} 0 & k < f(m-1) \\ \frac{k-f(m-1)}{f(m)-f(m-1)} & f(m-1) \leq k \leq f(m) \\ \frac{f(m+1)-k}{f(m+1)-f(m)} & f(m) \leq k \leq f(m+1) \\ 0 & k > f(m+1) \end{cases} \tag{3.7}$$

where M denotes the number of filters and $f()$ is a list of $M+2$ Mel-spaced frequencies [81]. This produces a Mel filterbank with an upper-frequency of 22.05 kHz, as shown in Figure 3.1.

In order to obtain the filterbank energies, each band of the Mel filterbank is multiplied with the power spectrum, and resulting coefficients are summed together. This produces 26 filterbank energy coefficients which describe the energy content of each filterbank. The log of each filterbank energy coefficient is then computed, producing 26 log filterbank energy values. A discrete cosine transform (DCT) is applied to these values, producing 26 MFCCs, of which the lower 13 MFCCs are used.

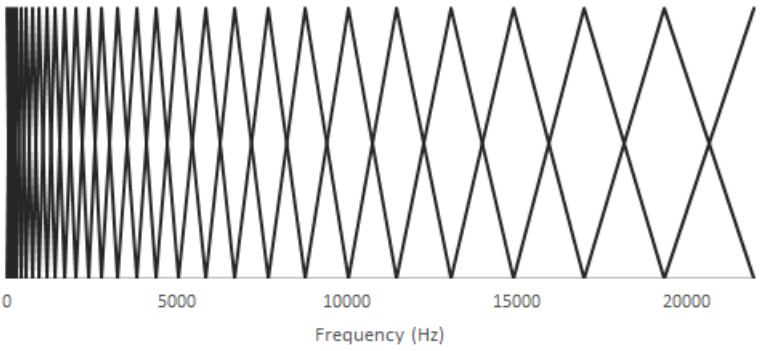


Figure 3.1: Diagram of Mel-scale filterbank.

For investigations using the Python programming language, the Python Speech Features library MFCC implementation has been used [82].

Feature Selection and Feature Vector Processing

The first phase in MFCC CC feature extraction is to obtain MFCC feature pairs which demonstrate significant difference in correlation between speech and non-speech data. These are the top n feature pairs with greatest difference in correlation, d, between speech and non-speech data sets. Correlation is calculated as the Pearson product-moment correlation coefficient, ρ. This is computed for all pairs of MFCCs in both the speech and non-speech data sets. Each correlation coefficient ρ in $\mathbf{P}$ is obtained from the covariance matrix, $\mathbf{\Sigma}$, of a pair of MFCC features, via the coefficient matrix $\mathbf{P}$:

$$P_{ij} = \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii} \times \Sigma_{jj}}} \tag{3.8}$$

The correlation coefficient has a value between -1 and 1, where 1 denotes total positive correlation, and -1 denotes total negative correlation. Two matrices are produced as a result of this phase: one inter-MFCC correlation matrix for speech data, $\mathbf{MFC}^{s}$, and one inter-MFCC correlation matrix for non-speech data, $\mathbf{MFC}^{ns}$. The difference matrix, $\mathbf{D}$, is obtained simply by:

$$\begin{aligned} \mathbf{S} &= \mathbf{MFC}^{s} - \mathbf{MFC}^{ns} \\ \mathbf{D} &= (|S_{ij}|) \end{aligned} \tag{3.9}$$

Higher values of d_{ij}, for $d_{ij} \in \mathbf{D}$, indicate greater variance in the MFCC pair relationships between speech and non-speech data, as shown in Figure 3.2. This in turn indicates that the pairs are more likely to provide information relating to the presence or absence of speech spectral data, thus facilitating more effective speech/non-speech discrimination.

MFCC	0	1	2	3	4	5	6	7	8	9	10	11	12
0	0	0.04	0.1	0.21	0.18	0.06	0.34	0	0.06	0.23	0.12	0.21	0.1
1	0.04	0	0.1	0.01	0.31	0.06	0.11	0.02	0.2	0.02	0.19	0.08	0.15
2	0.1	0.1	0	0.06	0.13	0.08	0.08	0	0.24	0.12	0.15	0.13	0.04
3	0.21	0.01	0.06	0	0.12	0.15	0.04	0.13	0.04	0.03	0.1	0.07	0.09
4	0.18	0.31	0.13	0.12	0	0.04	0.07	0.04	0.04	0.01	0.03	0.01	0.11
5	0.06	0.06	0.08	0.15	0.04	0	0.1	0.12	0.03	0.04	0.08	0.06	0.02
6	0.34	0.11	0.08	0.04	0.07	0.1	0	0.2	0.03	0.1	0.07	0.1	0.2
7	0	0.02	0	0.13	0.04	0.12	0.2	0	0.07	0.13	0.07	0.17	0.17
8	0.06	0.2	0.24	0.04	0.04	0.03	0.03	0.07	0	0.1	0.19	0.05	0.18
9	0.23	0.02	0.12	0.03	0.01	0.04	0.1	0.13	0.1	0	0.07	0.15	0.03
10	0.12	0.19	0.15	0.1	0.03	0.08	0.07	0.07	0.19	0.07	0	0.12	0.2
11	0.21	0.08	0.13	0.07	0.01	0.06	0.1	0.17	0.05	0.15	0.12	0	0.17
12	0.1	0.15	0.04	0.09	0.11	0.02	0.2	0.17	0.18	0.03	0.2	0.17	0

Figure 3.2: Matrix of MFCC pair correlation coefficient differences between speech and non-speech data. Darker squares indicate greater values.

The MFCC CC features are obtained for n MFCC pairs which have the greatest values of d. The cross-covariance vector is obtained by computing the cross-covariance of segments of the two signals along their length via a rectangular sliding window:

$$\begin{gathered}(\mathbf{f} \times \mathbf{g})_i := \sum_j \mathbf{f}_j \times \mathbf{g}_{i+j} \\ \mathbf{f} = \mathbf{v}^a_{k:k+w}, \mathbf{g} = \mathbf{v}^b_{k:k+w}, \forall k \in K - w\end{gathered} \tag{3.10}$$

where $\mathbf{v^a}$ and $\mathbf{v^b}$ are the MFCC vectors, k is the index, K is the length of the vectors, and w is the length of the sliding window.

Previous work has demonstrated the importance of temporal information for speech classification problems [36][119]. As such, temporal information is incorporated through using a window size, w, of 450 ms. This duration was determined based on an average phoneme duration of ≈176 ms [64]. A window size of 450 ms is therefore long enough to facilitate the rejection of brief speech-like phenomena, while still allowing for the detection of finer resolution (sub-1s duration) speech features, such as words or sub-word phoneme strings.

Classifier Training and Feature Optimization

The resulting MFCC-CC feature vectors are used to train a binary classifier to discriminate between speech and non-speech information in entertainment audio. In this work, two classification approaches have been explored: random forests and support vector machines (SVMs). Grid-search optimization has been used to find the optimal parameters for both the random forest and SVM classifiers. For the random forest, the number of estimators (trees per forest) was varied, using a range of between 10 and 500 estimators.

For the SVM, linear, polynomial and Radial Basis Function (RBF) kernels were investigated. For each kernel, various values for the C parameter were investigated, along with kernel-specific parameters d and γ for polynomial and RBF kernels respectively. In the context of SVMs, C denotes the regularization parameter which controls the trade-off between margin maximisation and training data errors [13]. The kernels used for the SVM are defined as:

Linear kernel:

$$K(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T\mathbf{x}' + c \tag{3.11}$$

Polynomial kernel:

$$K(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^T\mathbf{x}' + 1)^d \tag{3.12}$$

where d is the order of the polynomial.

Gaussian RBF kernel:

$$K(\mathbf{x}, \mathbf{x}') = \exp(\gamma||\mathbf{x} - \mathbf{x}'||_2^2) \tag{3.13}$$

where the free parameter γ affects the degree of influence for individual training examples, thus moderating support vector selection.

The number of MFCC feature pairs used for MFCC-CC feature creation was also explored in order to find the optimal number of MFCC feature pairs. The following sections discuss parameter tuning and MFCC-CC feature optimization.

3.3.3 Evaluation Design

The Partial Film dataset was used for evaluation of machine learning approaches. For each approach, k-fold cross-validation was used for classifier tuning, with each iteration investigating a range of classifier parameters. This method facilitates a comprehensive

evaluation of classifier performance through maximising the usefulness of the data - giving an accurate impression of classifier performance over the whole range of data.

Each iteration uses 90 minutes of data for the training set (from three films), and 30 minutes of data for the test set (from the remaining film). This ensures that the classifier is naïve to the test data and maximizes testing cycles for the evaluation test set. The mean of the performance metrics for all four evaluations is then obtained and used for performance evaluation.

For the random forest, the key parameter investigated is the number of estimators (or trees) per forest. For the SVM, two kernels are investigated: radial basis function (RBF) and polynomial. The RBF kernel is tuned using a range of C and γ values, while the polynomial is tuned using a range of C and $degree$ values.

Both the random forest and SVM approaches evaluate the performs of the machine learning algorithms over a range of MFCC-CC feature vectors, in order to determine the optimal number of MFCC-CC vectors required for good classifier performance. To do so, MFCC-CC feature vectors were added to the feature vector in order of significance, with the most significant being the vector with the greatest correlation coefficient difference, d, across speech and non-speech data.

The statistical parameters used for evaluation are accuracy, precision, recall, F-score and Receiver Operating Characteristics (ROC) curves. These were chosen as accuracy gives a reasonable impression of performance, while F-score and ROC evaluations give an estimate of how well the classifier will generalize. The use of these methods is further supported by their prevalence in the literature [36][119][125].

3.3.4 Evaluation Results

This section discusses the results of the investigations into random forest and SVM-based classification methods.

Random Forest Classification

The random forest approach was investigated using a range of estimators (trees per forest) to determine the optimal parameter for performance. The number of estimators used ranged from 10 to 500.

Figure 3.3 clearly demonstrates appreciable performance gain between 10 and 100 estimators, with performance metrics stabilising at approximately 150 estimators. This indicates an optimal number of estimators of $>$150. Previous work on random forest-based speech classifiers has found 200 estimators to be optimal [119]. Thus, given observations

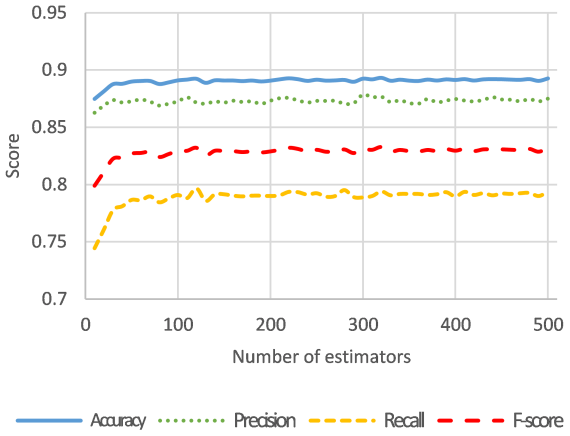


Figure 3.3: Random forest classification results using a range of estimators

from the literature and the results shown here, a value of 200 was chosen as the number of estimators for the random forest classifier.

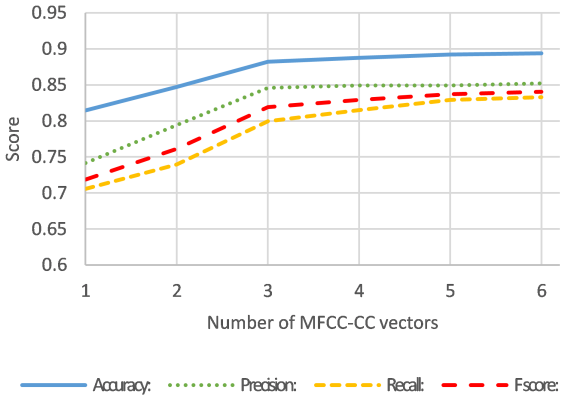


Figure 3.4: Random forest classification results using a range of MFCC-CC features

As demonstrated in Figure 3.4, pronounced improvement in performance can be observed between one and three MFCC-CC features, with performance levelling off at around five MFCC-CC features, after which performance enhancement is minimal. Therefore, 5 was deemed a suitable number of MFCC-CC features as this is past the point of significant performance gain by a reasonable margin.

Support Vector Machine Classification

The performance of the linear SVM was tested over a range of C values from 0.00001 to 100. As Figure 3.5 demonstrates, the best performance was obtained between C values

of 0.001 and 1.0, with a peak accuracy value of 0.864 at $C = 0.01$. The same trend can be observed for the F-score values in Figure 3.6, with better performance in the range $C = 0.001$ to 1.0, and performance trailing off thereafter. This is unsurprising, as higher values for C will tend to result in over-fitting, hence this drop-off in performance is likely attributed to the classifier beginning to develop a stronger bias to the training set.

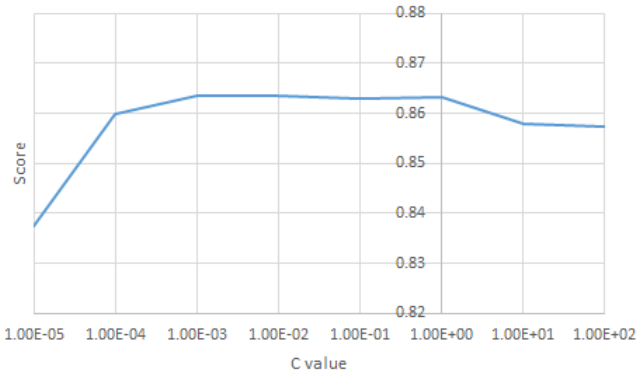


Figure 3.5: Accuracy scores for linear kernel SVM over a range of C values.

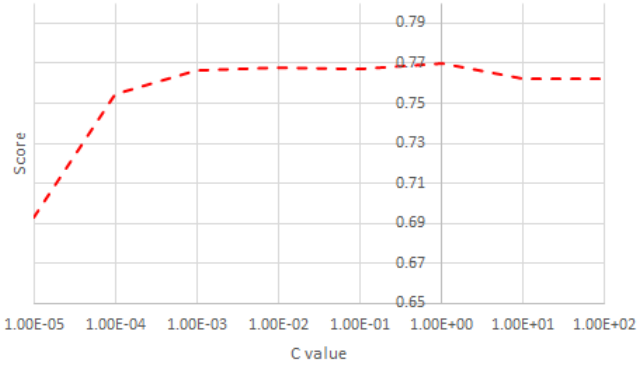


Figure 3.6: F-scores for linear kernel SVM over a range of C values.

As the heatmap in Figure 3.7 demonstrates, the polynomial kernel performed reasonably well, achieving a peak mean accuracy of 0.86 for speech classification on the Partial Film dataset. The peak accuracy statistics were obtained with low d values, and across a range of C values. This indicates optimal parameters of $C = 0.1$ to 1000 and $d = 1.0$. While the C parameter appears to be less critical - demonstrating strong performance over a range of values - low d values likely correspond to reduction in over-fitting of the SVM. As the d values are increased, poorer performance on the validation set becomes increasingly likely

due to the model's increased bias for the training set. This leads to over-fitting for higher values of d, and corresponds to the pattern observed in Figures 3.5 and 3.6.

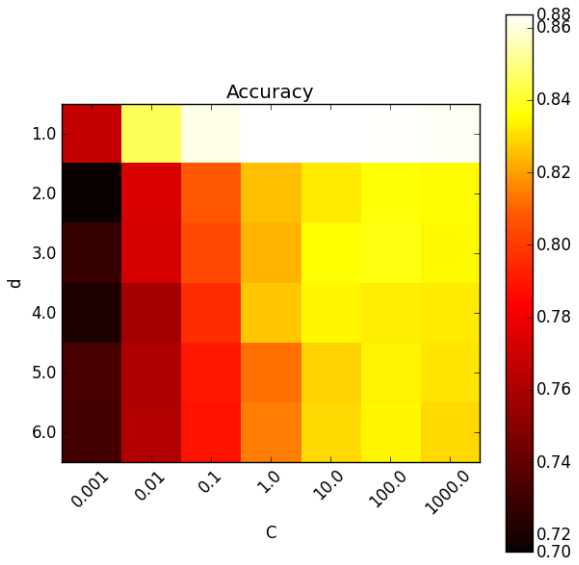


Figure 3.7: Heatmap of accuracy scores from SVM grid search with polynomial kernel SVM.

Figure 3.8 shows the performance of the RBF kernel SVM to be optimal over γ values of 0.001 and C values of 0.01 to 100. As with the polynomial kernel's d parameter, this range of γ values suggests that the classifier may begin to over-fit at higher γ values, resulting in increased training set bias. This effect is far more pronounced with the RBF kernel - with a sudden severe drop in accuracy once the parameters fall outside of their optimal values. Most importantly, this investigation demonstrates that the RBF kernel outperforms both the linear and polynomial kernels, as demonstrated in Table 3.3. As such, this section explores RBF kernel performance in more detail - examining the impact of the number of MFCC-CC features and classifier performance over a broader range of metrics.

Figure 3.9 shows classifier performance over a range of MFCC-CC features. Similarly to the results from the random forest classifier, the greatest degree of performance gain is achieved between 1 and 3 features, after which only marginal improvement in performance is observed. Unlike the random forest approach, the SVM demonstrates a slight dip in precison at 5 MFCC-CC features. Analysis of the scores on individual test sets demonstrates that this is due to a drop in precision score at 5 MFCC-CC features for the Shrek 3 data. Nevertheless, all other statistics demonstrate improvement between 4 and 5 features, hence

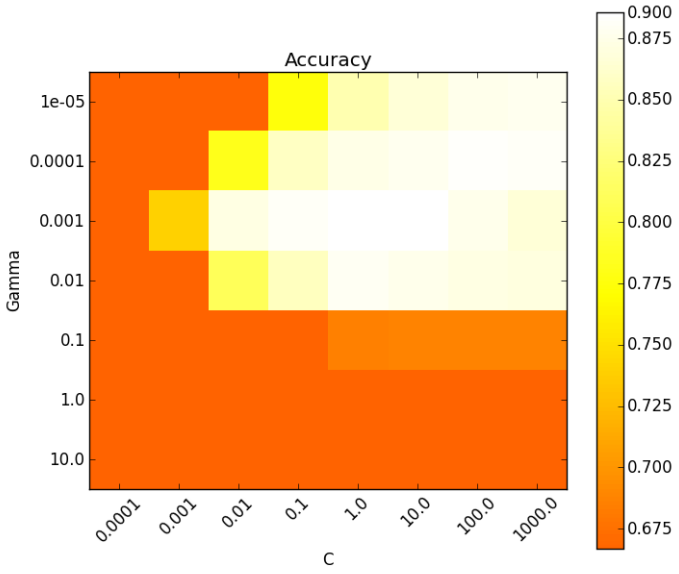


Figure 3.8: Heatmap of accuracy scores from SVM grid search with RBF kernel SVM using 5 MFCC-CC features.

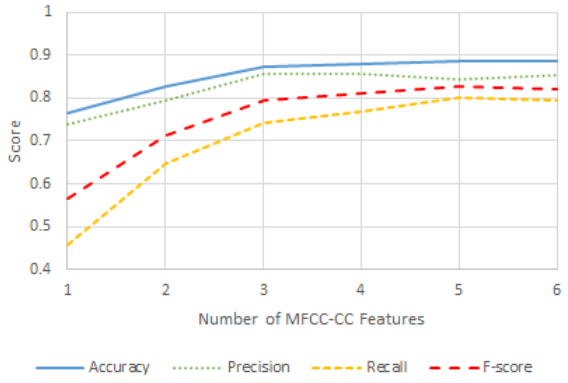


Figure 3.9: RBF kernel SVM performance over a range of MFCC-CC features using parameters $C = 1.0$ and $\gamma = 0.001$.

5 MFCC-CC features was selected as this is significantly past the 'knee' in performance gain, and fits the underlying trend for all but one of the test sets.

Conclusion

Optimal SVM performance was achieved using the RBF kernel with $\gamma = 0.001$ and $C = 1.0$, which outperformed the best polynomial score by $> 2\%$ across both accuracy

Classifier	**Accuracy**	**F-score**
SVM - Linear	0.864	0.768
SVM - RBF	0.886	0.827
SVM - Polynomial	0.864	0.769
Random Forest	**0.892**	**0.837**

Table 3.3: Results from tuned random forest and SVM approaches.

and F-score. While this performance is strong, it was exceeded marginally by the random forest classifier, which exhibited an accuracy of 0.892 to the SVM's accuracy of 0.886. Comparison of tables 3.4 and 3.7 shows that the random forest largely outperformed the SVM except for one film in the dataset for which it demonstrated some advantage on a few metrics. This was Knocked Up, for which the SVM achieved greater results in both accuracy and precision. The random forest also demonstrated slightly more consistent results, with all metrics following a consistent trend when evaluating the number of MFCC-CC features (whereas the SVM demonstrated a slight deviation from the trend with a drop in precision).

In order to assess the significance of the performance metrics presented here, the variance of the accuracy scores across the four films has been obtained. This has been done using 5 MFCC-CC features for both of the highest performing methods: the RBF SVM and the random forest classifier. For the random forest, a variance of 0.029 was obtained using 200 estimators. For the SVM, a variance of 0.001 was obtained using optimal SVM parameters: C = 1.0, and /gamma = 0.001. While the results are very close, the random forest has been chosen for future investigation as it is significantly faster to train than the SVM (with the random forest taking around 20 minutes compared the SVM taking up to several hours). Thus, the advantages in training time far outweigh the marginal gains of the low variance in results from the SVM. As such, a random forest with 250 estimators using 5 MFCC-CC features was selected as the classification algorithm of choice for the MFCC-CC VAD.

The evaluation provided here also serves to inform the degree of significance required for the MFCC-CC approach's performance to be considered superior to competing approaches. Thus, this work will only classify performance as improved if there is a minimum of a 0.03 improvement in performance between methods, in order to account for variation inherent to classification performance.

Test set	Accuracy	Precision	Recall	Fscore
Constantine	0.894	0.879	0.790	0.832
Shrek 3	0.845	0.781	0.741	0.761
Knocked Up	0.893	0.851	0.824	0.837
Blood Diamond	0.910	0.918	0.803	0.857
Mean	0.886	0.857	0.789	0.827

Table 3.4: Classification results from SVM with RBF kernel trained using optimal parameters from grid search cross-validation.

3.4 Experimental Design

Several investigations were undertaken to evaluate the performance of the MFCC-CC VAD with respect to other contemporary VAD approaches developed for entertainment media content. The initial investigation explores the performance of the MFCC-CC classifier and Sonnleitner *et al.*'s approach [119] on the Partial Film dataset, using leave-one-out cross-validation on two hours of data from the four feature films. This is followed by investigations using Whole Film Dataset 1, which explores the performance of the two classifiers on four whole feature films and compares their performance with the methods described by Eyben *et al.* [36]. This investigation uses the entire Partial Film dataset as training material, and Whole Film Dataset 1 as the test set. The final investigation uses leave-one-out cross-validation on Whole Film Dataset 2 in order to explore classifier performance on a greater range of feature films, and to explore the impact of training set size on classifier performance.

Each investigation evaluated performance using accuracy, precision, recall, F-score and ROC analysis.

3.5 Initial Investigation

3.5.1 *Sonnleitner* VAD

As Table 3.5 demonstrates, the approach from Sonnleitner *et al.*'s work [119] achieves a mean accuracy of 69.9%, and also demonstrates significantly lower values for precision, recall and F-score than reported in the paper. The foremost cause of this is likely the removal of the median filter, which would have had a smoothing effect on the data; reducing the impact of false positives and false negatives.

Another key factor is the type of data used. While the classifier may perform well when discriminating between music and speech, the approach may not be sophisticated

Test set	**Accuracy**	**Precision**	**Recall**	**Fscore**
Constantine	0.714	0.642	0.315	0.423
Shrek 3	0.701	0.642	0.228	0.337
Knocked Up	0.678	0.539	0.224	0.317
Blood Diamond	0.701	0.637	0.236	0.344
Mean	0.699	0.615	0.251	0.355

Table 3.5: Classification results from random forest trained on features described in Sonnleitner *et al.*'s work [119].

enough to achieve strong accuracy on entertainment media. This is because, unlike radio, which contains predominantly speech and music, entertainment media contains a variety of sounds, including modified voices, foley, and sound effects - as well as speech and music.

A further consideration is the difference in training set size between the example here and the training set used in Sonnleitner *et al.*'s work. While this likely contributed to the reduction in performance, the same training set has been used to evaluate this with respect to the MFCC-CC VAD, as such these results were deemed sufficient for comparison, as the MFCC-CC VAD has the same training set size constraints.

3.5.2 MFCC-CC VAD

The MFCC-CC VAD was developed to provide a VAD method with higher classification resolution than the approach detailed in Sonnleitner *et al.*'s paper [119], but with the aim of achieving better performance statistics than exhibited by the implementation in Section 3.5.1. Results from initial investigations were encouraging - as demonstrated when comparing Table 3.5 and 3.7, the MFCC-CC VAD significantly outperforms the implementation from the previous section, achieving accuracies $> 90\%$, and a mean accuracy of $\approx$89%. This is significantly better than the implementation based on Sonnleitner *et al.*'s paper, which achieves a peak accuracy of $\approx$71%, and a mean accuracy of $\approx$70% [119]. This performance advantage can be observed across all other metrics used, with particularly distinctive scores for the recall and F-score statistics. The low recall metrics for [119]'s approach suggests a high degree of misclassifications, particularly regarding false negatives - suggesting that it is unable to return a significant proportion of relevant classes. The poor precision score indicates that, while the approach returns some correct positive cases, it returns a significant proportion of false positives. In contrast, the MFCC-CC VAD demonstrates good results across all performance metrics, with a mean $> 80\%$ across all datasets for all statistics.

The MFCC-CC classifier was also evaluated using ROC curves (Figure 3.10), a common

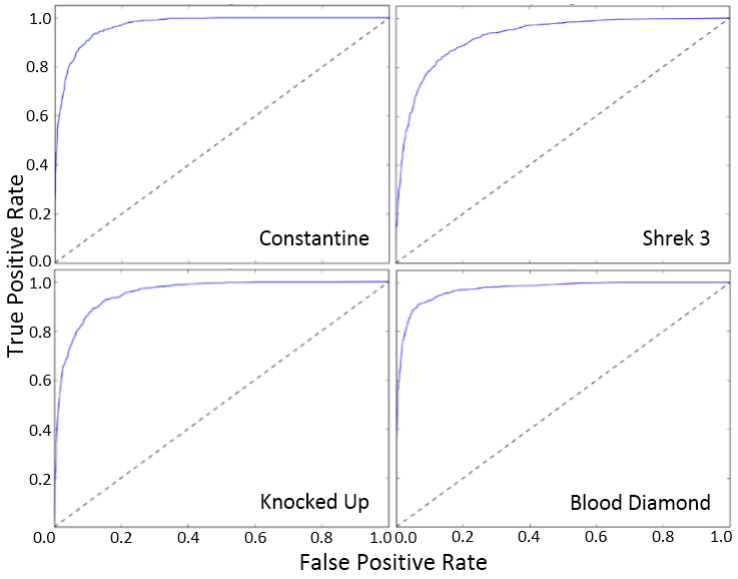


Figure 3.10: Receiver Operating Characteristic curves for MFCC-CC classification results from initial investigations.

Test set	Constantine	Shrek 3	Knocked Up	Blood Diamond	Mean
AUC	0.969	0.925	0.954	0.973	0.955
EER [%]	9.5	15.0	11.6	8.1	11.1

Table 3.6: Area under the curve and equal error rate from receiver operating characteristics plot.

method of assessing binary classifier performance. The ROC curves in Figure 4 indicate strong performance, with an average area under curve (AUC) of 0.955 (see Table 3.6), indicating that the classifier exhibits strong discrimination between the two classes. The Equal Error Rate (EER) observed here further indicates strong system accuracy, with an average EER of 11.1% achieved across the four test scenarios. This suggests better performance than the VAD in Eyben *et al.*'s work, which achieved an average EER of 33.2% on film audio data [36].

3.5.3 Conclusion

The approach from Sonnleitner *et al.*'s paper was particularly attractive given the classification results reported in the paper, and the fact that it was applied to more appropriate data than the other leading approach [36], which was developed for speech detection in acoustic environments (rather than for detecting speech in entertainment audio). Investigations into this approach demonstrate that it does not achieve statistics comparable to the paper when

Test set	Accuracy	Precision	Recall	Fscore
Constantine	0.903	0.902	0.794	0.844
Shrek 3	0.861	0.792	0.789	0.790
Knocked Up	0.881	0.783	0.889	0.833
Blood Diamond	0.924	0.920	0.845	0.881
Mean	0.892	0.849	0.829	0.837

Table 3.7: Classification results from random forest trained on MFCC-CC features.

applied to feature film audio data. While better results may be achievable with the addition of median filtering (as described in the paper), the median filtering would greatly limit classifier resolution, and is thus not desirable for the application scenario considered. As such, this approach was deemed unsuitable, and it was concluded that a more accurate classifier with greater classification resolution would be necessary.

The MFCC-CC approach was developed to provide the higher resolution functionality and to achieve more robust performance on the feature film audio data. Investigations into its performance on the partial film dataset demonstrate that it meets these requirements, significantly outperforming the approach proposed in the paper by Sonnleitner *et al.* [119] across all metrics and on all data. Further testing goes on to explore the performance of the MFCC-CC approach on a greater range of data, and to provide more comprehensive comparison against other contemporary approaches and state of the art.

3.6 Comparison with Contemporary and State of the Art Approaches

This section explores the performance of the MFCC-CC classifier on the first whole feature film dataset in order to provide a more comprehensive evaluation of its performance with respect to existing methods. The methods used for comparison were a long-standing state of the art VAD approach proposed by Sohn *et al.* [118], used to provide baseline performance statistics, as well as Eyben *et al.*'s [36] and Sonnleitner *et al.*'s [119] approaches, which have demonstrated strong performance on entertainment media.

Interestingly, the results in Table 3.8 indicate that the approach from Sonnleitner *et al.*'s work demonstrated competitive performance against both Eyben *et al.* and Sohn *et al.*'s work, despite the poor performance exhibited in Section 3.5. The MFCC-CC approach again exceeds the performance of Sonnleitner's approach - improving on all methods investigated, with greater AUC values and lower EER for all test sets. Another interesting point to note is that Eyben *et al.*'s approach yielded better results using the training set in

			AUC		
Test set	**Sohn***	**Eyben***	**Eyben**	**Sonnleitner**	**MFCC-CC**
I Am Legend	0.567	0.704	0.710	0.718	0.921
Kill Bill Vol. 1	0.554	0.627	0.642	0.800	0.893
Saving Private Ryan	0.577	0.743	0.708	0.717	0.946
The Bourne Identity	0.603	0.685	0.698	0.730	0.977
Mean	0.575	0.690	0.689	0.741	0.934
[%]			**EER**		
All	45.73	33.18	33.77	31.41	13.49

Table 3.8: Comparison of VAD approaches. * indicates results from [36] in which a different training set was used.

	AUC	
Test set	**Eyben**	**Sonnleitner**
I Am Legend	0.810	0.721
Kill Bill Vol. 1	0.720	0.734
Saving Private Ryan	0.792	0.722
The Bourne Identity	0.785	0.805
Mean	0.777	0.746
[%]	**EER**	
All	25.92	31.18

Table 3.9: Comparison of VAD approaches using median smoothing on the classifier output.

the paper, than when using the partial film dataset for training. This is likely due to the fact that the dataset used in the paper is substantially larger, at approximately 35 hours, indicating that a sufficient quantity of data can overcome some of the challenges presented by out-of-domain training. Nevertheless, the MFCC-CC approach still clearly outperforms all implementations detailed in Eyben *et al.*'s paper, indicating that the feature can be used for robust discrimination of speech activity.

Table 3.10 provides a more detailed performance comparison of the MFCC-CC approach and Sonnleitner *et al.*'s approach (as this demonstrated the most competitive results in Table 3.8). The MFCC-CC approach demonstrates some reduced performance when compared to the initial testing results in Table 3.7, however, this was anticipated given the limited training set and larger test set. Despite this, the approach continues to exhibit competitive results, outperforming Sonnleitner *et al.*'s classifier across all performance metrics. In particular, it can be seen that while Sonnleitner *et al.*'s approach demonstrates relatively strong accuracy scores, significantly greater F-score values for our approach can be observed, indicating more robust performance.

Test set	Accuracy		Precision		Recall		F-score	
IAL	**0.88**	0.81	**0.62**	0.47	**0.81**	0.17	**0.70**	0.25
KB.1	**0.84**	0.79	**0.64**	0.62	**0.72**	0.26	**0.68**	0.37
SPRn	**0.87**	0.77	**0.91**	0.45	**0.66**	0.29	**0.77**	0.35
TBI	**0.94**	0.76	**0.88**	0.45	**0.88**	0.25	**0.87**	0.32
Mean	**0.88**	0.78	**76**	0.50	**0.77**	0.24	**0.75**	0.32

Table 3.10: Performance statistics of MFCC-CC approach and classifier from Sonnleitner *et al.* [119] when applied to whole-film dataset. Left (bold) MFCC-CC results. Right: results Sonnleitner *et al.*'s approach.

As the MFCC-CC approach uses 450 ms of temporal data, the investigations into Eyben and Sonnleitner's methods were reproduced using median smoothing over the classifier output of the same duration. The results of these investigations are demonstrated in Table 3.9. As shown when comparing Table 3.8 and Table 3.9, both approaches demonstrate improved AUC and EER when using the median filtering. This is likely due to de-noising effect of the median filter: by assigning classes according to the most prevalent classification within the window, incorrect classifications have less impact, thus resulting in greater classification accuracy. Crucially, despite the use of additional temporal information in this investigation, the MFCC-CC approach still demonstrates significantly better performance across all datasets.

3.7 Six Film Cross-Validation Investigation

Whole Film Dataset 2 was used to investigate the impact of training set size on classifier performance. While the typical approach to learning curve investigations involves iteratively adding a small proportion of the training set and re-training and testing the classifier, the approach used here adds the data from whole feature films at each iteration. The reason for this is that it is difficult to sub-sample fairly from feature film content due to the high degree of variance in the audio content. As such, were sub-sampling used for each film, it is unlikely that the sample set would be representative of the variety of audio content present within the data. Therefore, the learning curve has been constructed through adding entire feature films, in order to ensure representative sampling of content and therefore provide a more accurate impression of performance. The following results therefore present the mean statistics obtained from k-fold cross-validation over a range of 6 films, wherein a range of 1-5 are used for training, with the remaining film held out as the test set.

Figures 3.11 and 3.12 demonstrate that performance improves both overall and on individual test sets as the size of the training set is increased. That said, performance gain

is fairly marginal - with a peak mean accuracy of 0.86 only marginally above the minimum observed mean accuracy of 0.85. This suggests that a single feature film contains sufficient training examples to achieve good classification performance.

Figure 3.12 also provides further detail into classifier performance on a broader range of data, demonstrating competitive results across all six test sets. Of particular interest is the strong performance observed when testing on data from The Fellowship of the Ring and Saving Private Ryan, both of which contain a substantial amount of sound effects and modified voices. This is encouraging, indicating that the classifier is capable of negotiating some of the key challenges encountered when detecting speech in feature-film audio data.

A further notable observation is the difference in results from testing on The Bourne Identity and Disney's Hercules, which achieve the strongest and weakest results respectively. The contrast in classifier performance implies relatively significant differences in the data, indicating that The Bourne Identity contains largely typical speech content, while Hercules contains a greater proportion of atypical speech content. Empirical analysis reveals that Hercules contains a large amount of music that incorporates spoken word vocal styles. This differentiates it from the other films in the dataset and is a likely contributor to the reduced classifier performance observed on this test set. As such, these results indicate that music containing spoken-word speech content should be considered as an additional challenge when discriminating between music and dialogue.

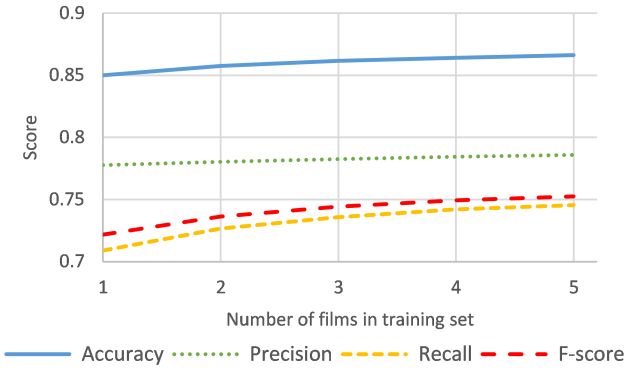


Figure 3.11: Mean MFCC CC classification results from six-film cross-validation over a range of training set sizes.

3.8 Non-English Speech Tests

While the VAD does not incorporate a language model, so far it has only been tested on largely English content. As such, to validate that it performs equivalently well on non-

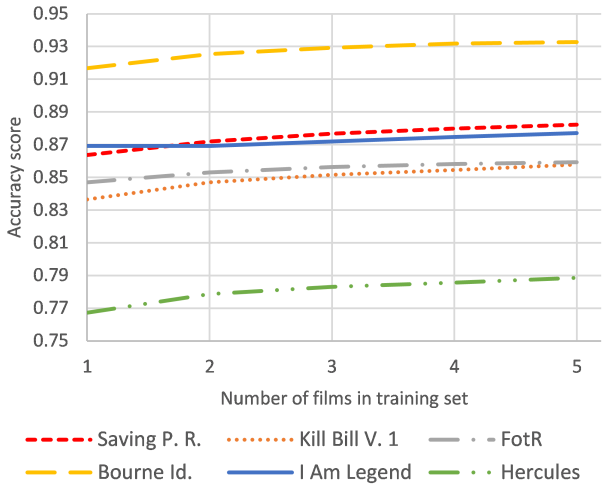


Figure 3.12: Accuracy of MFCC CC classifier from six-film cross-validation over a range of training set sizes.

Test set	**Accuracy**	**Precision**	**Recall**	**F-score**
The Girl with the Dragon Tattoo	0.921	0.906	0.887	0.897
Pan's Labyrinth	0.858	0.799	0.617	0.697

Table 3.11: Results from MFCC-CC VAD applied to non-English speech data.

English speech data, it has been tested on excerpts from two non-English films. These are The Girl with the Dragon Tattoo, which contains Swedish dialogue, and Pan's Labyrinth, which contains Spanish dialogue. The languages investigated were specifically chosen to cover both Germanic and Romance language groups: the two principal language groups in Europe. The test data comprised the first 30 minutes of each film, which was manually annotated to provide a ground truth for classifier evaluation. The classifier used is trained on all available data from the Whole Film Dataset.

As demonstrated in Table 3.11, the results for the two non-English datasets indicate similar VAD performance to the English data, with accuracies of 92% and 86% for The Girl with the Dragon Tattoo and Pan's Labyrinth respectively. These scores are both similar to the accuracy scores for the earlier investigations on English speech content, which achieved scores of $\approx 86 - 88\%$. The F-scores are also similar, with the lowest F-score achieved here being $\approx 70\%$ for Pan's Labyrinth, while The Girl with the Dragon Tattoo achieved an F-score of $\approx 89\%$. This resembles the range of F-score results from earlier investigations, such as demonstrated when comparing Table 3.11 to Table 3.10.

While this investigation only addresses a small subset of possible language groups, it clearly indicates that the VAD is capable of achieving strong results on languages which

are different to that used in the training set. This is encouraging as it demonstrates that the method is robust to the highly dynamic and variable audio used in entertainment media, and that it can be applied to languages which are not included in the training set. This makes it attractive for use in multimedia localisation and as the foundation of a language independent alignment solution.

3.9 Conclusion

This chapter has presented a novel approach for speech detection within film audio, and evaluated it with respect to a number of state of the art and contemporary approaches. The evaluations have shown that the approach is capable of strong performance over a range of feature film content, achieving accuracy scores of $> 90\%$ and improving on existing approaches across all performance metrics investigated. This approach is therefore deemed suitable for speech detection applications in feature film media. As such, the VAD will be used to obtain audio speech activity information which can be leveraged alongside transcript and visual speech information for use within an overall multimedia matching and synchronisation framework.

Chapter 4

Visual Speech Processing

4.1 Introduction

The previous chapter demonstrated that audio VAD is achievable via a number of methods, with the MFCC-CC method demonstrating encouraging performance with respect to other existing approaches. Despite the success of the proposed approach, there are still scenarios in which multimedia audio data is challenging for the audio VAD approach. Previous approaches have utilised visual speech information to help tackle some of the issues presented by complex or noisy audio signals in speech processing, such as work by Almajai *et al.* [2], Dov *et al.* [33] and Noda *et al* [94].

Given their prevalence in speech processing tasks, this work has explored the use of visual speech features for visual VAD (V-VAD). The principal here is that V-VAD can be used on content which is particularly challenging for audio VAD, such as content containing a significant degree of noise or sound effects. This is not a trivial task given the dynamic nature of entertainment media - with video content containing challenging lighting conditions, dynamic movement and frequent occlusions of key visual features.

Existing V-VAD approaches principally rely on appearance-based features, with several methods using 2D-DCTs as the primary feature [3] [75]. To address the challenging nature of visual speech in entertainment media, this work looks to exploit state-of-the-art methods for facial landmark localisation. These have been explored as they are more robust to noise induced by variable illuminations and occlusions (as discussed in Section 2.2.2), making them ideal for use with the challenging visual speech data in entertainment media. While the literature indicates that these landmark approaches may be better suited to entertainment media, appearance-based methods also provide useful information. As such, this work investigates both appearance and landmark-based features, and goes on to explore whether it is advantageous to combine both feature types.

4.2 Datasets

Two datasets have been used within this section: the Grid corpus, and a natural speech dataset comprising material obtained from video lectures. The Grid corpus is an audio-visual speech dataset assembled by the University of Sheffield [22]. The entire corpus contains 33 individuals each speaking 1000 sentences. For this work, four subsets of the Grid corpus have been used:

- **Grid Corpus Subset 1** The first subset is used for speaker independent testing. This comprises 1000 sentences from Grid subject 6. This subject was chosen as the same subject is used for speaker dependent investigations in Le Cornu *et al.*'s work [75], hence this allows for direct comparison with their approaches.
- **Grid Corpus Subset 2** This subset contains 10% of sentences from Grid subjects 1-7, 10 and 12. This is the subset of the corpus used in Le Cornu *et al.*'s paper [75], as such it has been used to compare directly with the VAD methods used in their paper.
- **Grid Corpus Subset 3** This subset uses the same subjects from Subset 1, however uses 100% of the available data. This has been used to validate that the 10% subset is representative of classifier performance on the whole dataset.
- **Grid Corpus Subset 4** The fourth Grid corpus subset is a gender balanced subset, comprising subjects 1-7, 15, 20 and 31. The latter three subjects were arbitrarily selected to create a gender balanced corpus of 5 male and 5 female subjects. This corpus is used to evaluate classifier performance on a gender balanced subset.

The natural speech dataset contains approximately 105 minutes of data from 7 different speakers. The video has been obtained (with permission) from video lectures from The University of Leeds, UK, and Duke University, North Carolina, USA. Whereas the Grid corpus subjects are restricted to stationary, frontal-face poses, this dataset has been assembled to provide visual speech information from unrestricted subjects. As such, the videos contain natural movements and head poses, as well as a variety of sub-optimal detection conditions including lighting variance, full and partial occlusions, and dynamic movement of both the camera and the subjects. The dataset contains approximately 10 minutes of speech and 5 minutes of non-speech for each speaker. This is as $\approx$5 minutes was the maximum amount of non-speech obtainable from most of the information sources, as many sources only contained $\approx$1 hour of data, during which pauses were rare. Furthermore, many of the sources change frequently from a view of the speaker to presentation slides, further

restricting the amount of usable data. Rather than using the natural speech/non-speech ratio, which in this case significantly favours speech, a ratio of 2:1 has again been used. This facilitates an accurate impression of classifier performance through ensuring there are a significant number of non-speech examples.

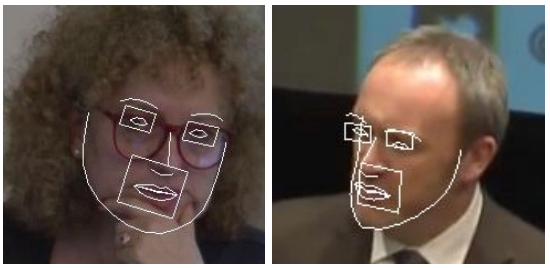


Figure 4.1: Examples from Natural Speech dataset: head poses, natural gestures and reflective glasses creating more challenging detection scenarios.

The data for the Grid Corpus datasets was segmented into speech and non-speech using the speech 'in' and 'out' times detailed in the data files provided with the Grid Corpus. The resulting data segments for each class were then concatenated, resulting in two pools of data: one for speech, and one for non-speech. For the Natural Speech Dataset, a similar approach for speech/non-speech segmentation used in Chapter 3 was applied: the data was first carefully segmented into speech and non-speech content, after which the resulting frames were concatenated to form two pools of data (speech and non-speech). In each case, the data was labelled according to its source file in order to ensure that, when the data was split into training and test data, none of the training data appeared in the test sets.

4.3 Feature Extraction and Selection

The first step in obtaining visual speech information is to extract relevant visual features. In the case of speech, these features should relate strongly to facial activity, and particularly to the mouth, as it is central to speech production. Several investigations have been undertaken to determine which computer vision techniques are most appropriate for feature extraction. Given work on audio/visual speech association in recent literature [2][75][94][73][123], two approaches have been considered: appearance-based features, via two dimensional DCTs (2D DCT), and landmark-based features, obtained using landmark localisation methods described by Kazemi *et al.* [66] and Saragih *et al.* [113].

4.3.1 Landmark Features

Two methods for landmark feature extraction have been used: a CLM-based approach and a pictorial structure model-based approach. These are introduced in Section 2.2.2. Each uses a simple face detector for initialisation, after which landmark localisation is applied to the ROI provided by the face detector. Each approach uses a different method for face detection, with the method from Kazemi *et al.* [66] using a HOG-based detector, and the method from Saragih *et al.* [113] using a Haar-like features detector. The implementation of Saragih *et al.*'s method used here was obtained via the FaceTracer C++ library [112], which facilitates an out-of-the-box implementation of the approach described in their paper. This is trained on the MultiPIE [43] dataset, thus providing 65 facial landmarks per face. The mouth-specific landmarks are numbered 48-64, thus the features used in this work are:

$$\mathbf{v}^{landmarks} = [\mathbf{l}_{48}, \mathbf{l}_{49}, \mathbf{l}_{50} ... \mathbf{l}_{64}] \tag{4.1}$$

where $\mathbf{l}_n$ comprises an (x,y) coordinate for the respective facial landmark location.

For the approach described in Kazemi *et al.*'s work [66], the DLib library [68] was used, which contains a C++ implementation of the approach. As the implementation comes with very few training examples, a training set was assembled using a variety of datasets labelled according to the i-BUG specification [108]. These datasets were:

- HELEN [74]
- iBUG [108]
- Annotated Facial landmarks in the Wild (AFW) [72]
- Labeled Face Parts in the Wild (LFPW) [10]

These datasets were chosen for training as they a) used 'faces in the wild' type data containing many examples of natural poses and illumination conditions, and b) were easily obtainable. The resulting training set consisted of 3837 annotated frames, comprising a variety of head poses and lighting conditions. The parameters used for training the approach were taken directly from the original paper [66] - with a tree depth of 5, and the learning rate, ν, set at 0.1. Oversampling has also been used to improve model training, with the oversampling amount set to 10. The model produces 68 facial landmarks, of which the mouth region landmarks are numbered 48-67. The resulting feature vector is therefore:

$$\mathbf{v}^{landmarks} = [\mathbf{l}_{48}, \mathbf{l}_{49}, \mathbf{l}_{50} ... \mathbf{l}_{67}] \tag{4.2}$$

4.3.2 Two Dimensional Discrete Cosine Transforms

The two dimensional DCTs (2D-DCT) coefficients are extracted from a rectangular region around the mouth. The region is obtained using a landmark-based approach to locate the mouth area. A rectangular region of interest is then defined around this region, with height and width set as $h_m = H/3$, and $w_m = W/2$, where H is the height of the detected face and h_m is the height of the region of interest around the mouth, and W is the width of the detected face, and w_m is the width of the region of interest around the mouth. A median filter is applied to the region of interest prior to extracting the 2D-DCT coefficients.

Previous literature has demonstrated good performance through the use of 14 2D-DCT mouth features [2]. As such, we employ the same approach for feature extraction here. The 2D-DCT matrix is obtained via:

$$t_{ij} = \begin{cases} \frac{1}{\sqrt{N}} & \text{if } i = 0 \\ \sqrt{\frac{2}{N}} cos \frac{\pi(2j+1)i}{2N} & \text{if } i > 0 \end{cases} \qquad 0 \leq i, j \leq N-1 \tag{4.3}$$

for each t_{ij} in transformation matrix **T**. The resulting 2D-transformed matrix is then given by:

$$\mathbf{C} = \mathbf{TZT}^T \tag{4.4}$$

where **Z** is an NxN-pixel image of the mouth region. The 2D-DCT produces a matrix which contains DCT coefficients ordered from high to low energy originating from the top left entry in the matrix - $(0, 0)$. The energy values decrease in zig-zag ordering, as demonstrated in figure 4.2.

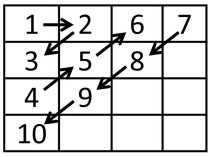


Figure 4.2: Example of energy-based ordering of DCT coefficients.

The resulting vector contains the first 14 DCT coefficients $c_1, c_2, c_3 ... c_{14}$, which forms the 2D-DCT vector for each video frame.

4.3.3 Feature Selection Via Audio-Visual Speech Correlation

The features have been evaluated according to their linear correlation with audio speech features. In this case, the audio feature used for correlation is MFCC0. This has been chosen following previous work [3], which showed that MFCC0 demonstrated the strongest correlation with visual speech features. A 10% subset of Grid Corpus Subset 4 was used to evaluate audio-visual speech correlation. The audio from the Grid corpus was originally sampled at 50 kHz, however has been down-sampled to 44.1 kHz. This has been done to conform with the sample rate used in the rest of the work.

Ordinary least squares has been used to model the data and obtain the multiple linear regression coefficients for each approach investigated, using MFCC0 as the dependant variable and the visual speech features as independent variables. Three facial feature approaches have been explored: one appearance-based feature, and two landmark-based features. The grey-level-based features are extracted using 2D-DCT, which has demonstrated strong performance in previous work [2][75], and the landmark-based features are extracted using the approaches from Kazemi *et al.* [66] and Saragih *et al.* [113], as described previously.

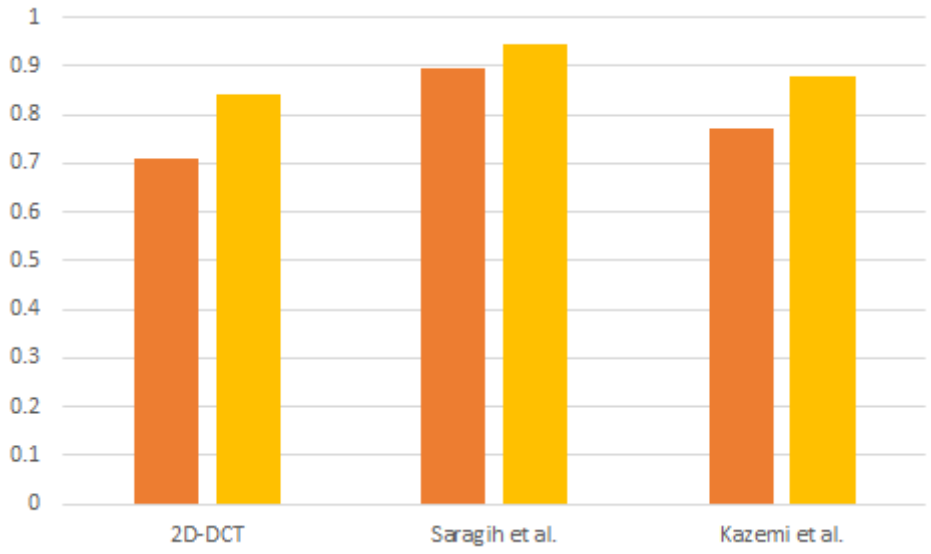


Figure 4.3: Multiple linear regression results for 2D-Discrete Cosine Transform, Saragih *et al.*'s landmarks [113] and Kazemi *et al.*'s [66] landmarks. Left bars (orange): R^2 terms. Right bars (yellow): correlation coefficients.

As demonstrated in Figure 4.3, the landmark-based features extracted from Saragih *et al.* and Kazemi *et al.*'s approaches correlate more strongly with audio speech features

when compared to the DCT features. This is potentially due to the DCT features incorporating information from the entire mouth region, whereas the landmark features facilitate extraction of lip-specific points while ignoring other, potentially noisy, information in the mouth region. However, the use of appearance-based features has clear advantages in a number of speech-oriented applications, such as automatic lip reading [49], where grey-scale information relating to visibility of teeth and mouth openness have proven to be crucial for accurate visual speech recognition. As such, this work will look to incorporate the DCT information alongside landmark features with the aim of providing a feature which incorporates the advantages of both approaches.

The results here also demonstrate that the approach from Saragih *et al.* achieves better correlation with audio speech features than the approach from Kazemi *et al.*'s work. A key contributing factor to the performance of Saragih *et al.*'s method is its ability to handle observations which do not adhere to the assumed model. This has been achieved by replacing the least-squares projection method with an M-estimator, as described in their paper [113]. While this modification was made to improve performance on data containing partial occlusions, it will also contribute to better model generalisation, and likely explains the enhanced correlation observed here, as visual speech features are extracted more consistently.

Given that the greatest audio-visual speech correlation was achieved using Saragih *et al.*'s approach, this has been chosen as the landmark localisation method for this work. The approach's strong performance on occluded data, as illustrated in their paper [113], was also influential when comparing the two methods used here.

4.4 Visual Voice Activity Detection

A number of methods for Visual Voice Activity Detection (V-VAD) have been proposed in the literature, such as Almajai *et al.*'s work [2], which uses V-VAD to enhance audio speech detection, and Le Cornu *et al.*'s work [75], which proposes a CNN-based approach for V-VAD, and demonstrates the strongest results, with an accuracy of $> 78\%$ on speaker independent data (the strongest results in the literature at the time of writing). While these results are encouraging, they have been obtained on a fairly ideal dataset - the Grid corpus - which consists of multiple speakers speaking a variety of brief, structurally similar sentences. Furthermore, all speakers are stationary and front-facing, and the dataset does not contain instances of occlusion, noise, significant lighting variability or dynamic movement. Given that the application scenario for this work is entertainment media, it would be beneficial to use a method which has been tested on, or indeed developed for,

more challenging speaker scenarios. This section discusses the development of a V-VAD approach designed to achieve strong performance for the task of speaker-independent visual speech classification for dynamic speaker applications.

4.4.1 Feature Extraction

As discussed in section 4.2.3, 2D-DCT and AAM demonstrated the strongest multiple correlation with audio speech features. Each approach has been used extensively in the literature, and each has a number of advantages and disadvantages. 2D-DCTs have proven to be useful in a variety of speech-based tasks, such as audio coefficient prediction [2], visual ASR [49][123] and V-VAD [75]. This demonstrates that appearance-based features can be used successfully within visual speech processing tasks, and supports findings in previous work as to the importance of grey-level features in this domain [17]. Landmark and shape-based approaches have also been successfully applied to visual speech processing problems, such as those described in work by Aubrey *et al.* [6] and Werda *et al.* [134], demonstrating that facial geometry can be exploited effectively for speech perception and processing tasks.

The advantages of appearance-based approaches are due to their incorporation of pixel-level features, making it possible to model mouth-specific information which can be leveraged within visual speech processing. This extends beyond more basic features such as mouth shape and openness, incorporating characteristics such as visibility of the teeth or tongue, which can be used to more accurately model visual speech features [49]. In turn, the disadvantage of appearance-based approaches is their reliance on a clear representation of the mouth region. As such, when this region is occluded, the data becomes highly noisy and potentially unusable.

Landmark-based features, on the other hand, are advantageous as they can compensate for occlusions or noise [66][113]. This makes them attractive for processing 'natural' visual speech (such as in entertainment media), where occlusions and dynamic movement may impede the performance of appearance-based approaches. The disadvantage to landmark-based approaches is that they rely entirely on shape-based information, making it difficult to accurately model visual speech characteristics due to the lack of appearance-level features.

Given these considerations, three feature representations have been explored: 2D-DCT, mouth landmarks and a combination of 2D-DCT and mouth landmarks. The hypothesis behind using the combined approach is that, where possible, the V-VAD will make use of 2D-DCT information, thereby facilitating a richer representation of visual speech characteristics. Where this is not possible (due to occlusion or other noise), the landmark

approach will still provide mouth shape information, thus reducing the impact of occlusion and noise on the V-VAD.

Visual Speech Feature Processing

In work by Vieriu *et al.* [130], an effective V-VAD is presented which processes grey-scale information from the mouth region to obtain statistical information which describes how visual speech characteristics change over time. Here, a similar method is proposed, this time using 2D-DCT, mouth landmarks and the combined 2D-DCT and mouth landmarks approach, rather than the features proposed in Vieriu *et al.*'s paper [130].

For the appearance-based features, 2D-DCTs are extracted in zig-zag order as described in Section 4.2.3, resulting in the feature vector:

$$\mathbf{v}^{dct} = [c_{(0,0)}, c_{(0,1)}, c_{(1,0)}, c_{(2,0)}, c_{(1,1)} \ldots] \tag{4.5}$$

The landmark feature vector simply comprises the mouth landmarks obtained from the tracker. As the tracker is trained on the CMU Multi-PIE dataset [43], the mouth landmarks are those numbered 48-64, hence the landmark feature is constructed as:

$$\mathbf{v}^{landmarks} = [\mathbf{l}_{48}, \mathbf{l}_{49}, \mathbf{l}_{50} \ldots \mathbf{l}_{64}] \tag{4.6}$$

where $\mathbf{l}_n$ comprises an (x,y) coordinate for the respective facial landmark location.

For the combined feature, the 2D-DCT and landmark features, $\mathbf{v}^{dct}$ and $\mathbf{v}^{landmarks}$ are combined for each frame, hence:

$$\mathbf{v}^{combined} = [\mathbf{v}^{dct}, \mathbf{v}^{landmarks}] \tag{4.7}$$

Given the importance of temporal information described in the literature [75, 121], the feature vectors are further processed to incorporate information from a range of frames (Figure 4.4). The approach for incorporating temporal information is based on Vieriu *et al.*'s work [130], and models the change over time by comparing all frames within a window to the first frame within the window (v_0). This is obtained using a sliding window of size w, for which the inter-frame difference, d, is obtained for the first frame, d_0, and each consecutive frame:

$$\mathbf{d}_i = \mathbf{v}_{i+1} - \mathbf{v}_0 \quad for \quad \mathbf{v}_{i:i+(w-1)} \quad in \quad \mathbf{v} \tag{4.8}$$

where $\mathbf{v}$ is the array of visual feature vectors and i is the index within the window. The feature is also appended with the mean, standard deviation, and first and second order

Figure 4.4: Illustration of frame window in which origin frame (v_0) is compared with subsequent frames to provide frame-difference feature.

temporal derivatives of **d**, where **d** is the final difference feature containing features $d_{0:w}$. Hence, the final feature vector is given as:

$$\mathbf{v}^{final} = [\mathbf{d}, \mathbf{d}^{\Delta}, \mathbf{d}^{\Delta\Delta}, \mathbf{d}^{\langle \mathbf{d} \rangle}, \mathbf{d}^{\sigma}] \tag{4.9}$$

In this work, a variety of window sizes are used to explore the impact of temporal information on V-VAD performance. This ranges from a window size of two to ten frames. For investigations using only two frames, the second order difference ($\mathbf{d}^{\Delta\Delta}$) is not used (as this cannot be computed for < 3 frames).

4.4.2 Experimental Design

Given the success of combining similar features with random forests in the literature [130], random forests have also been chosen as the classification method used here. Five V-VAD scenarios are explored here, using all Grid Corpus configurations described in Section 4.1, as well as the Natural Speech dataset.

Each test scenario uses a leave-one-out cross-validation approach, training on $n - 1$ samples and testing on the remaining sample. For each V-VAD scenario, the impact of temporal information and random forest estimators has been investigated. For the temporal information investigations, window sizes of 1, 3, 5, 7 and 10 have been used. This facilitates exploration of classifier performance from using only a single frame to using 400 ms of temporal information, as detailed in Table 4.1.

To investigate the impact of the number of estimators per forest, classifier performance

Window size (frames)	Temporal information at 25 fps (ms)
2	80
3	120
5	200
7	280
10	400

Table 4.1: Window sizes and equivalent durations used in V-VAD investigations.

over a range of estimators was explored on the speaker dependent and speaker independent Grid datasets. For the speaker dependent investigations, the number of estimators was increased incrementally, starting with 10 and 50 estimators, after which the number of estimators was increased by 50 for each subsequent test in the investigation, to a maximum of 450 estimators. As the V-VAD is being developed for speaker independent applications, speaker independent testing explored estimator impact at finer granularity - starting at 10 and 25 estimators, and increasing by 25 up to a maximum of 500 estimators - to obtain a more comprehensive impression of estimator impact on performance.

4.4.3 Speaker Dependent Results

Speaker dependent testing has been carried out using subject 6 from the Grid corpus. This subject was chosen as the same data is used for speaker dependent evaluation of the V-VAD approaches in Le Cornu *et al.*'s paper [75].

Estimator Tuning

Estimator results for a window size of 5 are presented here. This is as investigations in Section 4.3.3.2 section demonstrate that performance gain above 5 frames is marginal, and 5 frames achieves a good balance between resolution and performance. As demonstrated in Figures 4.5 and 4.6, all approaches demonstrate a gain in performance as the number of estimators is increased. This performance gain is more pronounced between 10 and 50 estimators, after which classifier performance begins to stabilise, with all approaches demonstrating less variability in performance above 250 estimators. Crucially, the results here show a clear advantage to using DCT over the landmark-based approach, with an approximate increase in accuracy of 2% across all estimator configurations tested. This improvement in performance supports the notion of appearance-level features providing a richer representation of visual speech characteristics, i.e. due to the visibility of tongue or teeth, as discussed in Hassanat *et al.*'s work [49]. The landmark-based features still achieve

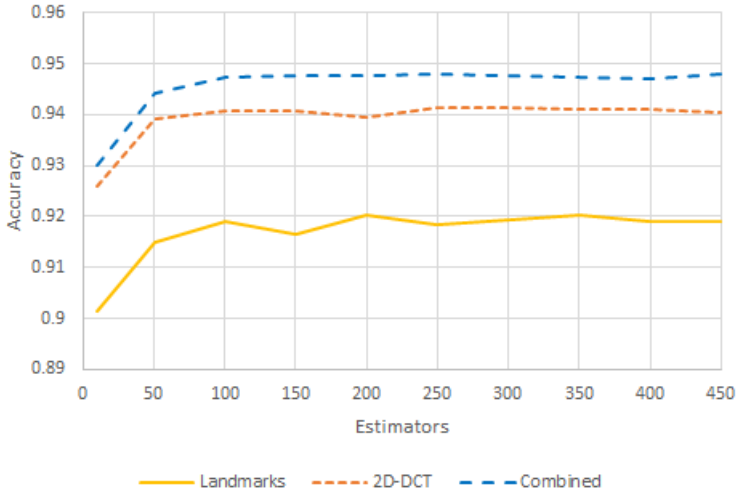


Figure 4.5: Accuracy results from visual speech feature comparison on speaker dependent dataset. Testing over a range of estimators with a window size of 5.

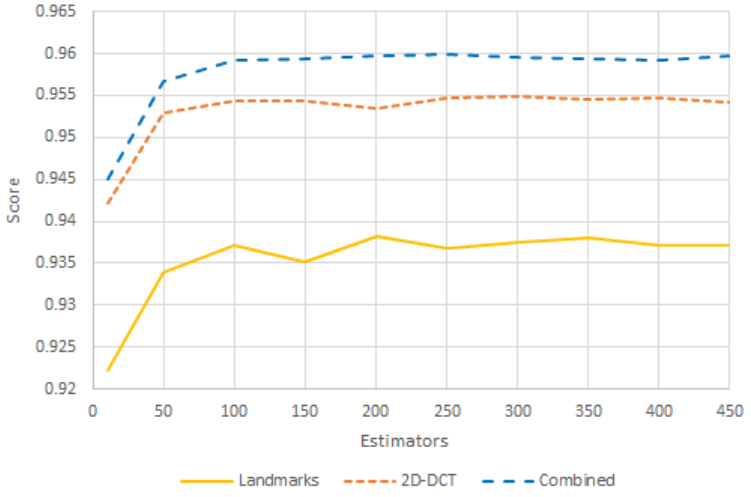


Figure 4.6: F-score results from visual speech feature comparison on speaker dependent dataset. Testing over a range of estimators with a window size of 5.

relatively strong results, with a peak accuracy of 92%, achieving performance statistics close to those observed in the literature [75].

Another key observation here is the performance gain achieved through the use of the combined approach. A marginal yet clear gain in performance can be observed, with the combined features attaining an accuracy approximately 0.6-0.8% greater than the 2D-DCT features, as well as a marginal improvement in F-score. This is encouraging, suggesting that performance gain can be achieved through combining shape/landmark and

Window size (frames)	Accuracy	Precision	Recall	F-score	ROC AUC
2	0.880	0.889	0.927	0.907	0.935
3	0.930	0.922	0.972	0.947	0.966
5	0.948	0.939	0.982	0.960	0.974
7	0.958	0.947	0.988	0.967	0.976
10	0.962	0.951	0.990	0.970	0.978
15	0.965	0.952	0.996	0.973	0.980

Table 4.2: Speaker dependent V-VAD results for combined feature classifier trained on 250 estimators.

appearance-based features, supporting the hypothesis behind this approach.

As negligible performance gain is achieved after 250 estimators, this configuration has been used in the following investigations.

Window Size Investigation

Further investigations into the performance of the combined classifier have been undertaken to explore the impact of temporal information. As Figures 4.7 and 4.8 demonstrate, both accuracy and F-score results improve with the inclusion of more temporal data.

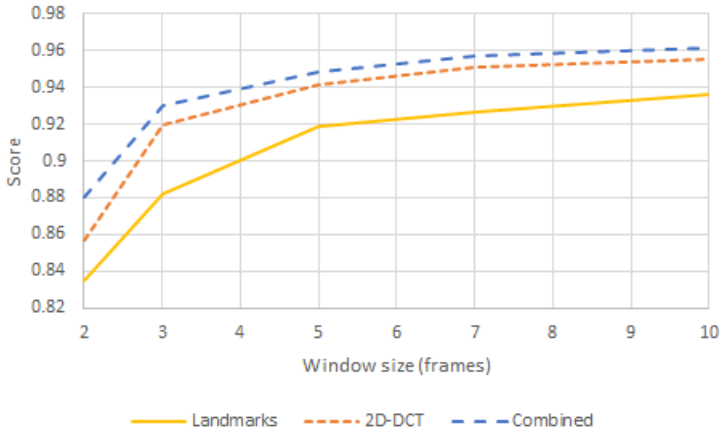


Figure 4.7: Accuracy results from visual speech feature comparison on speaker dependent dataset. Testing over a range of window sizes using 250 estimators.

Table 4.2 explores the impact of temporal information on classifier performance in more detail. The results demonstrate that all performance statistics improve as the number of frames is increased. Interestingly, the classifier performs relatively well given very little temporal information, with an accuracy of 0.88 when using information from single frames. The greatest improvement in performance is observed when moving from two frames to a

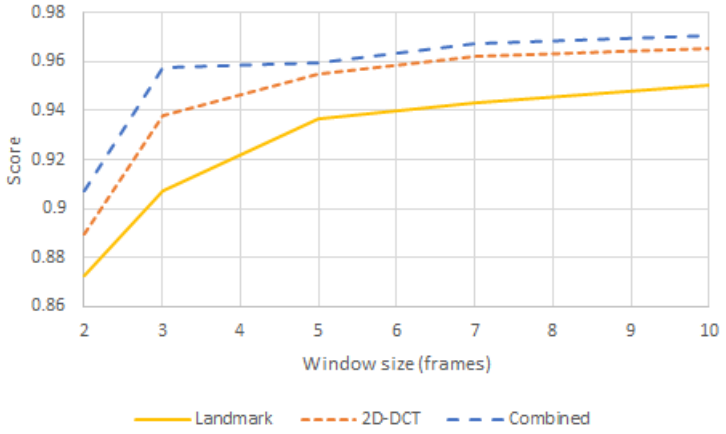


Figure 4.8: F-score results from visual speech feature comparison on speaker dependent dataset. Testing over a range of window sizes using 250 estimators.

Approach	**Accuracy**
Landmarks	0.919
DCT	0.942
Combined	0.948
GMM DCT*	0.926
GMM DCT Δ*	0.943
NN DCT*	0.960
NN DCT Δ*	0.968
CNN Static*	0.970
CNN Stack 3*	**0.977**

Table 4.3: Comparison of V-VAD results for speaker dependent tests using Grid subject s6. Results from Le Cornu *et al.*'s paper [75] indicated by *.

window size of five frames - producing an increase in accuracy of approximately 7%.

While the strongest results are observed with a window size of 15 frames, the performance gain is marginal when compared to using 10 frames (Table 4.2). As such, future investigations will explore the range of 2-10 frames, as the marginal performance increase does not justify the loss in resolution.

The speaker dependent VAD was evaluated with respect to leading contemporary approaches described in Section 2.2.3.2. A window size of 5 frames was used for the DCT, landmarks and combined approaches. As Table 4.3 demonstrates, the combined landmark and DCT approach comes closest to achieving the accuracy metrics from Le Cornu *et al.*'s work, with an accuracy between 1.3% and 3% below the NN and CNN results from their paper [75]. This demonstrates that, in a speaker dependent scenario, the NN and CNN approaches are able to more effectively model visual speech characteristics. As the

underlying aim of using the combined approach is to improve model generalisation through the incorporation of landmark-based features, it is possible that while the approach does not perform as well in the speaker dependent case, it may generalise better - thus improving upon the performance of these approaches in the speaker independent scenario.

4.4.4 Speaker Independent Results

Initial investigations explored classifier performance over a number of estimators and window sizes using Grid Corpus Subset 2. Figure 4.9 shows clear performance gain when increasing from 10 to 50 estimators, and between window sizes 2 and 5. The figure also demonstrates that the same pattern is reflected in the 2D-DCT and landmark-based approaches, however the landmark approach demonstrates improved performance over the 2D-DCT (contrary to the speaker dependent results). The following sections explore the impact of number of estimators and window size in more detail in order to determine suitable algorithm parameters. These investigations into classifier performance have been carried out on a number of speaker independent dataset configurations.

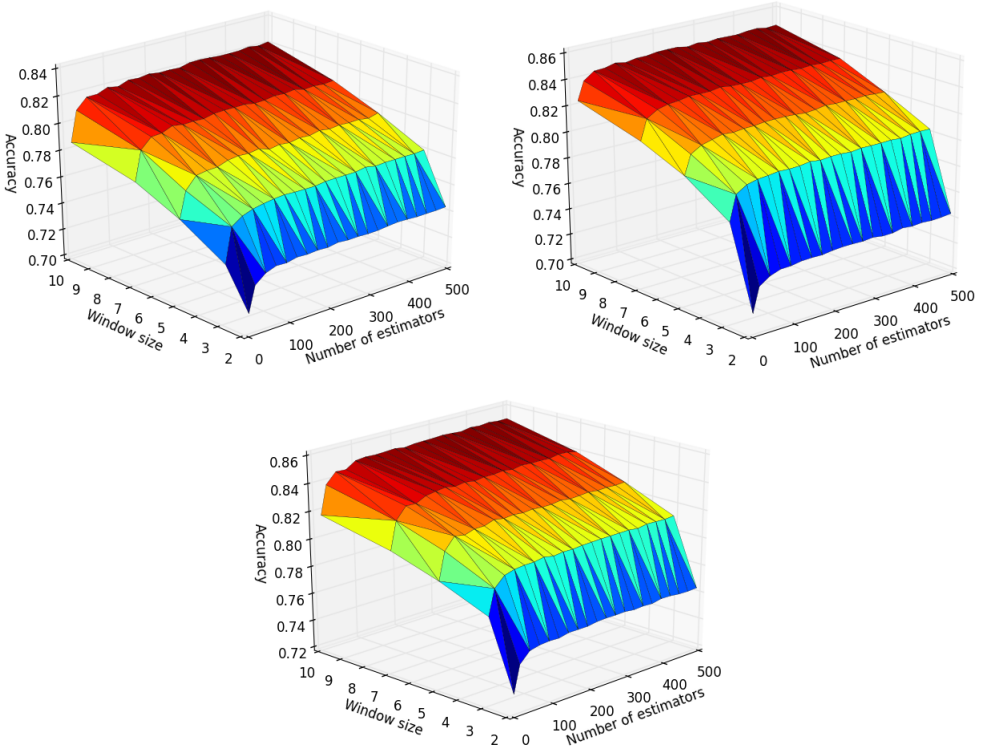


Figure 4.9: Speaker independent accuracy results using Grid dataset configuration from Le Cornu *et al.*'s paper [75]. Testing over a range of estimators and window sizes. Left: 2D-DCT, right: landmarks, Bottom: combined features.

Estimator Tuning

A window size of 5 has again been used for speaker independent estimator tuning. As Figure 4.10 demonstrates, the performance curves follow a similar trend to the speaker dependent results, with a significant initial increase in performance which then stabilises. Interestingly, the results here vary from the speaker dependent investigations, with the landmark-based approach achieving greater accuracy than the 2D-DCT approach. This indicates that landmark-based features may be particularly advantageous in speaker-independent scenarios, potentially due to the features' invariance to appearance factors (e.g. variable lighting, textures or facial features), which allows them to more effectively model speaker independent speech characteristics.

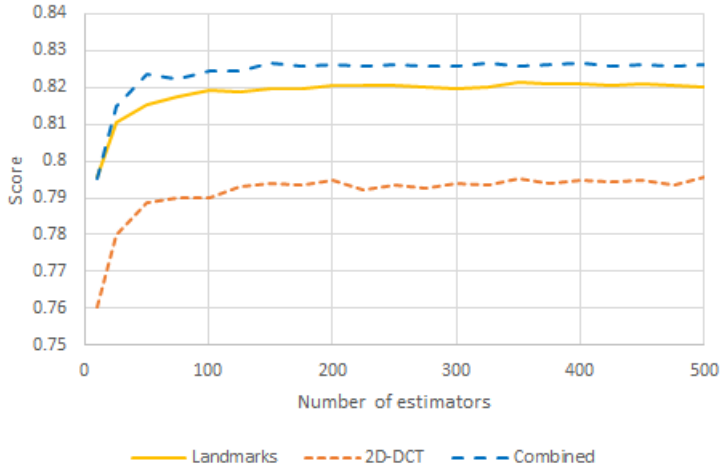


Figure 4.10: Speaker independent accuracy results using Grid dataset configuration from Le Cornu *et al.*'s paper [75]. Testing over a range of estimators with a window size of 5.

A similar trend can be observed in the F-score values, as demonstrated in Figure 4.11. This further supports the notion of the landmark-based approach being more robust for speaker independent scenarios due to minimising the impact of appearance-based phenomena.

Crucially, the combined approach has continued to demonstrate the strongest performance, with the accuracy and F-score values consistently exceeding the landmark-based approach by $\approx 0.5\%$ and $\approx 1\%$ respectively.

Window Size Investigation

The results in Figures 4.10 and 4.11 and the speaker dependent investigations demonstrate that good performance can be achieved using 250 estimators. As such, this number of

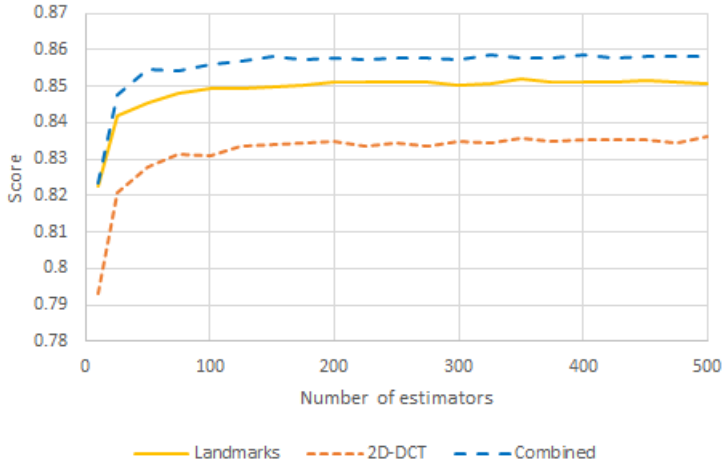


Figure 4.11: Speaker independent F-score results using Grid dataset configuration from Le Cornu *et al.*'s paper [75]. Testing over a range of estimators with a window size of 5.

Window size (frames)	Accuracy	Precision	Recall	F-score	ROC AUC
2	0.764	0.761	0.862	0.809	0.822
3	0.811	0.794	0.906	0.846	0.881
5	0.826	0.808	0.916	0.858	0.904
7	0.840	0.819	0.927	0.869	0.917
10	0.850	0.826	0.939	0.878	0.930

Table 4.4: Speaker independent V-VAD results for combined feature classifier trained on 250 estimators.

estimators has been chosen for the window size investigations.

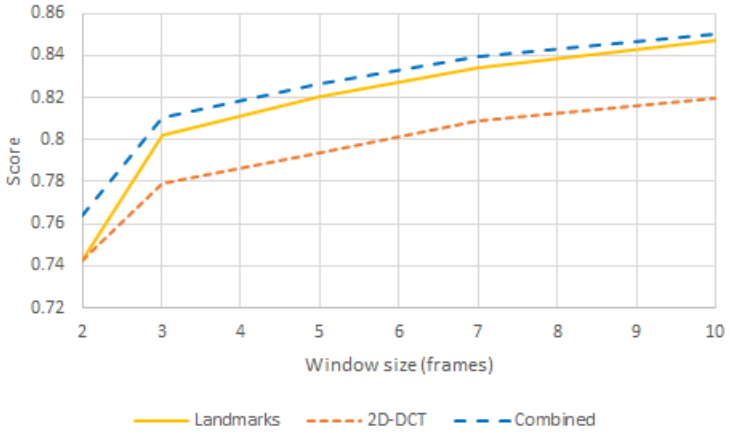


Figure 4.12: Speaker independent accuracy results using Grid dataset configuration from Le Cornu *et al.*'s paper [75]. Testing over a range of window sizes using 250 estimators.

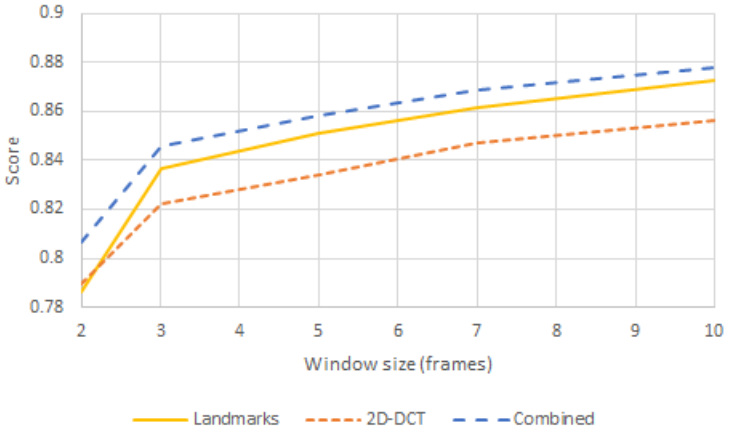


Figure 4.13: Speaker independent F-score results using Grid dataset configuration from Le Cornu *et al.*'s paper [75]. Testing over a range of window sizes using 250 estimators.

The use of greater window sizes continues to produce enhanced accuracy and F-score metrics with the speaker independent approach, as exhibited in Figures 4.12 and 4.13. Further insight into the impact of temporal information is provided in Table 4.4. Again, the trend from the previous section (Table 4.2) is reflected here, with the greatest increment in performance metrics observed in the step between one and three frames. All performance metrics demonstrate improvement as the number of frames is increased, however, performance gain between 7 and 10 frames is fairly marginal. As such, a frame size of 7 (280 ms) or 5 (200 ms) may be preferable - sacrificing marginal performance gain for a reasonably significant increase in resolution from 2.5 Hz to 5 Hz.

Table 4.5 compares speaker independent results of the three classification approaches with those from Le Cornu *et al.*'s paper [75]. While the NN DCT Δ approach demonstrates reasonable performance, all proposed approaches except for the 3 frame DCT approach, outperform the methods detailed in Le Cornu *et al.*'s work. Most significantly, this demonstrates that the combined approach is more effective on speaker independent data than the NN and CNN results proposed by Le Cornu *et al.*, further supporting the original hypothesis. This indicates that while the CNN and NN approaches achieve strong results on speaker dependent data, they do not generalise as well as the proposed approaches when considering speaker independent applications. This is likely due to a combination of factors:

- Previous work has demonstrated strong performance of random forests for both audio [119] and visual [130] VAD tasks. Thus, it is possible that the structure of the learning approach, i.e. a decision tree ensemble, is more effective at modelling speech characteristics for binary speech classificaton problems.

Approach	**Accuracy**
Landmarks (3 frames)	0.802
Landmarks (5 frames)	0.821
DCT (3 frames)	0.779
DCT (5 frames)	0.793
Combined (3 frames)	0.811
Combined (5 frames)	**0.826**
GMM DCT Δ*	0.705
NN DCT Δ*	0.787
CNN Static*	0.741
CNN Stack 3*	0.747

Table 4.5: Comparison of V-VAD results for speaker independent tests using 9 speaker Grid corpus configuration from Le Cornu *et al.*'s paper [75]. Results from Le Cornu *et al.* indicated by *.

- All approaches in Le Cornu *et al.*'s paper [75] use appearance-based features, and thus are not robust to more variable appearance features, such as would be present within a speaker independent dataset. As previously mentioned, the inclusion of landmark-based features likely helps to improve performance as the features are less affected by grey-scale variation.

The efficacy of landmark-based features for modelling speaker-independent data is further substantiated by the performance of the landmark-based approach, which significantly exceeds the performance of the GMM and both CNN-based approaches documented in Le Cornu *et al.*'s paper [75].

As Le Cornu *et al.*'s evaluation uses only 10% of the available data for the 9 selected subjects [75], further investigations have been carried out using 100% of data from these subjects. This has been done both to determine the impact that the amount of training data has on classifier performance, as well as to validate that the 10% subset of the data gives an accurate impression of classifier performance. These tests explore the performance of the landmarks, DCT and combined classifiers trained on 250 estimators over a range of window sizes.

Through comparing tables 4.5 and 4.6, marginal improvement can be observed from using 100% of the data from the 9 speaker subset, however the overall trend remains nearly identical. This indicates that the incorporation of more data has some impact on classifier performance, but suggests that the 10% subset used in earlier investigations gives a good general impression of performance across all metrics explored.

Investigations exploring the three feature sets also concur with earlier findings, with the landmark and combined approaches achieving greater performance when compared with

Window size (frames)	Accuracy	Precision	Recall	F-score	ROC AUC
2	0.779	0.784	0.870	0.824	0.840
3	0.830	0.822	0.915	0.865	0.901
5	0.848	0.837	0.927	0.879	0.923
7	0.859	0.846	0.936	0.888	0.935
10	0.870	0.854	0.947	0.897	0.947

Table 4.6: Speaker independent V-VAD cross-validation results for combined feature classifier trained on 250 estimators. Using 100% of Grid corpus data from users selected in Le Cornu *et al.*'s paper [75].

the DCT features. Interestingly, Figure 4.14 shows that, from a window size of 3 onwards, the accuracy of the landmark-based approach exceeds that of the combined approach, albeit marginally. A similar trend can be observed in Figure 4.15, with the F-score of the landmark approach slightly exceeding the metric of the combined method when using window sizes of 7 and 10. This indicates that the landmark features perform better with more contextual information. A possible explanation for this is that the landmark features are more robust to noise. Thus, as the number of frames per feature increases, the likelihood of noisy frames - and thus inaccurate DCT features - also increases. At this point, the DCT has a slight negative impact on the combined feature vector, whereas the landmark features alone are more robust to noisy grey-scale data.

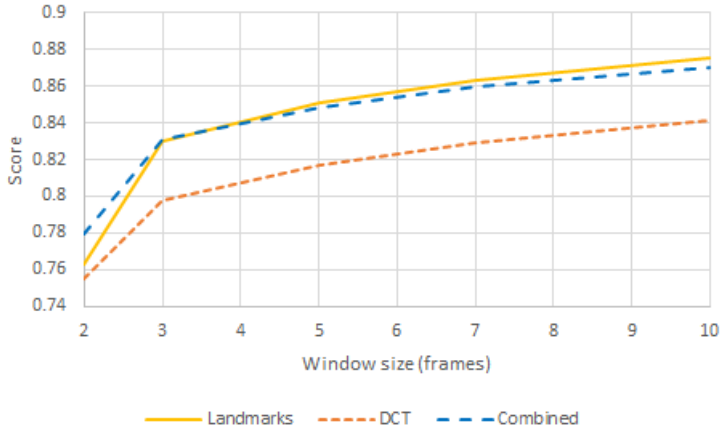


Figure 4.14: Speaker independent accuracy results using 100% of Grid corpus data from users selected in Le Cornu *et al.*'s paper [75]. Testing over a range of window sizes using 250 estimators.

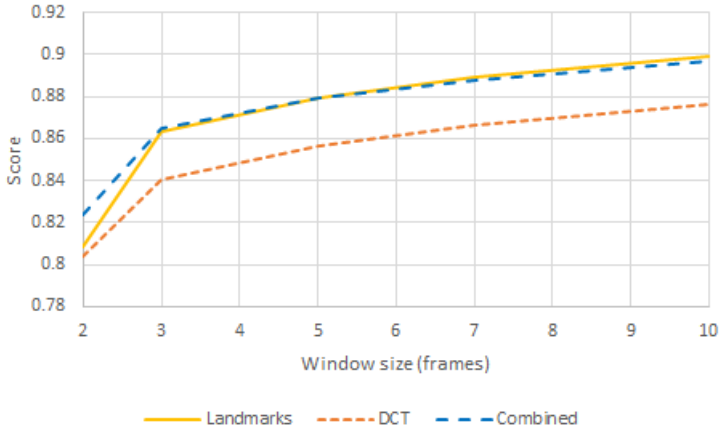


Figure 4.15: Speaker independent F-score results using 100% of Grid corpus data from users selected in Le Cornu *et al.*'s paper [75]. Testing over a range of window sizes using 250 estimators.

Gender Balanced Dataset

As the dataset used in Le Cornu *et al.*'s paper [75] is not gender balanced, the VAD investigations were repeated on the gender balanced subset of the Grid corpus to determine whether subject gender has any notable impact on classifier performance.

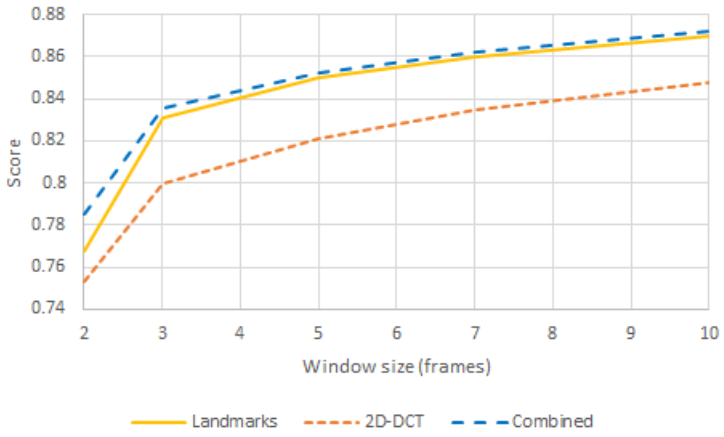


Figure 4.16: Speaker independent accuracy results using 100% of Grid corpus data from gender balanced subset. Testing over a range of window sizes using 250 estimators.

As the F-score and accuracy results in Figures 4.16 and 4.17 demonstrate, the landmark and combined approaches achieve the strongest results - further supporting the findings from earlier investigations. While the results from Grid subset 3 demonstrate a marginal improvement with frame sizes 7 and 10 using the landmark features alone, these results

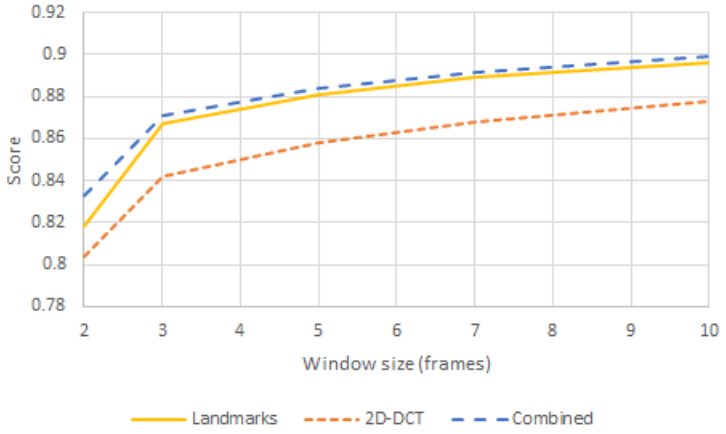


Figure 4.17: Speaker independent F-score results using 100% of Grid corpus data from gender balanced subset. Testing over a range of window sizes using 250 estimators.

corroborate more strongly with those from subsets 1 and 2, suggesting that the combined features perform better overall.

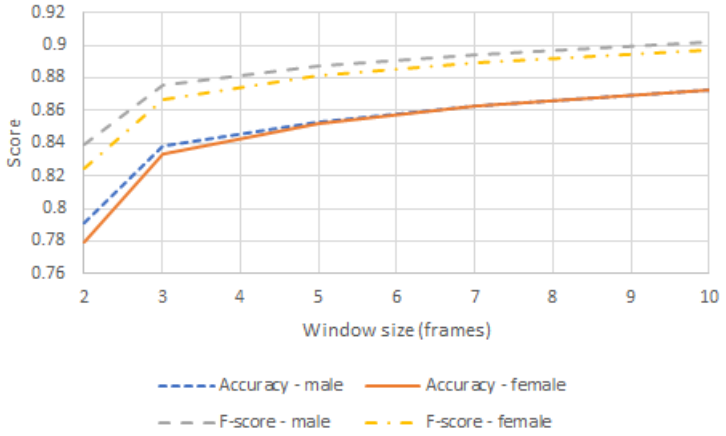


Figure 4.18: Mean of speaker independent accuracy and F-score results for male and female subsets of Grid Corpus Subset 4. Testing over a range of window sizes using 250 estimators.

Figure 4.18 shows the mean results for the male and female subsets of Grid Corpus Subset 4 over a range of window sizes. These results suggest that gender has minimal impact on classifier performance, with only subtly variable F-score results and accuracy metrics converging strongly at window sizes of 5+ frames. While the variance may simply be due to general inter-speaker variance (rather than variance influenced by gender), these results indicate that gender has little impact on classifier performance. Most importantly,

the investigations on the gender-balanced dataset achieve similar results to the unbalanced dataset, thus validating that the unbalanced dataset gives an accurate impression of classifier performance.

4.4.5 Natural Speech Dataset Results

This section explores the performance of the V-VAD on the Natural Speech dataset in order to evaluate its performance on more variable data incorporating more realistic speaker behaviour. As Figure 4.19 demonstrates, classifier performance begins to level off when between 10% and 20% of the training data is used, equating to approximately 100 minutes of data. As such, this has been used as a guideline for the amount of data necessary to provide a good impression of classifier performance, and ≈100 minutes of data has been put together for the Natural Speech dataset (approximately 15 minutes of data each for 7 speakers).

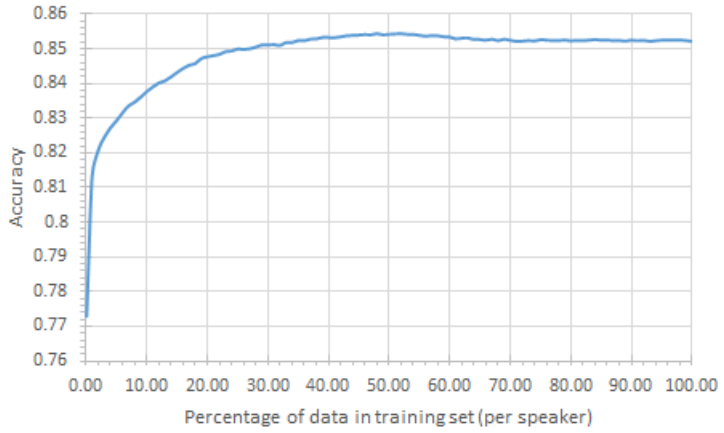


Figure 4.19: Learning curve of Grid Corpus Subset 4 using combined approach with window size of 5 and 250 estimators.

Figures 4.20 and 4.21 demonstrate that the impact of landmark features is particularly pronounced, with the landmark features alone achieving a notable improvement in performance over the DCT features commonly used in other work [2][75][91]. Furthermore, the performance of the combined features exceeds that of both the 2D-DCT features and landmark features over all window sizes tested. This indicates that the combined feature set provides a more robust representation of visual speech information for V-VAD in noisy, sub-optimal conditions. In turn, this supports the hypothesis that landmark features are advantageous for challenging visual speech detection tasks, and that performance can be enhanced by combining both appearance and landmark-based features.

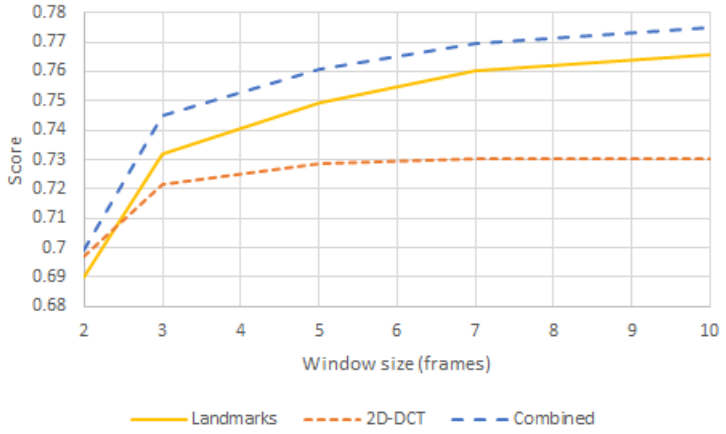


Figure 4.20: Mean accuracy results from V-VAD applied to Natural Speech dataset using 250 estimators over a range of window sizes using cross-validation.

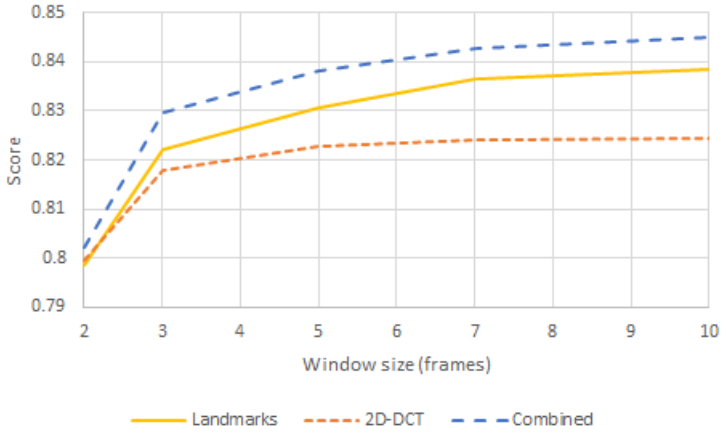


Figure 4.21: Mean F-score results from V-VAD applied to Natural Speech dataset using 250 estimators over a range of window sizes using cross-validation.

4.5 Conclusion

This chapter has presented a novel method for V-VAD which incorporates a combination of appearance-based and landmark-based features. The V-VAD has been tested on a variety of datasets, and has demonstrated improved performance when compared with other contemporary V-VAD approaches. Crucially, this work has shown that feature vectors which incorporate facial landmarks outperform commonly used appearance-based features in speaker-independent visual voice activity detection tasks. This performance advantage has also been demonstrated in more difficult speech detection tasks involving variable lighting and natural speaker behaviours, and the results indicate that landmark features are

Test set:	1	2	3	4	5	6	7
Accuracy	0.720	0.844	0.768	0.784	0.745	0.752	0.714
Precision	0.687	0.872	0.841	0.784	0.746	0.751	0.766
Recall	0.943	0.939	0.85	0.942	0.937	0.949	0.829
F-score	0.795	0.904	0.846	0.856	0.831	0.839	0.796

Table 4.7: Classifier results from Natural Speech dataset cross-validation using 250 estimators with a window size of 5 frames.

particularly valuable in achieving improved classification accuracy under these challenging conditions.

While the approach discussed here has achieved encouraging results, future work will look into using the combined features with more sophisticated machine learning algorithms, such as convolutional neural networks, as these have proven to be successful in the literature [75]. Furthermore, as more training data enhances classifier performance (as demonstrated in sections 4.3.4 and 4.3.5), it would be beneficial to continue to expand the dataset to facilitate training of a more accurate model.

Chapter 5

Language Independent Feature Matching and Alignment

5.1 Introduction

A key aim of this work is to develop a language-independent method for associating text (either transcript or subtitle files) with speech information. This has value in facilitating the following:

- A method for re-aligning audio to video of a different frame rate (e.g. aligning PAL to NTSC).
- A low-resource method for finding associations between text and speech content when no language model is available.
- A language-independent method for associating text and speech prior to linguistic processing for lexical discovery.
- A language-independent method for evaluating the accuracy of human-defined speech-to-text, e.g. for subtitles.

This chapter proposes language independent methods for feature matching and alignment which leverage information from speech detections to find corresponding patterns in textual media. While output from the speech detector does not provide the high level information used in language-dependent approaches, such as lexical content, it does provide useful information regarding speech within the audio signal in the form of speech in and out predictions. This section explores the process of feature matching and subsequent alignment using the predicted speech output and text resources such as subtitles and transcripts.

The chapter begins with an outline of the data representation methods used for this work, after which an approach for anchor point detection, signal segmentation and audio to text association is introduced. This is then applied in the context of multimedia alignment whereby a query signal, obtained from the text data, is mapped to a reference signal, obtained from the audio data. The chapter explores the application of the feature matching and alignment framework for both whole-film and scene-level data, and discusses the challenges and proposed solutions for each case. Section 5.5 also explores the feasibility of incorporating visual speech information, and goes on to demonstrate the challenges presented by feature film content in this regard. Lastly, the chapter explores an improved alignment method which utilises both whole-film and scene-level data to improve scale coefficient estimation, before concluding with a summary of the feature matching and alignment investigations.

The multimedia alignment method used throughout this chapter comprises three key steps: anchor point estimation, segment matching and scale coefficient estimation. Each step is evaluated individually to provide a detailed analysis of the alignment and association framework. First, evaluation of the anchor point estimation step is used to determine how close the estimated anchor points in the query signal are to their equivalent points in the reference signal, thus giving an impression of the anchor point estimation's accuracy. The evaluation of the segment matching step investigates the method's performance at general association by analysing the degree of overlap between segments matched by the algorithm. Lastly, the scale coefficient estimation task looks to generate a scaling coefficient to linearly adjust the time base of the query signal to match that of the reference signal. This is achieved by processing and filtering information from the anchor point and segment association steps. Crucially, evaluation of the scale coefficient estimation gives an impression of whether VAD and text data can be applied within an automatic multimedia alignment solution.

5.2 Data Representation

The speech detector provides speech activity data in the form of a binary signal, with 1 being speech and 0 being non-speech. As such, a similar signal has been generated from the text information. For the subtitle data, this is done using the speech 'in' and 'out' timestamps. For the transcript data, as out times are not provided, simulated durations have been used. This is done by counting the number of words following the onset in a given segment of speech, and dividing this number by a words-per-second estimate to provide the duration of speech activity in seconds. For English, this estimate is 2.5 words per second

[122]. An array of the same resolution as the VAD output is then created, and the speech in and out data is used to define elements of the array as either 1 (speech active) or 0 (speech inactive). This is achieved by converting the speech in and out times to array indices which match the resolution of the VAD output, after which all array items which lay between in and out times are defined as 1, while those outside of active subtitle regions (e.g. between 'out' and 'in') are defined as 0.

While this represents the text information in a manner similar to the VAD output, this approach has several problems. The first of these concerns transcript information: as the out time is only a rough estimate, the simulated signal varies with respect to the actual durations of speech segments. Secondly, while the subtitles provide in and out times, these timestamps represent the on/off times for the subtitles to be displayed on screen. These timings rarely match the actual speech content, as they are: a) re-worded to optimise their space requirements on screen, and b) displayed for longer than the corresponding speech, to allow viewers sufficient time to read the content. The final issue with this representation is due to noise from the speech detector. While this does demonstrate good performance on feature film data, the speech detections do not align perfectly with the ground truth, and the output can be noisy, containing both false positives and false negatives as demonstrated in Figure 5.1.

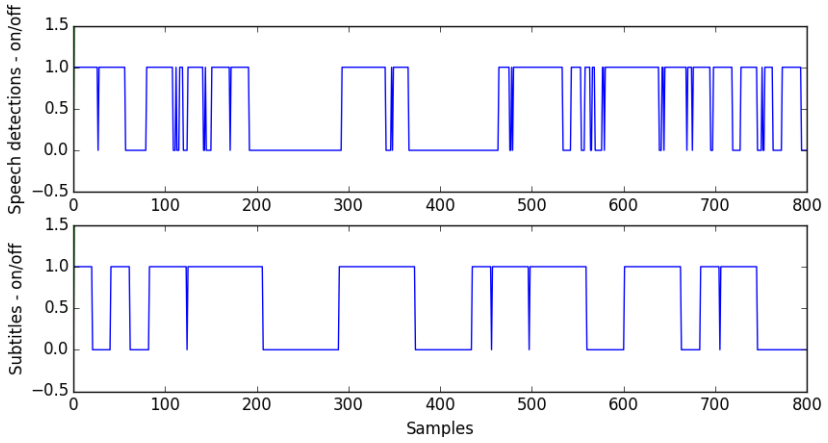


Figure 5.1: Plot of speech detections and subtitle in/out data represented as binary pulse signal.

Given these factors, sequence alignment techniques such as DTW cannot always be confidently used, as they will tend to misalign segments due to the variation between the binary speech and text signals. This problem is further compounded by the lack of shape in the binary signals. This results in many candidates which appear to match optimally

according to the DTW cost function, despite the fact that the features are not associated. This is illustrated in Figure 5.2, in which DTW is used to align the predicted speech (query) with the corresponding subtitles (reference). As there is no shift in time, these should align almost perfectly, producing a diagonal path through the cost matrix. Deviations from the diagonal correspond to points at which the signal is warped to fit an optimal alignment. Given that the signals correspond exactly to one-another, these deviations represent misalignments resulting from inconsistencies between the subtitle and speech detection signals.

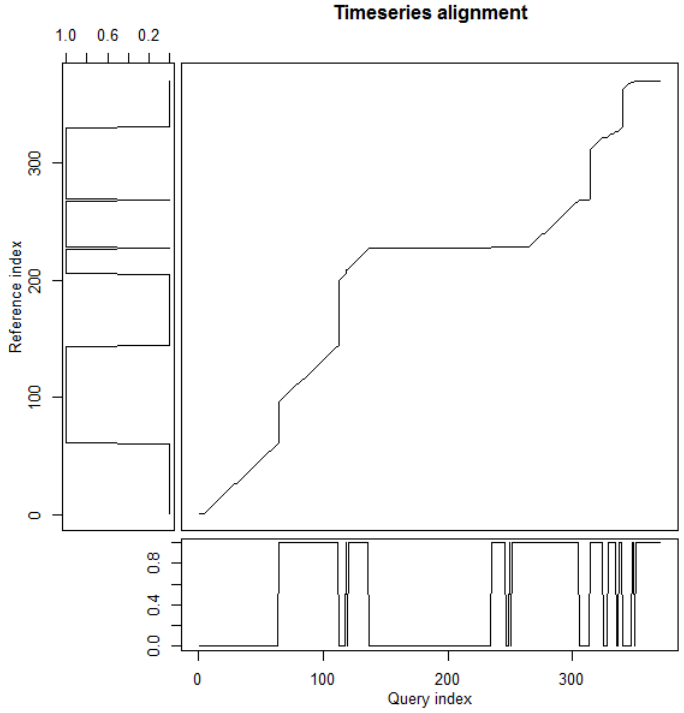


Figure 5.2: Example misalignment of two corresponding speech detection (query) and subtitle (reference) signals.

To address this, a sliding window is applied to the binary signals to obtain the sum of segments in signal $\mathbf{x}$ centred at indices i over windows of size w:

$$sum_i = \sum \mathbf{x}_{i-\frac{w}{2}:i+\frac{w}{2}} \tag{5.1}$$

This improves the performance of signal alignment by smoothing the detection noise and creating a signal with distinct shape characteristics. This reduces ambiguity when evaluating features between the two signals, improving the performance of the DTW feature matches and thus the subsequent alignment. This is illustrated in Figure 5.3, in which the

path through the cost matrix is improved, exhibiting less deviation from the diagonal.

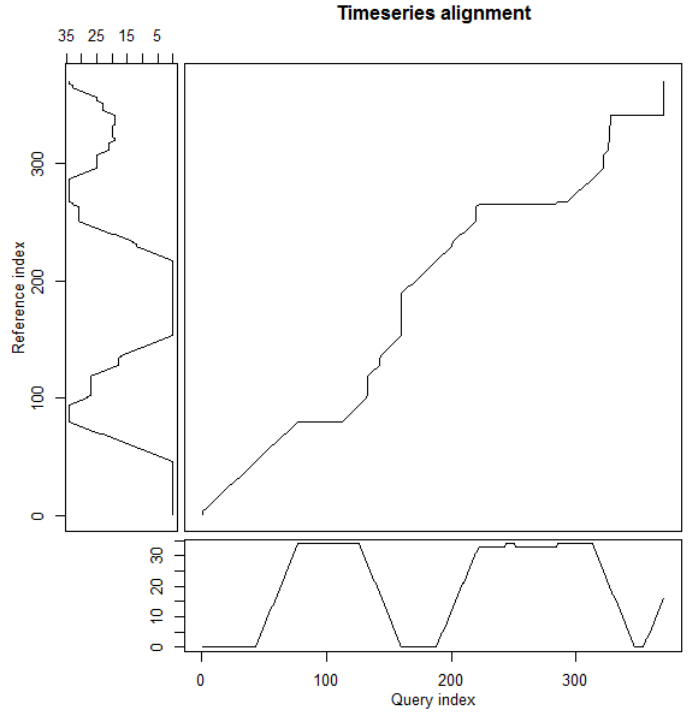


Figure 5.3: Example of DTW alignment on summed speech detection (query) and subtitle (reference) signals.

In this work, an implementation of the MFCC-CC VAD approach was used as the method for audio speech detection. The VAD was trained using the data from Whole Film Dataset 2 (As described in Section 3.2). For visual speech detection, the combined feature V-VAD approach from Chapter 4 was used. This was trained using the data from the Natural Speech Dataset (as described in Section 4.1). The Dynamic Time Warping (DTW) implementation used is from the R package developed by Toni Giorgino [39].

5.3 Anchor Point Detection and Signal Segmentation

One of the key goals of this work is to find associations between segments of audio and text. To do so, it is first necessary to identify segments which represent significant features within the audio and text signals. This is achieved by finding key anchor points within the data - points which can be used to define breaks between significant sections of dialogue. This has the additional advantage of minimizing DTW error through focusing on distinct points of interest, as significant features are less affected by speech detection noise. As

such, the similarity between text and audio data is likely to be greater around these points - this further enhances the accuracy of the associations produced by the DTW.

The anchor points are detected using a GMM to cluster minima features within the signal. This is done to identify extrema which correspond to significant features, such as breaks between dialogue sections (Figure 5.4), and thus which can be used to define segmentation points.

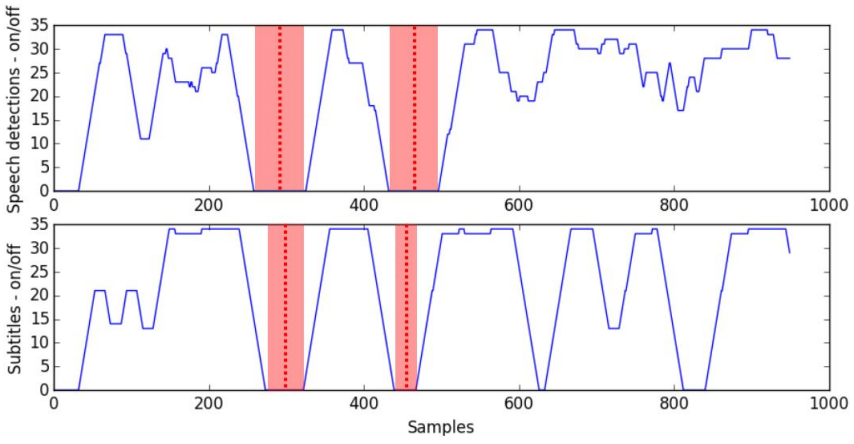


Figure 5.4: Example of corresponding signal minima.

The first step in this process is to apply smoothing to the signal to improve minima detection. This is achieved through applying convolution-based smoothing using a Hanning window. The smoothed output, $\mathbf{x}^s$, is obtained by convolving a scaled window with the signal $\mathbf{x}$. First, $\mathbf{x}$ is appended with reflected copies of itself at both ends of the signal to ensure minimisation of transient signal components at the beginning and end of the output signal.

$$\mathbf{x}' = [\mathbf{x}(-\tau), \mathbf{x}(\tau), \mathbf{x}(-\tau)] \tag{5.2}$$

The resulting signal, $\mathbf{x}'$, is then convolved with a Hanning window of size 17 samples. This window size was determined empirically through applying the approach to a range of speech detection and subtitle samples and varying the window size. The result is the smoothed signal, $\mathbf{x}^s$:

$$\mathbf{x}^s = \sum_{t=-T}^{T} h[n-t]x'[t] \tag{5.3}$$

where h is a Hanning window defined as:

$$h(n) = 0.5(1 - \cos(\frac{2\pi n}{N-1})) \tag{5.4}$$

This results in a smoothed signal which facilitates better extrema detection through smoothing 'flat' signal components, as illustrated in Figure 5.5.

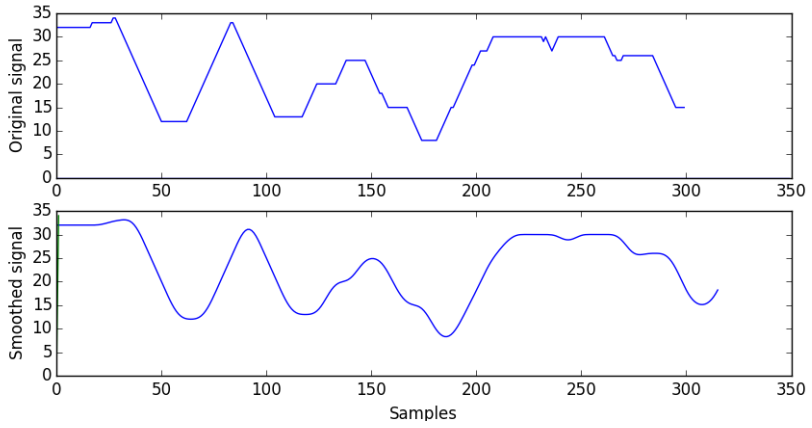


Figure 5.5: Example of signal before and after smoothing.

Once smoothed, the minima are obtained using a discrete wavelet transform-based approach.

5.3.1 Anchor Point Clustering

Following minima detection, the key anchor points are selected by constructing minima-centred features which are then clustered using a variational Bayes GMM (VB-GMM). The anchor point features were designed to convey key information about the anchor points, such as gradient and magnitude. The anchor point feature vectors, $\mathbf{a}^f$, are defined as:

$$\mathbf{a}^f = [\mathbf{a}^g, \mathbf{a}^m] \tag{5.5}$$

where $\mathbf{a}^m$ is a vector of magnitude information obtained from the smoothed signal as:

$$\mathbf{a}^m = \mathbf{x}^s_{t-\frac{w_a}{2}:t+\frac{w_a}{2}} \tag{5.6}$$

where w_a is the length of the window around the anchor point.

The gradient vector is obtained from $\mathbf{a}^m$ by computing the second-order differences for $\mathbf{a}^m_{1:N-1}$, forward differences at $\mathbf{a}^m_0$ and backward differences at $\mathbf{a}^m_N$, where N is the length of the magnitude vector, thus for each point, n to N, in $\mathbf{a}^m$:

$$\begin{aligned} if \quad n = 0 \quad & \mathbf{a}_n^g = \Delta^{\rightarrow}\mathbf{a}^m(n) = \Delta_1^{\rightarrow}[\mathbf{a}^m](n) \\ if \quad n = N \quad & \mathbf{a}_n^g = \Delta^{\leftarrow}\mathbf{a}^m(n) = \mathbf{a}^m(n) - \Delta_1^{\leftarrow}[\mathbf{a}^m](n) \\ else \quad & \mathbf{a}_n^g = \Delta\mathbf{a}^m(n) = \mathbf{a}^m(n + \tfrac{1}{2}) - \mathbf{a}^m(n - \tfrac{1}{2}) \end{aligned} \tag{5.7}$$

where Δ indicates the central difference, $\Delta^{\rightarrow}$ indicates the forward difference and $\Delta^{\leftarrow}$ indicates the backward difference.

where Δ indicates the forward difference,

This results in a feature vector, $\mathbf{a}^g$, which comprises the difference quotients of, and conforms to the same length as, vector $\mathbf{a}^m$. The final anchor point feature vector, $\mathbf{a}$, comprises all minima features $\mathbf{a}^f$ for the signal.

The features in $\mathbf{a}$ are then clustered using a GMM. While this could have been achieved using a simple rule based approach (e.g. via applying a threshold), the use of a GMM allows for other similar minima to be identified, such as significant but brief breaks in dialogue (e.g. the anchor point at ≈ 300 samples in Figure 5.6). The GMM produces a mixture of Gaussian distributions of the data, in which each distribution models a different type of extrema. In this work, a variational Bayes GMM (VB-GMM) was used. The VB GMM uses variational inference, an extension of the EM algorithm, to fit Gaussian components to the data. Like the EM algorithm, for each feature in the data, this calculates the probability that the feature was generated by a given component. Unlike the EM algorithm, variational inference also incorporates information from prior distributions to regularize model fitting. The initial priors are obtained by a Dirichlet process - a crucial reason for chooosing this approach, as this automatically determines the optimal number of components (or Gaussians) without the need for other more computationally expensive methods such as cross-validation [25] [14]. VB-GMM therefore produces a model which fits optimally, thus eliminating the need to re-fit and re-evaluate the model to find the optimal number of components, as would be necessary with a standard GMM. This is advantageous in the case of the anchor point data, as the variety of extrema 'shapes' varies greatly depending on the speech patterns in the content (and therefore the optimal number of components varies). This approach has been used to account for this variation, adapting to the data by procedurally determining the optimal number of components.

In this work, we initialise the VB GMM using a diagonal covariance matrix and the following parameters: $\alpha = 1.0$ and $max\_components = n/2$, where n is the number of anchor points in $\mathbf{a}$.

Once clustered by the GMM, the anchor points are selected according to the degree of the extrema. Observations from the data demonstrate that segments can be more effectively associated by larger features, such as long speech segments, rather than by less significant

features, such as fine variations in speech activity. As such, the segmentation focuses on finding more significant extrema, as these typically lie between more significant features, as demonstrated in Figure 5.6. To do so, the mixture with the lowest mean is selected, as this corresponds to the anchor points with the lowest magnitude, and thus to the minima that are likely to lie between, rather than among, segments of continuous speech. While basic minima separation could be achieved by implementing a simple threshold, the incorporation of the GMM and gradient data allows for other similar minima to be identified, such as significant but brief breaks in dialogue, e.g. the anchor point at ≈ 300 samples.

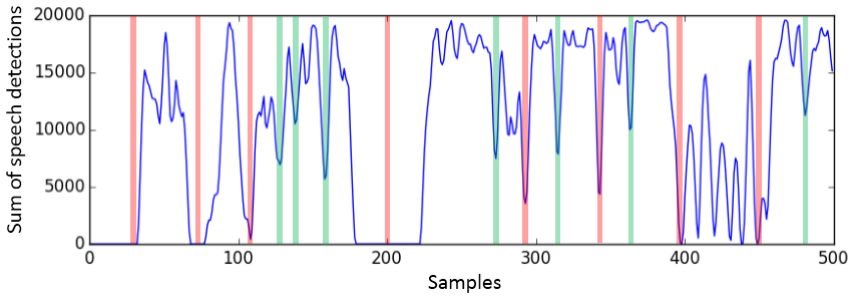


Figure 5.6: Example of more extreme minima separating significant regions of speech activity. Red: extreme minima. Green: less extreme minima.

5.3.2 Audio to Text Association

One of the goals of this work is to provide a language independent method for audio and text association. This is achieved using anchor point-based segmentation in combination with DTW. Anchor points are used to define breaks between segments, and thus to define the segments to be matched. Once the anchor points have been identified, DTW is applied to the audio and text signals. Given a source and a reference signal, anchor points in the reference signal are used to find matching points in the source signal. In this case, we use the speech detector data as the source signal, as this is prone to false detections, whereas the subtitle data is assumed to be free of noise, and therefore provides a better reference for anchor point selection.

This results in a vector mapping all anchor points obtained from the text to corresponding points in the audio, as demonstrated in Figure 5.7. The association between audio and text segments can therefore be obtained from the anchor point mapping.

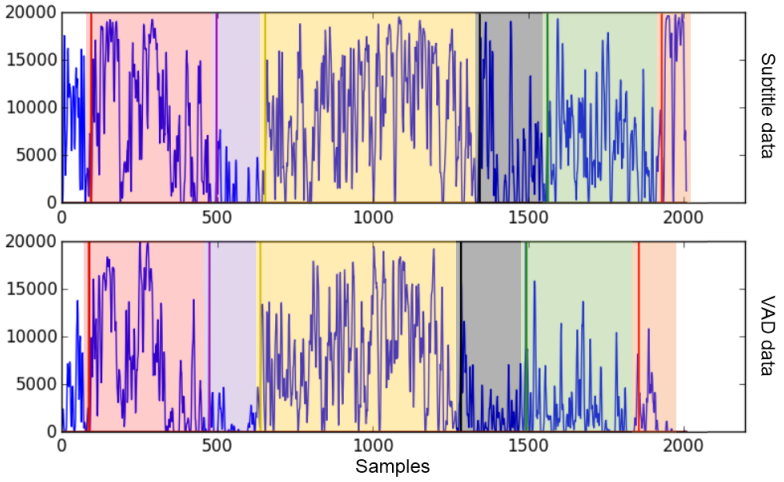


Figure 5.7: Example of anchor point matching via DTW on feature film data. Segments are coloured to reflect the mapping between the subtitle and VAD data.

5.4 Audio to Text Association of Whole Film Content

Two potential applications have been explored for audio-to-text association. The first application is focused on audio/text segment matching, and is evaluated by comparing the audio and text anchor mappings to the ground truth. The second application focuses on exactly aligning the signals, for use in automatic subtitle or audio alignment. This is evaluated by calculating the difference between an estimated scale coefficient and a target scale coefficient. Due to the application context proposed by the industrial partner, this work assumes a linear relationship between the time bases of the text and audio signals, as such it seeks to find a single scale coefficient to align the data.

Given the length of feature film data, and the fact that we are interested in general, rather than fine, alignment, the audio and text signal lengths are scaled down by a factor of 50. This retains the signal shape while reducing the size of the vectors to be processed, thus reducing the time and resources required for processing.

To evaluate the performance of the matching and alignment methods, they have been applied to data from the following four feature films:

- Frankenweenie
- Brave
- John Carter
- Pirates of the Caribbean: Dead Man's Chest

The first three films were chosen as text content was provided by the industrial partner. The last film was chosen to balance the dataset between live action and animated films, and was chosen arbitrarily as the content was readily available.

5.4.1 Start Point Alignment

Prior to segment matching, the start points of the signals are aligned to enhance the efficacy of the matching and alignment algorithms. This is achieved by using DTW to align the first anchor point. While the signals could be aligned by shifting them to match the corresponding non-zero components at the beginning of the signals, there is no guarantee that these are equivalent, i.e. there could be false detections from the speech detector at the signal start which would result in an incorrect initial alignment. The use of anchor points and DTW is therefore preferable for initial alignment as it ensures that the features are aligned according to similarity, and thus increases the likelihood of accurate initial alignment.

5.4.2 Anchor Point Evaluation

Anchor point accuracy is evaluated by calculating the error between the target anchor point and the scaled source anchor point in minutes. Given a target anchor point, a^t, a source anchor point, a^s, and a scaling coefficient, c, the error is calculated as:

$$e = |a^t - (a^s c)| \tag{5.8}$$

This is computed for each anchor point pair across both segments for all films in the dataset. As Figure 5.8 demonstrates, Frankenweenie achieves the strongest results, with the lowest mean and standard deviation results for anchor point error, while poorest performance is exhibited on John Carter, with the mean error at $\approx 0.69mins$. The mean of both the mean and standard deviation across all samples is $\approx 0.5mins \approx 0.45mins$ respectively. Given an average segment duration of $\approx 13mins$, this indicates reasonable segmentation performance, with the source anchor points being, on average, within 5% of the segment length. This in turn suggests that the segments are strongly associated, and that they should achieve reasonable results for segment matching. However, this degree of error may be too great to achieve an accurate estimate of scale, suggesting that the approach may not be sufficient for accurate scale estimation.

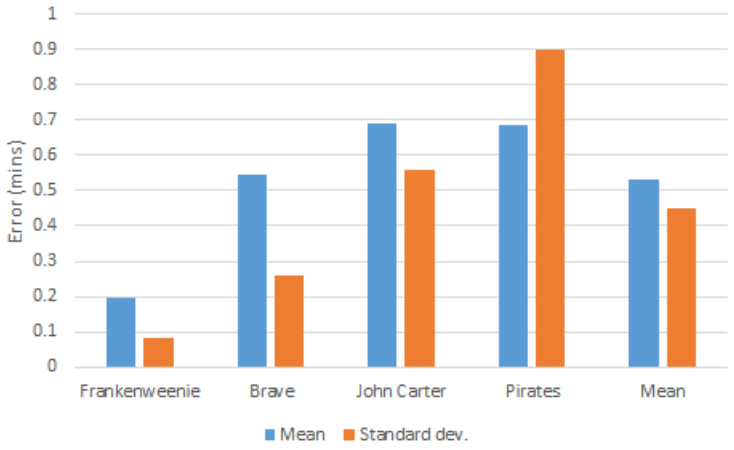


Figure 5.8: Anchor point evaluation results for VAD and subtitle alignment of whole film data.

5.4.3 Segment Matching

Segment matching is performed using the anchor point detection and DTW approach described previously. The segment matching described here is being considered for rough audio to text association, and as such does not evaluate word-level accuracy, but is instead concerned with segment-level accuracy.

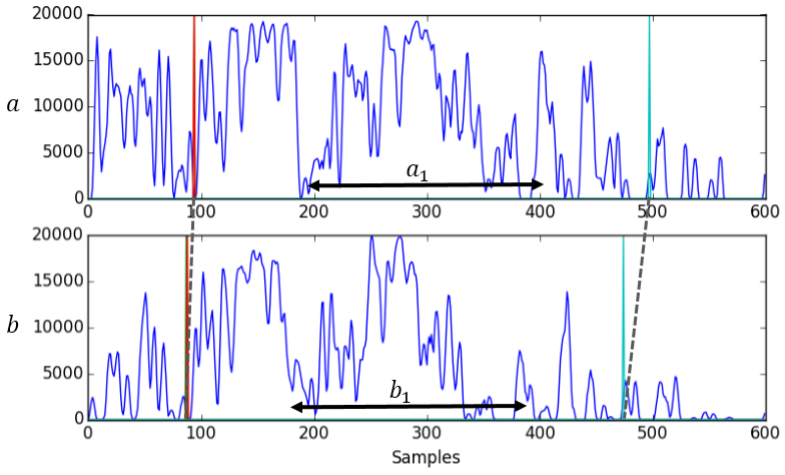


Figure 5.9: Segment matching example mapping a segment in signal b (segment b_1) to a segment in signal a (segment a_1). Segments are marked by segment start (red) and end (blue) anchor points.

Figure 5.9 illustrates how segment matches are evaluated. Given a pair of segments

across the target and source signals, a and b, the source segment is scaled by the scaling coefficient, c. If scaled segment bc overlaps the target segment by $>= 80\%$, the segments qualify as a match. The value of 80% was chosen as this permits some margin for variation while ensuring that the segments contain a significant degree of mutual content. The segment matches are further validated by manually comparing the audio and text content for each match to ensure that matching segments are appropriately classified.

The results presented here are given as the Positive Predictive Value (PPV) i.e.:

$$PPV = \frac{correct \quad matches}{all \quad matches} \tag{5.9}$$

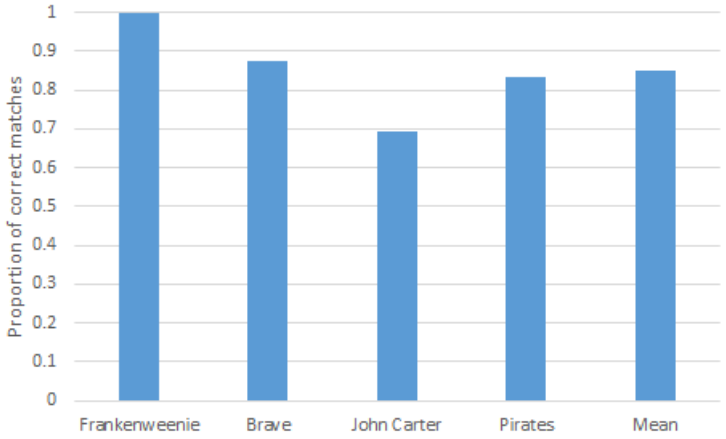


Figure 5.10: Segment matching results use subtitle and VAD data.

As Figure 5.10 demonstrates, the algorithm correctly matches a significant proportion of segments - achieving a mean of $\approx 85\%$ correct matches overall. As with the anchor point evaluation, the strongest results can be observed for Frankwenweenie, while the weakest results were obtained using data from John Carter. This is to be expected, as a greater degree of anchor point error indicates poorer inter-signal segmentation. Overall these results demonstrate that the combination of anchor point segmentation and DTW-based mapping can be used for successful audio-to-text association, with $> 80\%$ correct matches achieved for a majority of the data sets tested.

5.4.4 Scale Coefficient Estimation

One of the aims of this work is to provide a low resource, language independent method of re-aligning subtitle content to films distributed in different formats. One such example is the conversion of NTSC to PAL, which involves altering the frame rate from 24 fps to

25 fps. This alters the time base by a factor of 1.04167 which, while subtle, results in misalignment, particularly with longer media such as feature films. To correct for this, we propose an approach which utilises the combination of unsupervised anchor point selection and DTW described previously.

Thus, the approach first performs anchor point selection on the subtitle data, after which DTW is applied to the signals to obtain the corresponding points in the speech detection data. The scale coefficient estimate, sc_t, is then simply obtained by:

$$sc_t = \frac{a_t^{sub}}{a_t^{spd}} \tag{5.10}$$

where a_t^{sub} is the time at a given subtitle anchor point and a_t^{spd} is the time at the corresponding speech detection anchor point.

Given that sections 5.3.2 and 5.3.3 demonstrate that anchor point misalignment will occur, the final scale factor is computed by taking the median of all scale factor estimates, sc_t for all anchor point pairs in a^{sub} and a^{spd}. This has proven to be more effective than taking the mean, which is less accurate due to outlying scale estimates produced by misalignments, as demonstrated in Figure 5.11.

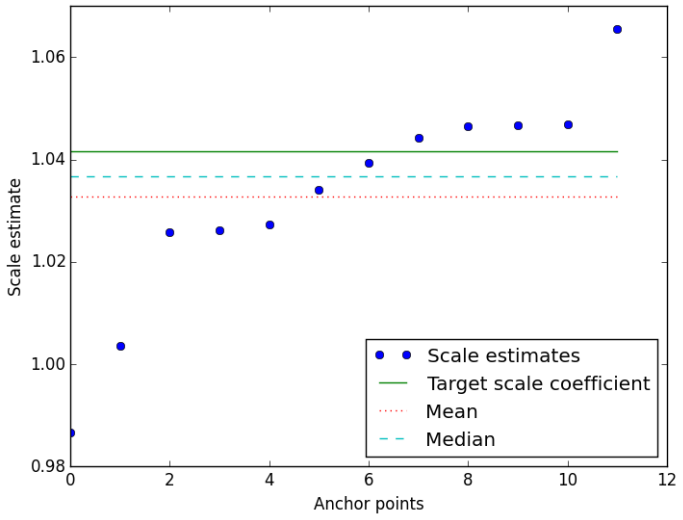


Figure 5.11: Plot of scale estimates obtained from anchor point matches.

For the films in the dataset, the subtitles have been obtained from PAL sources, while the films themselves are NTSC. The subtitles therefore need to be rescaled by factor of 1.04167 to align with the PAL framerate. The alignment method is evaluated using the scale coefficient error, sc^{err}, which is defined as:

Dataset	Scale estimate	Scale error	Scale error %
Frankenweenie	1.0451	0.0034	0.3294
Brave	1.0509	0.0093	0.8893
John Carter	1.0386	0.0031	0.2945
Pirates of the Caribbean	1.0448	0.0032	0.305
Mean	-	0.0047	0.455

Table 5.1: Scale factor estimate results.

$$sc^{err} = |sc^{target} - sc^{estimate}| \tag{5.11}$$

where sc^{target} is the target scale coefficient of 1.04167 and $sc^{estimate}$ is the scale coefficient estimate produced by the alignment process.

As demonstrated in Table 5.1, the alignment method does a reasonable job of estimating the scale factor, with a mean error of 0.45% across all datasets. Interestingly, while the worst results for the matching investigation were obtained for John Carter, it obtains the best results for scale factor estimation. This indicates that, while it didn't achieve as many strong segment matches, the anchor points corresponding to the scale estimate median were more accurately associated.

While the estimates produced here are close to the target value of 1.04167, they would be unsuitable for automatically realigning subtitle data. This is due to the cumulative effect of the error over the duration of the media. For example, consider a set of subtitles scaled to a 60 minute piece of media which need to be rescaled to its NTSC counterpart. The counterpart will have a duration of $3750s$. Rescaling this incorporating an error of $\approx 0.9\%$ (e.g Brave) would result in a duration of $\approx 3780s$. Thus, the subtitles towards the end of the media would be out of synchronization by around 30s - demonstrating that the approach is not suitably accurate for automatic subtitle re-alignment.

The underlying reasons for this error are likely twofold:

1. **Subtitle timing**: as mentioned earlier, subtitles are not designed to align perfectly with speech, and are instead created with readability in mind. While these give a reasonable impression of the location of dialogue, they will not align perfectly, and therefore will introduce error into the system.

2. **Speech detection errors**: while the speech detection approach has proven to work well on entertainment media, it achieves an accuracy of between 85% and 90%. Thus, the speech detector will contain errors, both in the form of false positives and false negatives, which may subsequently contribute to alignment errors.

Nevertheless, this does provide a method for rough alignment and association of audio and text information, and is therefore useful as a means of aiding post-production through automatically selecting and roughly aligning associated segments.

5.4.5 Transcript Alignment and Matching

As transcript data was provided by the industry partner, investigations into the use of transcript data, rather than subtitle data, for alignment and matching were carried out. These use the same processes used for subtitle alignment and matching, but replace the subtitle signal with a simulated speech in/out signal generated from the transcripts. As the transcripts only contain speech *in* times, the *out* time is simulated by simply counting the number of words in a passage and dividing this by the average number of words per second in English, which is 2.5 [122]. Thus, for each transcript entry, the duration is given by:

$$duration\ (seconds) = \frac{n\ words}{2.5} \tag{5.12}$$

This can then be used to create a pulse signal such as those generated by the subtitle and VAD data.

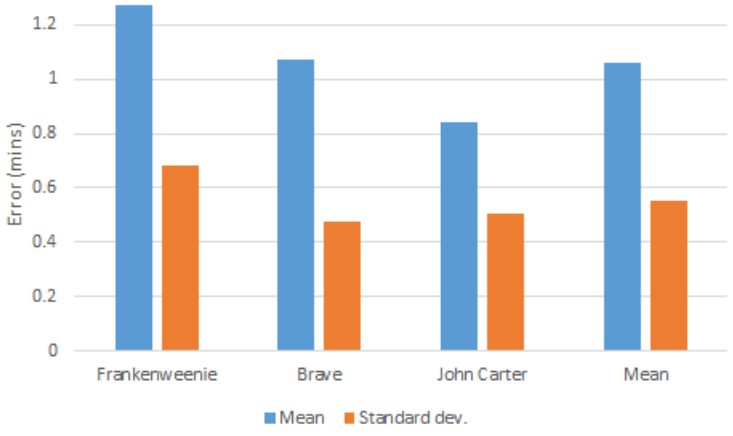


Figure 5.12: Anchor point evaluation results for VAD and transcript alignment of whole film data.

As with the subtitle oriented approach, the anchor points are first evaluated by computing the error between target and source anchor points. As Figure 5.12 shows, there is a rise in anchor point error across the board when using the transcript data.

As Figure 5.13 illustrates, the segment matching results are similar to the trend demonstrated in the anchor point evaluation. This is as John Carter achieves the strongest results,

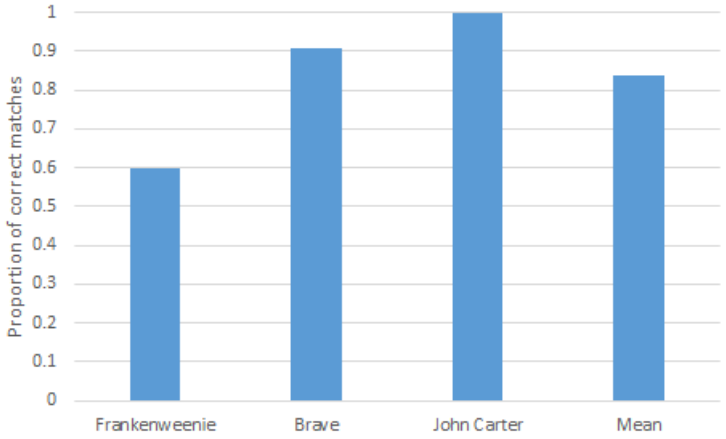


Figure 5.13: Segment matching results using transcript and VAD data.

Dataset	Scale estimate	Scale error	Scale error %
Frankenweenie	1.0678	0.0262	2.491
Brave	1.0647	0.0231	2.195
John Carter	1.0533	0.0116	1.108
Mean	-	0.02	1.932

Table 5.2: Scale factor estimate results.

while fewer successful matches are obtained for Brave and Frankenweenie. A similar pattern can also be observed for the scale factor estimation results in Table 5.2. These results demonstrate a fall in scale estimate accuracy across the board when compared with the subtitle data, indicating that the transcript data is not as useful for scale coefficient estimation. This is probably due to the use of the average speaking rate value used to estimate speech *out* times. The speech *out* times produced by this approach are likely to vary greatly in accuracy depending on the speaker and the type of dialogue. On the other hand, while the subtitle information is modified for readability, this will still match the action on screen, thus its timing variations will more closely match the variations in the spoken dialogue.

While it would have been interesting to explore the use of transcript data further, the results here indicate that subtitles are more useful for segment matching and scale estimation. Further to this, the industry partner was primarily interested in applications of subtitle data, as transcripts themselves are rarely used in the localisation process. As such, the remaining work focuses on the use of subtitles as the form of textual information.

Film	**Dataset #**	**Duration (minutes)**
Silence of the Lambs	1	01:35
Silence of the Lambs	2	02:31
Silence of the Lambs	3	02:28
Pulp Fiction	4	02:07
Pulp Fiction	5	03:36
John Carter	6	02:04
John Carter	7	02:59
John Carter	8	01:42
A Few Good Men	9	02:03
A Few Good Men	10	03:10

Table 5.3: Dataset used for development and testing of fine alignment approach.

5.5 Scene-Level Alignment

While general alignment of audio and text information is useful, fine alignment is also desirable. This is as it facilitates associations of finer features, such as speech content within a scene, rather than general alignment which associates segments from the film as a whole (segments which may or may not correspond to scenes or have other semantic meaning). Alignment of finer features is therefore useful for realigning subtitles of individual sections, and for automatically identifying audio/text associations for the post-production process.

Interestingly, while DTW is effective for aligning media with longer durations, it is less effective for aligning shorter excerpts. This section evaluates the performance of DTW for aligning shorter pieces of multimedia, and proposes a novel approach for alignment based on the anchor point features discussed previously.

Short scenes from a number of films were selected for use in developing a new alignment approach and for testing the performance of DTW on short extracts. As the work also looks to investigate the incorporation of visual speech features, scenes involving faces were specifically selected. As such, it was necessary to use live action films, as the V-VAD approach will not work on animated media. The details of the dataset assembled for this investigation are given in Table 5.3.

For the scene-level data, rescaled data has not been used. As such, both the reference and query signals are of the same scale, thus the target scale coefficient is 1.0.

As with the whole film content, the scene-level content is also used to evaluate segment matching and scale coefficient estimation for alignment, using the same methods used for whole film investigations.

5.5.1 Segment-Based Alignment

Due to the poor performance observed when using DTW for aligning scene-level data (as demonstrated in Section 5.4.3), a segment-based alignment approach was developed. This works similarly to DTW in that it uses a cost matrix to map similar features between the two signals. Unlike DTW, this approach focuses on matching individual segments, and ignores finer signal features. In this way, it is more robust to fine-level noise, as it only considers larger features.

Another parameter used to ensure focus on broader features is the minimum distance between segmentation points. This is used to ensure that anchor points are separated by a minimum distance of 5 seconds. In this way, only minima which separate utterances of significant duration are considered, thus increasing the likelihood that the resulting segments are associated with speech content, rather than noise such as false detections. The value of 5 seconds was chosen according to empirical observations, which demonstrated that segments of shorter durations were more prone to being affected by speech detection noise.

The proposed approach for segment-based matching and alignment is as follows. After anchor point segmentation, the cosine similarity between each segment in the reference signal is computed for each segment in the query signal. The segments are constructed using the anchor points as the start points, and the end of the signal as the end point. Thus, each segment can be described as:

$$\mathbf{seg}_t = \mathbf{signal}_{a_t:end} \tag{5.13}$$

where a_t is the anchor point at time t. This results in a cosine similarity matrix which can then be used to find optimal matches between segments. To reduce the likelihood of incorrect alignments, we constrain the search space to the $n/2$ nearest segments, where n is the total number of segments. As with DTW, the segments are matched using a dynamic programming-based algorithm to navigate the similarity matrix. This algorithm is defined as:

```
query_anchors -> vector containing query signal anchor indices
reference_anchors -> vector containing reference signal anchor indices
cost_matrix -> cosine distance matrix
search_limit -> n/2
min_index -> index of minimum value


while i < cost_matrix[:,i] and j < cost_matrix[:,j]:
    obtain index of minimum value in cost_matrix[i][j:j+search_limit]
    update mapping_vector with mapping = [i, min_index]
    if length of reference_segment > length of query_segment:
        increment j to index of closest anchor in reference_anchors
        i++
    else if length of query_segment > length of reference_segment:
        increment i to index of closest anchor in query_anchors
        j++
    else:
        i++
        j++
```

Here, the cost matrix is navigated first by the reference segments via i, for which the minimum distance in the query segments is obtained by evaluating:

$$argmin(cost\_matrix[i][j : j + search\_limit]) \tag{5.14}$$

This gives the minimum cost matrix value for query segments $j : j + search\_limit$, from which the resulting i and j indices can be used to obtain the reference-to-query signal mapping, shown here in the form $[i, min\_index]$. Following this, the algorithm updates the query and reference anchor vectors according to the position of the subsequent anchors. This is done by first evaluating whether the distance between the current and subsequent query or reference anchors is longer, and then by incrementing the other anchor vector appropriately. For example, if the reference segment is longer than the query segment, the algorithm iterates the reference anchor vector by 1, and the query anchor vector to the index of the closest query anchor point, defined as

$$argmin(|reference\_anchor[i + 1] - query\_anchor[j]| \quad for \quad j \quad to \quad J) \tag{5.15}$$

This is done as it is assumed that a linear relationship exists between the two signals. Therefore, if the distance between the current and subsequent reference anchor points is greater than the distance between the query anchor points, it is likely that the reference signal does not contain a corresponding anchor point, and thus the query anchor point vector should increment to an anchor point which more closely matches the anchor point

in the reference vector. This is illustrated in Figure 5.14.

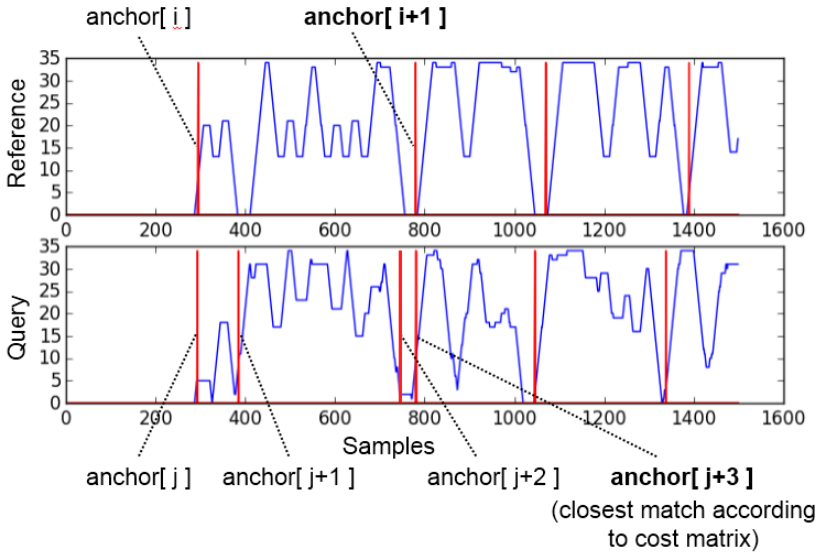


Figure 5.14: Illustration of anchor incrementation by 'closest' anchors. Bold typesetting is used to indicate the matching anchor points discussed in the text.

As shown here, the reference anchor point at $[i+1]$ can be more closely associated with the query anchor point at $[j+2]$ than at $[j+1]$. As such, j is incremented to the more closely matching anchor point in the query anchor point vector, and the anchor at $[j+1]$ is discarded by the alignment process as it does not closely match an equivalent reference anchor point. The closest anchor point is then identified by evaluating cost matrix for $j : j + search\_limit$, which in this case is $anchor[j+3]$.

As illustrated previously in the pseudocode, this process is repeated for all segments in the reference and query signals, producing a vector which contains the reference-to-query mapping for each anchor point and corresponding segment. In order to compare the performance of this approach with DTW, both methods have been evaluated for scene-level segment matching and scale coefficient estimation.

5.5.2 Anchor Point Evaluation

As with the whole film investigations, audio/text association is first investigated by evaluating anchor point selection. Given that the scene-level data is of higher resolution, the anchor point error is given in seconds.

Figure 5.15 demonstrates that, while the proposed approach exhibits greater error for some samples in the dataset (e.g. sample 7), overall it achieves better performance than

the DTW-based approach. This performance advantage is reasonably significant, with an average mean and standard deviation of $\approx 2s$ compared with values of $> 3s$ for the DTW-based approach.

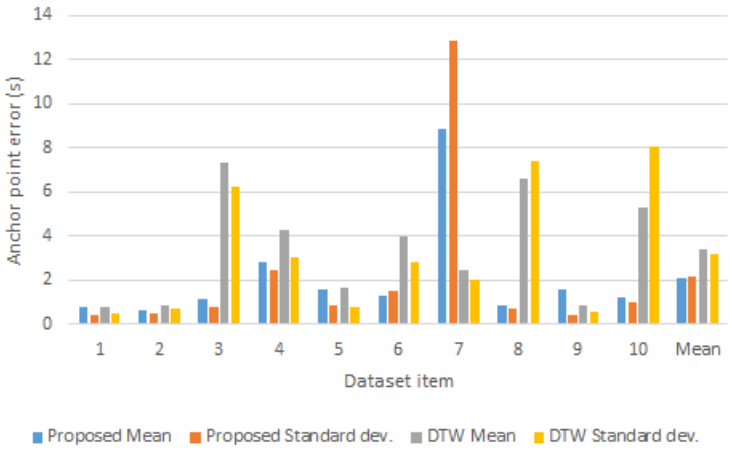


Figure 5.15: Anchor point evaluation results for VAD and subtitle alignment of scene-level data.

While this demonstrates a clear improvement over the DTW approach, a mean error of $2s$ is fairly significant given the fine resolution of the data. This is likely an unavoidable product of the higher resolution data, as subtitles will frequently be of greater duration than the corresponding audio by several seconds in order to ensure readability. This, and the increased impact of speech detection errors on finer resolution data, are likely the key factors contributing to anchor point errors.

5.5.3 Segment Matching

This section explores both DTW and the proposed method for the task of segment matching. Figure 5.16 demonstrates a similar trend to that observed in the anchor point evaluation, with DTW demonstrating enhanced performance on some samples, and the proposed approach achieving better overall results. Another notable factor is the significant drop in matching performance between the DTW here and the whole film results. Whereas the DTW-based approach achieved $> 80\%$ correct matches for the whole film data, it achieves an average of only 50% correct matches for the scene level data (Figure 5.16). This clearly illustrates DTW's significantly reduced performance on the scene level data and validates the decision to develop an alternative approach. In contrast, the proposed approach achieves a mean of $\approx 75\%$ correct matches. This demonstrates marginally worse performance than

achieved for the whole film data, a result that is likely due to the increased impact of subtitle and audio inconsistencies on finer resolution data.

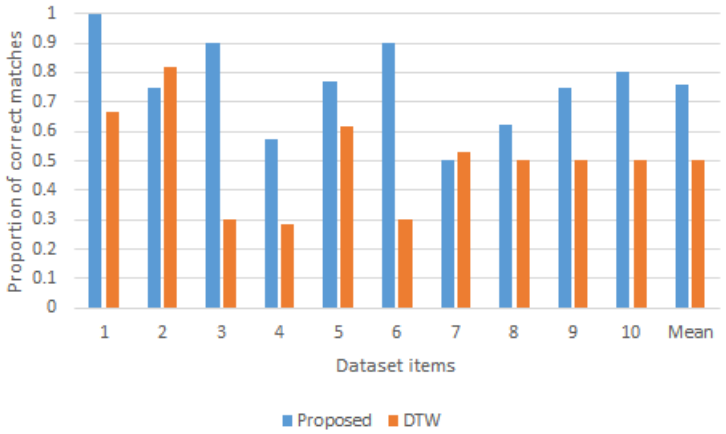


Figure 5.16: Segment matching results for scene-level data.

The cause of the DTW's poor performance is likely due to reduced smoothing of the scene level features, as these use a much smaller summing window. This results in a greater proportion of flat signal segments (Figure 5.17), which have proven to reduce the performance of DTW's warping algorithm in previous work [71]. In addition to this, the reduced smoothing also increases the impact of noise, i.e. inconsistencies between the VAD and subtitle signals. This is likely also a contributing factor, as previous work demonstrates that DTW has a tendency to latch onto noise, particularly in data which contains a significant proportion of flat segments [90].

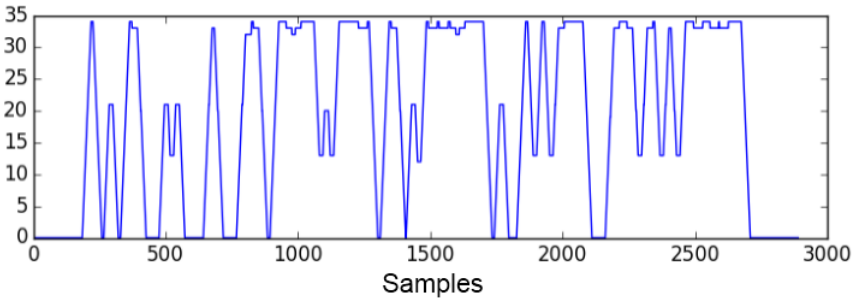


Figure 5.17: Example signal containing flat features.

This could be improved with the use of a larger summing window, however the window size was selected to maximise smoothing while minimising reduction of signal features for the finer level data. As such, while a greater window size would enhance the smoothing

effect, it would also smooth distinct features necessary for accurate segment matching and alignment.

On the other hand, the proposed approach performs relatively well as it focuses on segments rather than on mapping every point from the two signals. This allows it to ignore finer-resolution noise, making it more effective in the presence of noisy data. The result is a matching approach that does not get stuck in finer level inconsistencies between the two signals, and thus produces better overall performance.

5.5.4 Scale Coefficient Estimation

Scale coefficient estimation was evaluated using the same method described in Section 5.1.2. As demonstrated in Figure 5.18, the proposed approach generally performs better for scale coefficient estimation when compared with DTW, with a scale estimate error of 0.013 compared with the DTW-based methods error of 0.027. Most importantly, while the DTW-based approach achieves better estimates on some data, the proposed approach is far less variable, with the maximum scale error never exceeding 0.03, while the DTW-based method exhibits more significant errors of > 0.06. This demonstrates that, not only does the proposed approach achieve better results overall, but it exhibits far more consistent performance, indicating that it is a more robust method for scene-level scale estimation and alignment.

Another point to note is that the scale estimate error here is greater than the error observed for the whole film data. This is unsurprising given that the matching investigation has already revealed poorer performance for segment association, which in turn will result in less accurate scale estimation.

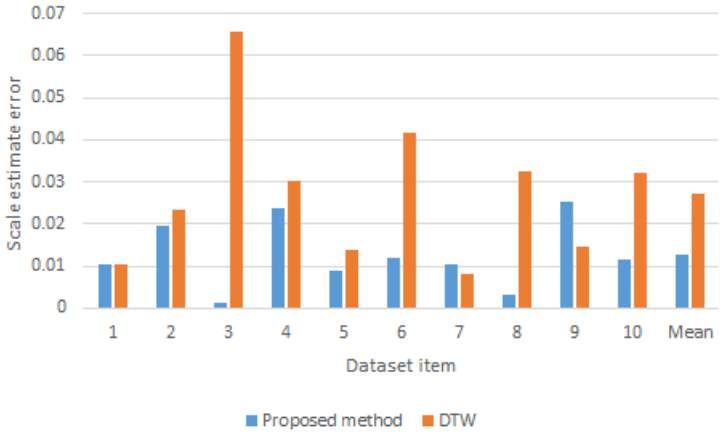


Figure 5.18: Scene-level scale estimate results.

5.5.5 Incorporating Visual Features

Two investigations into the incorporation of visual speech information have been explored. Both use the V-VAD detector output, in a similar way to which the VAD detector output has been used. The first investigation explores audio to visual alignment of content using VAD and V-VAD output. The second explores the use of V-VAD as a method of enhancing audio to text alignment. This section uses the sample dataset as the previous section (detailed in Table 5.3).

Audio to Video Alignment

The audio to video alignment method uses the same as the audio to text method described 5.4.1, however in this case the output from the V-VAD is used in place of the text data. As with the earlier approach, the V-VAD and VAD data is processed using a summing window of size 17 prior to the segmentation and matching steps. The results show the scale coefficient error, given as the difference between the estimated scale coefficient and the target scale coefficient.

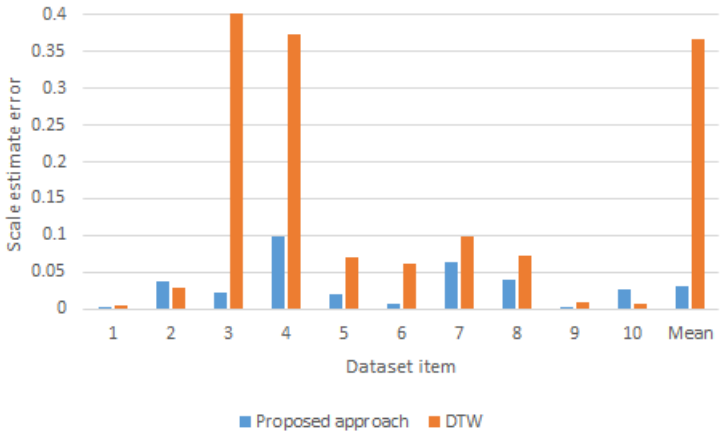


Figure 5.19: Audio to video alignment error. *Note: DTW results for item 3 = 2.9, and thus exceed the scale of the plot.*

As demonstrated in Figure 5.19, the audio to video alignment approach can achieve reasonable results, with an error of as little as 0.001 (0.1%) when using the proposed approach. As with the audio to text alignment on scene-level data, the proposed segment-oriented approach outperforms DTW for alignment tasks. These results are encouraging, and suggest that the combination of audio VAD and V-VAD methods developed through this work has potential for use in audio to video alignment.

V-VAD Enhanced Audio to Text Alignment

The V-VAD approach described in Chapter 4 has also been explored as a method of enhancing audio to text alignment. This has been achieved by incorporating the V-VAD output as part of a confidence scoring mechanism for anchor point filtering. To do so, the Pearson correlation is computed for a window of size 10 around each anchor point in the VAD output to provide the correlation coefficient for the V-VAD and VAD data at this point. The correlation coefficient at each anchor is then compared to the mean of all anchor-centred correlation coefficients. If the correlation coefficient is less than the mean, the anchor point is discarded.

The underlying principal here is to utilise the additional modality to filter out false detections: if both the V-VAD and VAD indicate speech activity, the region is less likely to contain false detections. Thus, by using the VAD/V-VAD correlation, it is possible to determine the extent to which the detectors agree, and thus filter the anchor points accordingly.

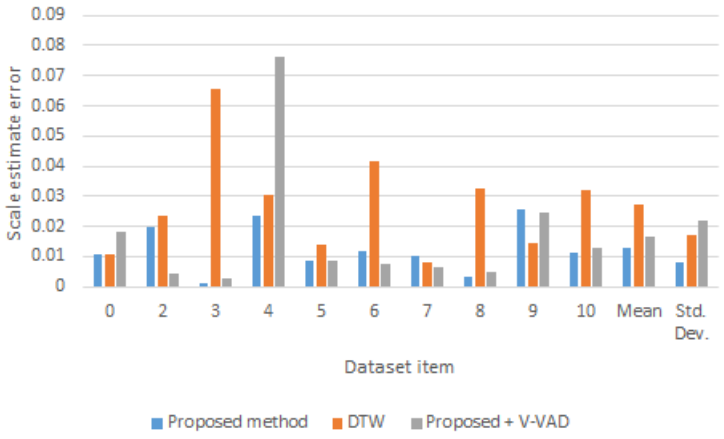


Figure 5.20: Scene-level scale estimate results for proposed approach, V-VAD enhanced approach and DTW-based approach.

As Figure 5.20 demonstrates, the V-VAD enhanced approach achieves better results than the DTW-based approach, but does not perform quite as well as the VAD-only approach overall. This can be explained by variable quality V-VAD data, as illustrated by comparing the standard deviation of the three approaches. While the proposed method has the lowest standard deviation, the V-VAD enhanced approach exhibits the highest value across the dataset. As the data demonstrates, the V-VAD enhanced approach performs very well on certain items, achieving the strongest performance on items 2, 6 and 7. This indicates that the V-VAD performed well on these datasets, thus enhancing the scale estimation process.

On the otherhand, the V-VAD enhanced approach produced poorer results on items 1, 4 and 9 - indicating that the V-VAD performed badly on these samples. This can be confirmed by comparing the V-VAD and VAD sum plots for the corresponding samples from the dataset.

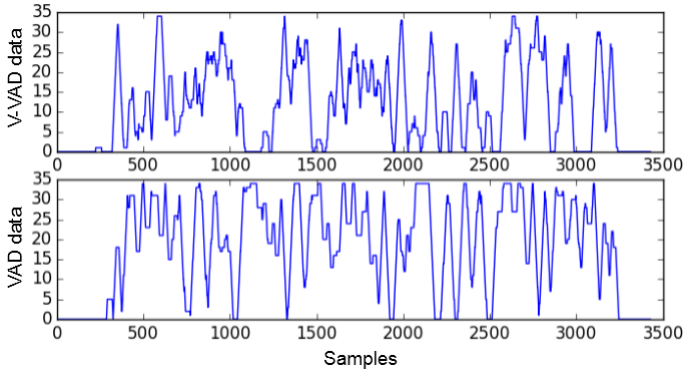


Figure 5.21: Plot of V-VAD sum data and VAD sum data from dataset item 2.

As Figure 5.21 demonstrates, there is consistent V-VAD output for the sample, indicating that the face detection and feature localisation performed favourably on the data. The V-VAD was therefore more successful, and could be used as a means of confidence scoring, producing stronger anchors which could be used to enhance the alignment process.

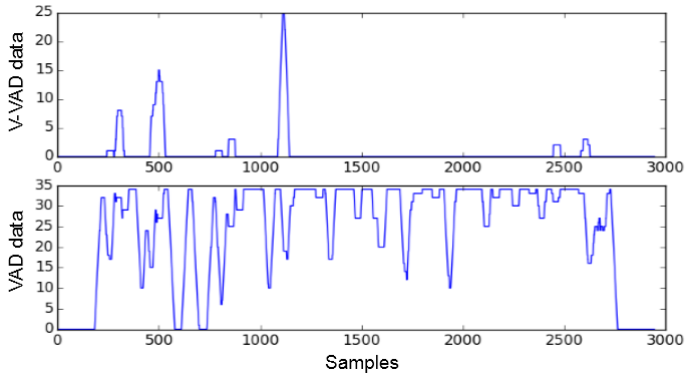


Figure 5.22: Plot of V-VAD sum data and VAD sum data from dataset item 4.

In contrast, as shown in Figure 5.22, the V-VAD and VAD plots for sample 4 are starkly dissimilar - with the V-VAD missing a significant amount of information. As such, the resulting confidence estimates are inaccurate, resulting in a poor selection of anchors and an inaccurate scale estimate.

Conclusion

Section 5.4.5 has explored the use of V-VAD data both as a means of audio to video alignment and as a means of enhancing the audio to text alignment process. The results demonstrate that the V-VAD is capable of strong performance in these tasks on some samples, while overall it produces poorer results when compared to the VAD-only text alignment method. Observations of the V-VAD and VAD data in Figures 5.21 and 5.22 indicate that the variable performance is due to poor V-VAD performance on some data. This is likely due to poor face detection and feature localisation data, produced by particularly challenging video content. As such, we conclude that a more accurate V-VAD approach, coupled with more robust face detection and feature localisation methods, is required before a V-VAD based approach can be confidently used on feature film media.

Nevertheless, this work demonstrates that good performance can be achieved with V-VAD data for both audio to video alignment and enhanced audio to text alignment. The investigations here also further support the use of the proposed alignment approach, with it achieving significantly better results for audio to video alignment when compared with DTW.

Given the results in this section, the V-VAD approach has not been used for further alignment investigations.

5.6 Improving General Alignment Through Incremental Scene-Level Alignment

The general alignment method discussed in Section 5.1.2 produces a scale error of up to 0.879%, which is only sufficient for rough alignment of content. To improve upon this, an incremental alignment method has been explored. This method uses the initial scale estimate as a starting point - rescaling the signal prior to applying the scene-level alignment method incrementally to signal segments. This facilitates two key functions:

1. It provides a more precise estimate of segment associations between the text and audio material by focusing on finer signal features.

2. It provides a better estimate of the scale factor through using information obtained from a range of signal segments.

The first step in this process is to apply the general alignment method and re-scale the signal to the resulting scale estimate. This produces a roughly aligned signal, which

improves the fine alignment process through ensuring that the beginning of the signals are within reasonable alignment.

Once the initial estimate has been applied to roughly align the subtitle signal, the fine alignment process is applied incrementally to segments of 10 minutes' duration. This duration was chosen to maximize the likelihood of reliable speech detection content while minimizing the impact of the scale estimate error. For example, given the worst estimate from Table 5.1, the alignment would be out of sync by approximately 7 seconds towards the end of the 10 minute segment. This is a significant improvement on using the original signal, which would have been out of alignment by 30 seconds at this point. While a shorter segment would improve on this alignment error, the use of a longer segment increases the likelihood that there will be sufficient speech and dialogue data for the alignment process, and also that there will be enough speech detection data of sufficient quality for the alignment process to be effective (i.e. a reasonable amount of low noise speech detection information).

For each segment, the fine alignment approach is applied, and a scale estimate is produced. The whole signal is then updated by applying this estimate to the segment, and the scale estimate is stored in an array. The pre and post alignment correlation is also calculated for each signal segment, from which the alignment correlation difference is obtained:

$$corr^{diff} = corr^{post} - corr^{pre} \tag{5.16}$$

where $corr^{pre}$ and $corr^{post}$ are the pre and post-alignment correlation values obtained as the Pearson product-moment correlation coefficient.

This gives an impression of the efficacy of the alignment process for each segment, and is used to filter the resulting scale estimates for overall realignment. The correlation difference is stored in an array, and the process is repeated for all signal segments.

Once the fine alignment process has been applied to all signal segments, the resulting scale estimates and correlation differences are used to filter the results and obtain a second scale estimate, which is used to correct the initial estimate. The scale estimates are selected according to correlation improvement, the idea being that the best estimate will correspond to the most accurately aligned segment, and will thus produce the greatest improvement in correlation.

As such, the estimate corresponding to the greatest value of $corr^{diff}$ is obtained by finding the respective index, and using this to obtain the scale estimate from a vector of all scale estimates, $\mathbf{e}$:

Scale estimate		
Dataset	**Initial**	**Corrected**
Frankenweenie	1.0451	1.0453
Brave	1.0509	1.0452
John Carter	1.0386	1.0413
Pirates of the Caribbean	1.0448	1.0437

Table 5.4: Scale factor estimates before and after incremental alignment.

Scale estimate error [%]		
Dataset	**Initial**	**Corrected**
Frankenweenie	0.3294	0.3448
Brave	0.8893	0.337
John Carter	0.2945	0.0335
Pirates of the Caribbean	0.3054	0.198
Mean	0.455	0.2283

Table 5.5: Scale factor estimate error results before and after incremental alignment. Error given as percentage of target scale factor.

$$s^c = \mathbf{e}[index^d(argmax(corr_i^{diff} \quad for \quad all \quad corr_i^{diff} \quad in \quad \mathbf{d}))] \tag{5.17}$$

where s_c is the estimate from the incremental alignment process and **d** is a vector containing values of $corr^{diff}$ for all segments. The final scale factor, s^f, is then obtained by applying s^c to the initial scale estimate, s^i, obtained from the general alignment process:

$$s^f = s^i \times s^c \tag{5.18}$$

As demonstrated in Tables 5.4 and 5.5, this improves overall alignment, with improvements observed for all data except Frankenweenie, for which the scale error increases marginally. The improvement is fairly significant, with the mean scale error over all datasets reducing from 0.455% to 0.228%. This shows that the incremental alignment approach has been successful in improving on the initial scale estimates, demonstrating that better alignment can be achieved through combining the general and fine alignment approaches.

While the incremental alignment approach produces a notable improvement in estimating the scale factor, the error is still too great for practical use. This can be illustrated when considering the earlier example of an hour of data. While scaling content of duration $3600s$ by a factor of 1.04167 would produce duration $\approx 3750s$, scaling the content by the

least accurate scale estimate of 1.0453 (corresponding to Frankenweenie) would scale the content to a duration of $\approx 3763s$. While this is a clear improvement over the initial scale factor, the alignment is still out by around $13s$ towards the one hour mark. Nevertheless, the incremental approach comes very close on some examples, such as John Carter, for which it is out of sync by under $2s$ towards the one hour mark. This suggests that the approach is more effective for some content, likely due to a combination of better speech detection performance and the use of subtitles which more closely align with the dialogue audio.

5.7 Conclusion

In this chapter we have explored the use of VAD and text data within an audio to text alignment framework that does not require a language model. While the results weren't suitably accurate for a fully automated audio to text alignment solution, we were able to significantly improve on the initial misaligned signals in linear alignment tasks. The methods therefore have potential as part of a semi automated solution, providing a pre-alignment estimate for use within the post-production workflow. Alternatively, this could be used as a pre-processing phase prior to a more comprehensive approach based on lexical discovery. While alignment is not perfect, the more accurately aligned signal components (e.g. the first quarter or so of data) could be used for initial discovery of speech and text associations. This could then be used to build a basic speech recognition model, which could then be applied to find word-level associations to align the remaining data.

The chapter also evaluated VAD and subtitle-based approaches for whole film and scene-level matching of audio and text segments. This demonstrated good performance in both use cases, and suggests that the anchor point segmentation and matching approaches are effective for association, if not for perfectly accurate alignment.

A crucial discovery regarding scene-level matching and alignment was the poor performance of DTW on the finer level features. This was likely due to increased impact of inconsistencies between VAD and text data, due to the summing process having a reduced smoothing effect on the higher resolution features. This was successfully addressed with the development of an alternative, segment-based matching algorithm. This approach demonstrated significant improvement over the DTW approach for the scene-level features in both matching and alignment investigations. Furthermore, this approach went on to demonstrate improved performance over the DTW-based approach for audio-to-video alignment.

The chapter also explored the use of V-VAD in audio-to-video alignment, and demon-

strated promising results when using the anchor-based alignment approach. As with the audio-to-text alignment, this method was not suitably accurate for a wholly automated solution, but very promising results were observed on some data.

V-VAD data was also explored as a means of enhancing audio-to-text alignment, through applying VAD/V-VAD correlation as a means of quantifying anchor point accuracy. This demonstrated some promise, but as with the other V-VAD-based approach, exhibited significant variability across datasets due to inconsistent V-VAD performance. As such, while the use of V-VAD for alignment tasks is promising, performance could likely be enhanced through the following:

1. The use of more sophisticated machine learning algorithms to model visual voice activity could lead to improved V-VAD performance, and thus improve the overall performance of the V-VAD-to-VAD alignment method.

2. The incorporation of a more robust face detection and landmark localisation solution could enhance performance, as reducing missed detections would enhance performance on more difficult data (e.g. dataset 5).

3. While the VAD approach has demonstrated strong performance, it is still affected by missed detections and false detections - thus further VAD enhancement would likely also improve performance. This could be achieved by training the VAD on more data, or by employing genre-specific feature selection, as discussed in Section 3.

In summary, this chapter has explored a number of methods for finding associations between and aligning multimedia content. The most encouraging results were demonstrated by the combined alignment approach, which utilised general alignment as an initial estimate, before refining this through incremental scene-level alignment. Despite these results coming close to the desired alignment, the approach is not accurate enough for use in a wholly-automated alignment solution. This accuracy should be achievable with the integration of a lexical discovery-based approach utilising low resource speech recognition models, for which the proposed approach could serve as a preprocessing phase. Thus, we conclude that the approach is most suitable as a means of automatically segmenting and roughly aligning data within a semi-automated, rather than wholly-automated, solution for subtitle alignment.

Chapter 6

Conclusions and Future Work

This thesis has presented novel work in both audio and visual voice activity detection, and has explored the use of these techniques within a multimedia alignment framework. The work demonstrates that performance improvements can be obtained for both visual and audio Voice Activity Detection (VAD) in challenging speech conditions through engineering noise-robust features. We also show that, while a fully automated alignment solution could not be realised, improvements on initial alignment can be obtained via alignment strategies which do not rely on language models.

In Chapter 3, a novel feature for audio VAD in film multimedia was presented. This was developed to address the problem of speech detection in entertainment audio, which has proven to be a non-trivial task for a number of contemporary and state-of-the-art VAD approaches [36][119]. The proposed solution was a novel set of features - Mel Frequency Cepstral Coefficient (MFCC) Cross Covariance features - which combines a correlation-based preprocessing step with cross-covariance modelling of inter-MFCC relationships to provide a highly discriminative feature vector. Investigations demonstrate that this works successfully with both support vector machines and random forests, with the latter achieving marginally better performance. Furthermore, we demonstrate that this outperforms state-of-the-art VAD methods on multimedia speech detection tasks, achieving significantly greater performance metrics on a commonly used feature-film dataset.

Chapter 4 explored the combination of landmark and appearance-based features for visual speech detection tasks. A number of case studies were presented, ranging from straightforward visual VAD tasks on the Grid corpus [22], to more challenging speech detection scenarios involving natural speaker poses, dynamic gestures and variable illumination conditions. The results demonstrated that, across all investigations, the combination of landmark and appearance-based features yielded performance improvements when compared with using either feature set individually. Crucially, we demonstrated the value

of state-of-the-art landmark localisation techniques, with the landmark-based approach significantly outperforming 2D-DCT features for Visual-VAD (V-VAD) tasks on challenging data. Furthermore, the combined features proposed achieved better results when compared with recent methods for Grid corpus V-VAD tasks.

Lastly, Chapter 5 presents an alignment framework designed with language independence in mind. Here, we demonstrated that rough alignment of audio and text media could be obtained using anchor-point selection and Dynamic Time Warping (DTW). This also proved to be effective for identifying associated sequences of audio and text data, however was not suitable for wholly automated content alignment. This chapter went on to present techniques for scene-level alignment, developing a segment-wise method which outperformed DTW on scene-level association tasks. Visual VAD information was also explored for both audio-to-visual alignment and as a means of enhancing audio-to-text alignment. Results demonstrated that, while the landmark localisation methods were state-of-the-art, they were still not sufficiently robust for use with entertainment multimedia. Finally, an incremental alignment method was presented, which improves on the initial alignment through incorporating incremental scene-level alignment. This produced clear improvements in alignment, but was still not sufficient for use as a wholly automated solution. We conclude that while the proposed alignment framework does not offer a complete solution, it may be useful as a means of initial alignment prior to the application of lexical discovery based techniques. In this way, a language independent approach could be developed through utilising rough audio and text alignment to build a sparse language model, such as described in [120].

6.1 Application Contexts

In this section, we explore several potential applications for the methods developed through this work.

6.1.1 Automatic Content Segmentation

The audio VAD developed through this project could improve multimedia post-production workflows through automatically segmenting content into speech and non-speech. This would be particularly useful for translators working on adaptation, as it would highlight crucial sections of dialogue. This would guide their focus, and would reduce the time required for adaptation as they would not need to make an initial pass over the film to mark up key areas of dialogue.

6.1.2 Subtitle Validation

When developing subtitles for multimedia, the content is typically checked by a number of people in order to ensure that it is of sufficient quality. The audio VAD and segment association/alignment described in this work could improve this process by providing a means of automatically estimating the accuracy of subtitle associations, and the degree to which the subtitles correspond to audio speech activity. While this wouldn't be a wholly automated process, it could provide a confidence measure which could reduce the number of passes necessary for translators and other post-production engineers.

6.1.3 Enhancement of Automatic Transcription Methods

Many methods for automatic transcription do not include a speech/non-speech segmentation phase, and instead apply Automatic Speech Recognition (ASR) directly to the content [16][65][4]. This could produce alignment errors, particularly in audio containing highly variable content as is the case with entertainment multimedia. As such, these processes may benefit from a segmentation phase, to separate speech activity from non-speech content. The idea here would be to improve ASR performance by constraining the input to content that is more likely to correspond to speech, thus reducing the negative impact of noise or sound effect content on speech recognition.

6.1.4 Improving ADR Through AV-VAD

While this work demonstrated that current methods for extracting visual speech features are not sufficient for entertainment media, future improvements to these could results in stronger performance. This could then be harnessed to improve automated Automatic Dialogue Replacement (ADR) solutions through incorporating visual speech information. This could be combined with audio speech data to provide a method for robust audio realignment of ADR recordings. The idea here would be to utilise the audio information where possible (using similar methods to current audio-to-audio alignment [86][79]), but leverage visual speech information in areas where the original recording is too noisy for robust alignment.

6.1.5 Pre-Processing for Language Independent Alignment

As previously mentioned, while the alignment method discussed in Chapter 5 is not suitable for a fully automated solution, it could be utilised as a pre-processing step for a more comprehensive alignment framework. Several existing audio-to-text alignment

methods, such as those proposed by Stan *et al.* [120], have been developed for low-resource applications. In these cases, they require an initial transcription from which to build a sparse dictionary. This is used to build a speech recogniser, which can subsequently be applied for word and sentence-level alignment between the audio and text data. As the alignment strategy here can roughly align content, it is capable of producing reasonable alignments for towards the beginning of content (e.g. the first ten minutes of a feature film). As such, this could automatically provide the initial transcription data required for a more comprehensive lexical discovery-based approach, removing the need for manual transcription alignment.

6.2 Future Work

This work has investigated broad range of disciplines, including computer vision, audio processing and sequence alignment. These have been explored with the goal of developing a method for aligning multimedia content that does not rely on a language model. While some progress was made towards this goal, the method produced is not capable of wholly automated alignment. As such, the key interest for future work is in developing this into a wholly automated solution. This would involve exploring methods for lexical discovery, and investigating their potential for use in multimedia content. Assuming that they can be leveraged for more refined word-level alignment, these could then be integrated to achieve the alignment resolution necessary to provide a comprehensive automated solution for subtitle localisation.

Another key area for further development is visual voice activity detection. While the work here demonstrated encouraging results, the proposed feature set was only tested with one classification algorithm. Given that other methods have proven useful in the literature [2][91][75][121], it would be sensible to investigate V-VAD performance using other machine learning algorithms, particularly more sophisticated approaches such as deep neural network based methods.

There are also several interesting areas for further development of the audio VAD approach. One of these is in exploring whether the correlation-based features can be optimised for subsets of multimedia content (e.g. for genre-specific classification). If this is the case, more robust models can be trained to produce classifiers tailored according to multimedia content. These could then be used within a comprehensive ensemble-based approach, by which incoming audio data is filtered according to which type of sub-content it most closely matches. The data would then be fed to a specially adapted classifier, thus producing more accurate estimates of speech activity.

As with the V-VAD, use of other classification algorithms with the audio VAD could also be explored. Of particular interest here are context-sensitive deep learning approaches, such as RNNs, which may be able to produce enhanced performance due to their capability for modelling sequential data.

Bibliography

[1] Timo Ahonen, Abdenour Hadid, and Matti Pietikainen. Face description with local binary patterns: Application to face recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(12):2037–2041, 2006.

[2] Ibrahim Almajai and Ben Milner. Using audio-visual features for robust voice activity detection in clean and noisy speech. In *Signal Processing Conference, 2008 16th European*, pages 1–5. IEEE, 2008.

[3] Ibrahim Almajai, Ben Milner, and Jonathan Darch. Analysis of correlation between audio and visual speech features for clean audio feature prediction in noise. In *INTERSPEECH*. Citeseer, 2006.

[4] Xavier Anguera, Jordi Luque, and Ciro Gracia. Audio-to-text alignment for speech recognition with very limited resources. In *INTERSPEECH*, pages 1405–1409, 2014.

[5] Bishnu S Atal and Manfred R Schroeder. Adaptive predictive coding of speech signals. *Bell System Technical Journal, The*, 49(8):1973–1986, 1970.

[6] Andrew Aubrey, Bertrand Rivet, Yulia Hicks, Laurent Girin, Jonathon Chambers, and Christian Jutten. Two novel visual voice activity detectors based on appearance models and retinal filtering. In *Signal Processing Conference, 2007 15th European*, pages 2409–2413. IEEE, 2007.

[7] James K Baker. The dragon system–an overview. *Acoustics, speech and signal processing, IEEE transactions on*, 23(1):24–29, 1975.

[8] Tom Barker and Tuomas Virtanen. Non-negative tensor factorisation of modulation spectrograms for monaural sound source separation. In *INTERSPEECH*, pages 827–831, 2013.

[9] I Bazzi and J Glass. Modeling out-of-vocabulary words for robust speech recognition. In *Proc. of ICSLP*, 2000.

[10] Peter N Belhumeur, David W Jacobs, David J Kriegman, and Narendra Kumar. Localizing parts of faces using a consensus of exemplars. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(12):2930–2940, 2013.

[11] Pascal Belin, Robert J Zatorre, Philippe Lafaille, Pierre Ahad, and Bruce Pike. Voice-selective areas in human auditory cortex. *Nature*, 403(6767):309–312, 2000.

[12] Donald J Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA, 1994.

[13] Christopher M Bishop. Pattern recognition and machine learning. pages 340–344, 2006.

[14] Christopher M Bishop. Pattern recognition and machine learning. pages 474–480, 2006.

[15] P Jeffrey Bloom. Use of dynamic programming for automatic synchronization of two similar speech signals. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'84.*, volume 9, pages 69–72. IEEE, 1984.

[16] Norbert Braunschweiler, Mark JF Gales, and Sabine Buchholz. Lightly supervised recognition for automatic alignment of large coherent speech recordings. In *INTERSPEECH*, pages 2222–2225, 2010.

[17] Luca Cappelletta and Naomi Harte. Phoneme-to-viseme mapping for visual speech recognition. In *ICPRAM (2)*, pages 322–329, 2012.

[18] Richard M Chamberlain and John S Bridle. Zip: a dynamic programming algorithm for time-aligning two indefinitely long utterances. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'83.*, volume 8, pages 816–819. IEEE, 1983.

[19] Bingjie Cheng and Shangping Zhong. A novel chicken voice recognition method using the orthogonal matching pursuit algorithm. In *2015 8th International Congress on Image and Signal Processing (CISP)*, pages 1266–1271. IEEE, 2015.

[20] Siew Wen Chin, Kah Phooi Seng, Li-Minn Ang, and King Hann Lim. Improved voice activity detection for speech recognition system. In *Computer Symposium (ICS), 2010 International*, pages 518–523. IEEE, 2010.

[21] William S Cleveland and Susan J Devlin. Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American statistical association*, 83(403):596–610, 1988.

[22] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5):2421–2424, 2006.

[23] Timothy F Cootes, Gareth J Edwards, and Christopher J Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (6):681–685, 2001.

[24] Timothy F Cootes, Christopher J Taylor, David H Cooper, and Jim Graham. Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59, 1995.

[25] Adrian Corduneanu and Christopher M Bishop. Variational bayesian model selection for mixture distributions. In *Artificial intelligence and Statistics*, volume 2001, pages 27–34. Morgan Kaufmann Waltham, MA, 2001.

[26] David Cristinacce and Timothy F Cootes. A comparison of shape constrained facial feature detectors. In *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on*, pages 375–380. IEEE, 2004.

[27] David Cristinacce and Timothy F Cootes. Facial feature detection and tracking with automatic template selection. In *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*, pages 429–434. IEEE, 2006.

[28] David Cristinacce and Timothy F Cootes. Feature detection and tracking with constrained local models. In *BMVC*, volume 2, page 6. Citeseer, 2006.

[29] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[30] Li Deng and Douglas O'Shaughnessy. *Speech processing: a dynamic and optimization-oriented approach*. CRC Press, 2003.

[31] Oscar Déniz, Gloria Bueno, Jesús Salido, and Fernando De la Torre. Face recognition using histograms of oriented gradients. *Pattern Recognition Letters*, 32(12):1598–1603, 2011.

[32] Simon Dixon and Gerhard Widmer. Match: A music alignment tool chest. In *ISMIR*, pages 492–497, 2005.

[33] David Dov, Ronen Talmon, and Israel Cohen. Audio-visual voice activity detection using diffusion maps. *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, 23(4):732–745, 2015.

[34] Peter Elias. Predictive coding–i. *Information Theory, IRE Transactions on*, 1(1):16–24, 1955.

[35] Antti Eronen and Anssi Klapuri. Musical instrument recognition using cepstral coefficients and temporal features. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, volume 2, pages II753–II756. IEEE, 2000.

[36] Florian Eyben, Felix Weninger, Stefano Squartini, and Bjorn Schuller. Real-life voice activity detection with lstm recurrent neural networks and an application to hollywood movies. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 483–487. IEEE, 2013.

[37] File-Extensions.org. Digital video and movie file extension list, 2016.

[38] Theodoros Giannakopoulos, Aggelos Pikrakis, and Sergios Theodoridis. A dimensional approach to emotion recognition of speech from movies. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 65–68. IEEE, 2009.

[39] Toni Giorgino et al. Computing and visualizing dynamic time warping alignments in r: the dtw package. *Journal of statistical Software*, 31(7):1–24, 2009.

[40] James R Glass. A probabilistic framework for segment-based speech recognition. *Computer Speech & Language*, 17(2):137–152, 2003.

[41] Jean-Philippe Goldman. Easyalign: an automatic phonetic alignment tool under praat. 2011.

[42] Ralph Gross, Iain Matthews, and Simon Baker. Generic vs. person specific active appearance models. *Image and Vision Computing*, 23(12):1080–1093, 2005.

[43] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010.

[44] NIST Multimodal Information Group. 2005 nist speaker recognition evaluation training data, 2005.

[45] Leon Gu and Takeo Kanade. A generative shape regularization model for robust face alignment. In *Computer Vision–ECCV 2008*, pages 413–426. Springer, 2008.

[46] Alfred Haar. Zur theorie der orthogonalen funktionensysteme. *Mathematische Annalen*, 69(3):331–371, 1910.

[47] Abdenour Hadid, Matti Pietikäinen, and Timo Ahonen. A discriminative feature space for detecting and recognizing faces. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–797. IEEE, 2004.

[48] Bo Han and Yupin Luo. Accurate face detection by combining multiple classifiers using locally assembled histograms of oriented gradients. In *Audio, Language and Image Processing (ICALIP), 2012 International Conference on*, pages 106–111. IEEE, 2012.

[49] Ahmad BA Hassanat and Sabah Jassim. Visual words for lip-reading. In *SPIE Defense, Security, and Sensing*, pages 77080B–77080B. International Society for Optics and Photonics, 2010.

[50] Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85, 2005.

[51] Timothy J Hazen. Automatic alignment and error correction of human generated transcripts for long speech recordings. In *INTERSPEECH*, volume 2006, pages 1606–1609, 2006.

[52] Hynek Hermansky. Perceptual linear predictive (plp) analysis of speech. *the Journal of the Acoustical Society of America*, 87(4):1738–1752, 1990.

[53] Hynek Hermansky and Nelson Morgan. Rasta processing of speech. *Speech and Audio Processing, IEEE Transactions on*, 2(4):578–589, 1994.

[54] Kevin Hilman, Hyun Wook Park, and Yongmin Kim. Using motion-compensated frame-rate conversion for the correction of 3: 2 pulldown artifacts in video sequences. *Circuits and Systems for Video Technology, IEEE Transactions on*, 10(6):869–877, 2000.

[55] Tomlinson Holman. *Sound for film and television*. Taylor & Francis, 2010.

[56] IMDB.com. Internet movie database, 2016.

[57] Fumitada Itakura. Minimum prediction residual principle applied to speech recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 23(1):67–72, 1975.

[58] Fred Jelinek. Speech recognition by statistical methods. *Proceedings of the IEEE*, 64:532–556, 1976.

[59] Oliver Jesorsky, Klaus J Kirchberg, and Robert W Frischholz. Robust face detection using the hausdorff distance. In *Audio-and video-based biometric person authentication*, pages 90–95. Springer, 2001.

[60] Hongliang Jin, Qingshan Liu, Hanqing Lu, and Xiaofeng Tong. Face detection using improved lbp under bayesian framework. In *Image and Graphics (ICIG'04), Third International Conference on*, pages 306–309. IEEE, 2004.

[61] Hongliang Jin, Qingshan Liu, Xiaoou Tang, and Hanqing Lu. Learning local descriptors for face detection. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pages 928–931. IEEE, 2005.

[62] Michael Jones and Paul Viola. Fast multi-view face detection. *Mitsubishi Electric Research Lab TR-20003-96*, 3:14, 2003.

[63] Michael Kass, Andrew Witkin, and Demetri Terzopoulos. Snakes: Active contour models. *International journal of computer vision*, 1(4):321–331, 1988.

[64] Hiroaki Kato, Minoru Tsuzaki, and Yoshinori Sagisaka. Effects of phoneme class and duration on the acceptability of temporal modifications in speech. *The Journal of the Acoustical Society of America*, 111(1):387–400, 2002.

[65] Athanasios Katsamanis, Matthew Black, Panayiotis G Georgiou, Louis Goldstein, and S Narayanan. Sailalign: Robust long speech-text alignment. In *Proc. of Workshop on New Tools and Methods for Very-Large Scale Phonetics Research*, 2011.

[66] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1867–1874, 2014.

[67] Brian King. Sneak peek - rubbadub, 2011.

[68] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.

[69] Tomi Kinnunen, Evgenia Chernenko, Marko Tuononen, Pasi Fränti, and Haizhou Li. Voice activity detection using mfcc features and support vector machine. In *Int. Conf. on Speech and Computer (SPECOM07), Moscow, Russia*, volume 2, pages 556–561, 2007.

[70] Joachim Koehler, Nelson Morgan, Hynek Hermansky, H Guenter Hirsch, and Grace Tong. Integrating rasta-plp into speech recognition. In *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*, volume 1, pages I–421. IEEE, 1994.

[71] Leona KOPTIKOVA, Jindrich HLADIL, Petr Cejchan, Martin VONDRA, VICH Robert, and Ladislav Slavik. The dynamic time-warping approach to comparison of magnetic-susceptibility logs and application to lower devonian calciturbidites (prague synform, bohemian massif). *Geologica Belgica*, 2010.

[72] Martin Köstinger, Paul Wohlhart, Peter M Roth, and Horst Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 2144–2151. IEEE, 2011.

[73] Yuxuan Lan, Richard Harvey, B Theobald, Eng-Jon Ong, and Richard Bowden. Comparing visual features for lipreading. In *International Conference on Auditory-Visual Speech Processing 2009*, pages 102–106, 2009.

[74] Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, and Thomas S Huang. Interactive facial feature localization. In *Computer Vision–ECCV 2012*, pages 679–692. Springer, 2012.

[75] Thomas Le Cornu and Ben Milner. Voicing classification of visual speech using convolutional neural networks. In *FAAVSP-The 1st Joint Conference on Facial Analysis, Animation and Auditory-Visual Speech Processing*, 2015.

[76] Rainer Lienhart and Jochen Maydt. An extended set of haar-like features for rapid object detection. In *Image Processing. 2002. Proceedings. 2002 International Conference on*, volume 1, pages I–900. IEEE, 2002.

[77] Qingju Liu, Wenwu Wang, and Philip Jackson. A visual voice activity detection method with adaboosting. In *Sensor Signal Processing for Defence (SSPD 2011)*, pages 1–5. IET, 2011.

[78] Beth Logan et al. Mel frequency cepstral coefficients for music modeling. In *ISMIR*, 2000.

[79] Synchro Arts Ltd. Advanced automatic audio alignment, 2016.

[80] Lie Lu, Hong-Jiang Zhang, and Hao Jiang. Content analysis for audio classification and segmentation. *Speech and Audio Processing, IEEE Transactions on*, 10(7):504–516, 2002.

[81] James Lyons. Mel frequency cepstral coefficients, 2016.

[82] James Lyons. Python speech features, 2016.

[83] Dau-Cheng Lyu, Ren-Yuan Lyu, Yuang-Chin Chiang, and Chun-Nan Hsu. Cross-lingual audio-to-text alignment for multimedia content management. *Decision Support Systems*, 45(3):554–566, 2008.

[84] John Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4):561–580, 1975.

[85] John D Markel and AH Jr Gray. *Linear prediction of speech*, volume 12. Springer Science & Business Media, 2013.

[86] David Mellor. Wordfit automatic dialouge synchronization. *Audio Media*, pages 87–90, 1991.

[87] Paul Mermelstein. Distance measures for speech recognition, psychological and instrumental. *Pattern recognition and artificial intelligence*, 116:374–388, 1976.

[88] Kieron Messer, Jiri Matas, Josef Kittler, Juergen Luettin, and Gilbert Maitre. Xm2vtsdb: The extended m2vts database. In *Second international conference on audio and video-based biometric person authentication*, volume 964, pages 965–966. Citeseer, 1999.

[89] Stephen Milborrow and Fred Nicolls. Locating facial features with an extended active shape model. In *Computer Vision–ECCV 2008*, pages 504–513. Springer, 2008.

[90] Todd Mytkowicz, Amer Diwan, Matthias Hauswirth, and Peter F Sweeney. Aligning traces for performance evaluation. In *Parallel and Distributed Processing Symposium, 2006. IPDPS 2006. 20th International*, pages 8–pp. IEEE, 2006.

[91] Rajitha Navarathna, David Dean, Sridha Sridharan, Clinton Fookes, and Patrick Lucey. Visual voice activity detection using frontal versus profile views. In *Digital Image Computing Techniques and Applications (DICTA), 2011 International Conference on*, pages 134–139. IEEE, 2011.

[92] John A Nelder and Roger Mead. A simplex method for function minimization. *The computer journal*, 7(4):308–313, 1965.

[93] Elias Nemer, Rafik Goubran, and Samy Mahmoud. Robust voice activity detection using higher-order statistics in the lpc residual domain. *Speech and Audio Processing, IEEE Transactions on*, 9(3):217–231, 2001.

[94] Kuniaki Noda, Yuki Yamaguchi, Kazuhiro Nakadai, Hiroshi G Okuno, and Tetsuya Ogata. Audio-visual speech recognition using deep learning. *Applied Intelligence*, 42(4):722–737, 2015.

[95] Timo Ojala, Matti Pietikäinen, and Topi Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7):971–987, 2002.

[96] OpenCV. Face detection using haar cascades, 2015.

[97] Alex Park, Timothy J Hazen, and James R Glass. Automatic processing of audio lectures for information retrieval: Vocabulary selection and language modeling. In *ICASSP (1)*, pages 497–500, 2005.

[98] Eric K Patterson, Sabri Gurbuz, Zekeriya Tufekci, and John N Gowdy. Cuave: A new audio-visual database for multimodal human-computer interface research. In *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, volume 2, pages II–2017. IEEE, 2002.

[99] Julien Pinquier, Christine Sénac, and Régine André-Obrecht. Speech and music classification in audio documents. In *IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS SPEECH AND SIGNAL PROCESSING*, volume 4, pages 4164–4164. IEEE; 1999, 2002.

[100] R Venkatesha Prasad, Abhijeet Sangwan, HS Jamadagni, MC Chiranth, Rahul Sah, and Vishal Gaurav. Comparison of voice activity detection algorithms for voip. In *Computers and Communications, 2002. Proceedings. ISCC 2002. Seventh International Symposium on*, pages 530–535. IEEE, 2002.

[101] Recognition PRIMA Perception and Integration for Smart Environments. Talking face video, 2016.

[102] Lawrence R Rabiner, Aaron E Rosenberg, and Stephen E Levinson. Considerations in dynamic time warping algorithms for discrete word recognition. *The Journal of the Acoustical Society of America*, 63(S1):S79–S79, 1978.

[103] Thiruvengatanadhan Ramalingam and P Dhanalakshmi. Speech/music classification using wavelet based feature extraction techniques. *Journal of Computer Science*, 10(1):34–44, 2014.

[104] Sami Romdhani, Shaogang Gong, Ahaogang Psarrou, et al. A multi-view nonlinear active shape model using kernel pca. In *BMVC*, volume 10, pages 483–492, 1999.

[105] Henry A Rowley, Shumeet Baluja, and Takeo Kanade. Neural network-based face detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(1):23–38, 1998.

[106] Henry A Rowley, Shumeet Baluja, and Takeo Kanade. Rotation invariant neural network-based face detection. In *Computer Vision and Pattern Recognition, 1998. Proceedings. 1998 IEEE Computer Society Conference on*, pages 38–44. IEEE, 1998.

[107] Anindya Roy and Sébastien Marcel. Haar local binary pattern feature for fast illumination invariant face detection. In *British Machine Vision Conference 2009*, number LIDIAP-CONF-2009-048, 2009.

[108] Christos Sagonas and Stefanos Zafeiriou. ibug - facial point annotations, 2016.

[109] Kirill Sakhnov, Ekaterina Verteletskaya, and Boris Simak. Approach for energy-based voice detector with adaptive scaling factor. *IAENG International Journal of Computer Science*, 36(4):394, 2009.

[110] Kirill Sakhnov, Ekaterina Verteletskaya, and Boris Simak. Dynamical energy-based speech/silence detector for speech enhancement applications. In *Proceedings of the World Congress on Engineering*, volume 1, page 2. Citeseer, 2009.

[111] Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 26(1):43–49, 1978.

[112] Jason Saragih and Kyle McDonald. Face tracker, 2016.

[113] Jason M Saragih, Simon Lucey, and Jeffrey F Cohn. Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision*, 91(2):200–215, 2011.

[114] Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.

[115] Steven W Smith. *Digital signal processing: a practical guide for engineers and scientists*. Newnes, 2003.

[116] Pieter Soens and Werner Verhelst. Robust temporal alignment of spontaneous and dubbed speech and its application for automatic dialogue replacement. In *Signal Processing Conference, 2010 18th European*, pages 80–84. IEEE, 2010.

[117] Pieter Soens and Werner Verhelst. On split dynamic time warping for robust automatic dialogue replacement. *Signal Processing*, 92(2):439–454, 2012.

[118] Jongseo Sohn, Nam Soo Kim, and Wonyong Sung. A statistical model-based voice activity detection. *Signal Processing Letters, IEEE*, 6(1):1–3, 1999.

[119] Reinhard Sonnleitner, Bernhard Niedermayer, Gerhard Widmer, and Jan Schlüter. A simple and effective spectral feature for speech detection in mixed audio signals. In *Proceedings of the 15th International Conference on Digital Audio Effects (DAFx12)*, 2012.

[120] A Stan, Y Mamiya, J Yamagishi, P Bell, O Watts, RAJ Clark, and S King. Alisa: An automatic lightly supervised speech segmentation and alignment tool. *Computer Speech & Language*, 35:116–133, 2016.

[121] Fei Tao, John HL Hansen, and Carlos Busso. An unsupervised visual-only voice activity detection approach using temporal orofacial features. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[122] Steve Tauroza and Desmond Allison. Speech rates in british english. *Applied linguistics*, 11(1):90–105, 1990.

[123] Kwanchiva Thangthai, Richard Harvey, Stephen Cox, and Barry-John Theobald. Improving lip-reading performance for robust audiovisual speech recognition using dnns. 2015.

[124] Thomas Tran, Soroosh Mariooryad, and Carlos Busso. Audiovisual corpus to analyze whisper speech. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8101–8105. IEEE, 2013.

[125] Andreas Tsiartas, Prasanta Ghosh, Panayiotis G Georgiou, and Shrikanth Narayanan. Bilingual audio-subtitle extraction using automatic segmentation of movie audio. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5624–5627. IEEE, 2011.

[126] George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *Speech and Audio Processing, IEEE transactions on*, 10(5):293–302, 2002.

[127] Werner Verhelst. Automatic post-synchronization of speech utterances. In *EUROSPEECH*, 1997.

[128] Werner Verhelst and Marcel Borger. Intra-speaker transplantation of speech characteristics. an application of waveform vocoding techniques and dtw. 1991.

[129] Werner Verhelst and Marc Roelands. An overlap-add technique based on waveform similarity (wsola) for high quality time-scale modification of speech. In *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*, volume 2, pages 554–557. IEEE, 1993.

[130] L Vieriu. Real-time voice activity detection using a simple webcam. *Proceedings of WCSIT*, 2014.

[131] Ric Viers. *Sound Effects Bible*. Michael Wiese Productions, 2011.

[132] Paul Viola and Michael Jones. Robust real-time object detection. *International Journal of Computer Vision*, 4, 2001.

[133] Yang Wang, Simon Lucey, and Jeffrey F Cohn. Enforcing convexity for improved alignment with constrained local models. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[134] Salah Werda, Walid Mahdi, and Abdelmajid Ben Hamadou. Lip localization and viseme classification for visual speech recognition. *arXiv preprint arXiv:1301.4558*, 2013.

[135] Norbert Wiener. *Extrapolation, interpolation, and smoothing of stationary time series*, volume 2. MIT press Cambridge, MA, 1949.

[136] Norbert Wiener. *DAFX: Digital Audio Effects*. 2011.

[137] Ellen Wixted. Interview with the creator of the automatic speech alignment feature in audition cs6, 2012.

[138] Hongming Zhang and Debin Zhao. Spatial histogram features for face detection in color images. In *Advances in Multimedia Information Processing-PCM 2004*, pages 377–384. Springer, 2004.

[139] Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2879–2886. IEEE, 2012.