Terms and Conditions of Use of Digitised Theses from Trinity College Library Dublin

Copyright statement

All material supplied by Trinity College Library is protected by copyright (under the Copyright and Related Rights Act, 2000 as amended) and other relevant Intellectual Property Rights. By accessing and using a Digitised Thesis from Trinity College Library you acknowledge that all Intellectual Property Rights in any Works supplied are the sole and exclusive property of the copyright and/or other IPR holder. Specific copyright holders may not be explicitly identified. Use of materials from other sources within a thesis should not be construed as a claim over them.

A non-exclusive, non-transferable licence is hereby granted to those using or reproducing, in whole or in part, the material for valid purposes, providing the copyright owners are acknowledged using the normal conventions. Where specific permission to use material is required, this is identified and such permission must be sought from the copyright holder or agency cited.

Liability statement

By using a Digitised Thesis, I accept that Trinity College Dublin bears no legal responsibility for the accuracy, legality or comprehensiveness of materials contained within the thesis, and that Trinity College Dublin accepts no liability for indirect, consequential, or incidental, damages or losses arising from use of the thesis for whatever reason. Information located in a thesis may be subject to specific use constraints, details of which may not be explicitly described. It is the responsibility of potential and actual users to be aware of such constraints and to abide by them. By making use of material from a digitised thesis, you accept these copyright and disclaimer provisions. Where it is brought to the attention of Trinity College Library that there may be a breach of copyright or other restraint, it is the policy to withdraw or take down access to a thesis while the issue is being resolved.

Access Agreement

By using a Digitised Thesis from Trinity College Library you are bound by the following Terms & Conditions. Please read them carefully.

I have read and I understand the following statement: All material supplied via a Digitised Thesis from Trinity College Library is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of a thesis is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form providing the copyright owners are acknowledged using the normal conventions. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone. This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.
Classification and Segmentation Methods with Application in Audio and Acoustic Signal Processing

A thesis submitted to the University of Dublin for the degree of Doctor of Philosophy

Darren Francis Kavanagh
Trinity College Dublin, June 2011
To my family,
In fond memory of my grandfather Jim Cross.
Abstract

This thesis concerns the exploration of novel interactive educational resources to enhance literacy and language learning.

The core system that arises from this research is called "ReciTell", which is a rich media based Virtual Learning Environment (VLE). The technology provides an innovative learning environment through the complementary fusion of visual, auditory and kinaesthetic learning modalities, which is aimed at literacy and language learning. This is achieved by proposing an approach that emulates traditional teaching methods within an advanced technological environment. Creating rich media based educational resources is complex and labour intensive and as a result they tend to be expensive to produce and require a specialized skill-set and media design knowledge.

Considering this problem, this thesis focuses on the research of suitable signal processing methods for automating the media production process. Automation of this process is extremely advantageous as it allows for the streamlined production of interactive educational applications in an efficient and cost effective way. The main body of the thesis addresses the most difficult problem in relation to automating the production, which was the development of an automatic speech segmentation methodology. This speech segmentation system automatically determines the boundary points for the beginning and end of each word in natural voice recordings. This information is an important component for audio-visual synchronisation for interactive educational purposes in the end-applications.

This thesis proposes an automatic speech segmentation method that uses a Dynamic Programming (DP) technique known as Dynamic Time Warping (DTW). While the segmentation algorithm explored using state of the art speech features such as Mel-Frequency Cepstral Coefficients (MFCC) and Perceptual Linear Prediction (PLP) Features, some additional investigation into other feature extraction and speech characterization methods has also been conducted. The segmentation methodology has incorporated Principal Component Analysis (PCA) approaches to perform speech feature transformation and dimensionality reduction. This is favourable as it results in an efficient and dimensionally reduced feature space, $\mathbb{R}^j \rightarrow \mathbb{R}^f$ for $f \leq j$. This effectively acts to remove any information redundancies that are usually present within the features, thus helping to make the algorithm more computationally efficient and robust. The segmentation methodology that is proposed is an unsupervised text-dependent approach, as the text stream for the spoken audio is known a-priori. This type of approach lends itself well to its role within the end-application, i.e. Automatic Content Generation (ACG) system.

The effectiveness of this segmentation method has been validated by appropriate evaluation and analysis of the segmentation results. The results strongly show the method's ability and robustness at automatically segmenting speech, with its performance showing high segmentation accuracy and low segmentation error. The experimentation has clearly demonstrated the
suitability of the method for fulfilling its important role within the ACG system.

The three educational applications that have been explored as part of this research are also reported. These applications include, (i) the core system, the “ReciTell” system, which is in the area of digital technologies for commercial educational applications, (ii) iMARK for user generated content, to greatly facilitate the creation of digital learning objects, and (iii) The Book: Discovering Sounds Initiative (TBDSI), for studies in medieval manuscripts, which allows for greater accessibility to these precious resources and increases the learning outcomes with the provision of sophisticated rich media accompaniments.

In addition to the work described above, another scientific research problem was investigated as part of this thesis. This research was carried out on an industrial application in the area of estimating the Remaining Useful Life (RUL) of rotating machines. The ability to predict the RUL of Rotating Machines is a highly desirable function of Automated Condition Monitoring (ACM) systems. Typically, vibration signals are acquired through contact with the machine and then used for monitoring. In this research, a novel implementation of the ubiquitous feature extraction approach Envelope Analysis (EA) is applied to acoustic noise signals (< 25 kHz) in order to predict the RUL of a rotating machine. This approach is compared to a Data-Driven approach to feature extraction which utilizes an Information Theoretic approach. The results indicate that sufficient information exists in the acoustic signal emitted from a machine, to determine the RUL using both approaches. Suitable comparison of the results for the two approaches is presented.
Declaration

I hereby declare that this thesis has not been submitted as an exercise for a degree at this or any other University and that it is entirely my own work.

I agree that the Library may lend or copy this thesis upon request.
Acknowledgements

Firstly, I owe a particular debt of gratitude to my family, for their support, encouragement, and patience while this manuscript was being prepared. I wish to thank all fellow students and staff at Trinity College Dublin (TCD) for the great experiences shared, friendships gained, and warm sense of camaraderie that prevails in the research laboratories, across campus and college life. To express my gratitude to all of the postgraduate students at the 1937 Reading Room TCD for the great interdisciplinary atmosphere that is present there which has provided an excellent studious climate for the composition of this thesis towards the latter stages.

I want to thank Mr. John Squires and Mr. Liam O'Sullivan for the different projects that I had the pleasure of working together with them on.

A special thanks to Bell Laboratories Alcatel-Lucent for giving me the opportunity to collaborate with industry researchers through a research internship. Many thanks to Dr. Patricia Scanlon, Professor Alan Lyons, Professor Lawrence Cowsar and all of the team at Bell Laboratories.

I wish to thank the generosity of the funding bodies that have supported this work. I am truly grateful for their support. These funding bodies include the Irish Research Council Science Engineering and Technology (IRCSET), National Digital Learning Resources (NDLR), and the Trinity Long Room Hub, Fig. 1.

Lastly, I am profoundly grateful to my supervisor, Professor Francis Boland, for all his help, guidance and encouragement over the course of my postgraduate studies. For the many discussions and white-board sessions that we had along the way, I found the whole experience to be extremely rewarding and has really deepened my knowledge in the field of scientific research and technology driven innovation.

Figure 1: The Funding agencies that supported this work are the Irish Research Council for Science and Engineering (IRCSET) forming part of the Embark Initiative, the National Digital Learning Resources (NDLR), and the Long Room Hub at Trinity College Dublin.
# Contents

<table>
<thead>
<tr>
<th>Contents</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>List of Acronyms</strong></td>
<td>xi</td>
</tr>
<tr>
<td><strong>1 Introduction</strong></td>
<td>1</td>
</tr>
<tr>
<td>1.1 Speech Segmentation</td>
<td>2</td>
</tr>
<tr>
<td>1.1.1 The Research Problem</td>
<td>3</td>
</tr>
<tr>
<td>1.2 Automated Condition Monitoring of Rotating Machines</td>
<td>4</td>
</tr>
<tr>
<td>1.3 Thesis Outline</td>
<td>4</td>
</tr>
<tr>
<td>1.4 Contributions of this thesis</td>
<td>6</td>
</tr>
<tr>
<td>1.5 Publications</td>
<td>7</td>
</tr>
<tr>
<td><strong>2 Acoustic Signal Sources and Models</strong></td>
<td>9</td>
</tr>
<tr>
<td>2.1 Acoustic Signals</td>
<td>9</td>
</tr>
<tr>
<td>2.2 Speech Communication</td>
<td>11</td>
</tr>
<tr>
<td>2.3 Theory of Speech Production</td>
<td>12</td>
</tr>
<tr>
<td>2.3.1 Physiology of Human Speech Production</td>
<td>12</td>
</tr>
<tr>
<td>2.3.2 Manner and Places of Articulation</td>
<td>16</td>
</tr>
<tr>
<td>2.3.3 Prosodies</td>
<td>18</td>
</tr>
<tr>
<td>2.3.4 Co-articulation</td>
<td>18</td>
</tr>
<tr>
<td>2.3.5 Spectrogram</td>
<td>19</td>
</tr>
<tr>
<td>2.3.6 Linear Model</td>
<td>21</td>
</tr>
<tr>
<td>2.3.7 Nonlinear Models</td>
<td>21</td>
</tr>
<tr>
<td>2.4 Synthesized Voice</td>
<td>25</td>
</tr>
<tr>
<td>2.4.1 Background to Speech Synthesis</td>
<td>25</td>
</tr>
<tr>
<td>2.4.2 Modern Speech Synthesis Technology</td>
<td>27</td>
</tr>
<tr>
<td>2.5 Conclusion</td>
<td>28</td>
</tr>
<tr>
<td><strong>3 Automatic Speech Segmentation: Review of Motivation and Methods</strong></td>
<td>29</td>
</tr>
<tr>
<td>3.1 Review of the Literature</td>
<td>29</td>
</tr>
<tr>
<td>3.2 Motivations and Thematic Applications</td>
<td>31</td>
</tr>
</tbody>
</table>
CONTENTS

3.3 Segmentation Methodologies ................................................................. 38
  3.3.1 Supervised and Unsupervised ......................................................... 38
  3.3.2 Text-dependent and Text-independent Approaches ...................... 38
  3.3.3 Top-down and Bottom-up Segmentation ....................................... 39
  3.3.4 Phoneme Segments ......................................................................... 40
  3.3.5 Automated Segmentation Approaches .......................................... 41
  3.3.6 Speech Features for Segmentation ............................................... 44
  3.4 Segmentation Performance Metrics .................................................... 44
  3.5 Conclusion .......................................................................................... 45

4 Speech Analysis and Feature Extraction ................................................. 47
  4.1 Signal Representations: The Philosophy of Features ......................... 48
    4.1.1 Pre-emphasis ................................................................................. 50
  4.2 Short-term Processing of Signals ....................................................... 51
    4.2.1 Short-time Fourier Transform ...................................................... 54
    4.2.2 Energy based Octave Band Analysis ........................................... 57
    4.2.3 Cepstrum ..................................................................................... 59
    4.2.4 Linear Predictive Analysis ............................................................. 66
    4.2.5 Dynamic Features ....................................................................... 70
    4.2.6 Time Domain Features ................................................................. 73
  4.3 Principal Component Analysis ............................................................. 74
    4.3.1 Motivation and Goals behind PCA .............................................. 77
    4.3.2 The PCA Procedure ..................................................................... 78
  4.4 Conclusion .......................................................................................... 82

5 Speech Segmentation Using Dynamic Time Warping ............................... 83
  5.1 Proposed Methodology ........................................................................ 83
    5.1.1 Defining the Problem .................................................................. 86
    5.1.2 Dynamic Time Warping Approach .............................................. 87
    5.1.3 Feature Transformation and Subset Selection ................................ 95
  5.2 Acceptable Tolerances ....................................................................... 97
  5.3 Speech Corpus used for Experimentation ......................................... 100
  5.4 Results .............................................................................................. 102
    5.4.1 Results Format ............................................................................. 102
    5.4.2 Univariate Feature Approach ...................................................... 103
    5.4.3 Baseline Multivariate Feature Approach ..................................... 104
    5.4.4 Dynamic Multivariate Feature Approach .................................... 105
    5.4.5 PCA Dynamic Multivariate Feature Approach ............................ 105
  5.5 Discussion ......................................................................................... 106
List of Acronyms

ACM  Automated Condition Monitoring
ACG  Automatic Content Generator
AM-FM  Amplitude Modulation - Frequency Modulation
ANN  Artificial Neural Network
ASR  Automatic Speech Recognition
BD  Brightness Difference
CC  Complex Cepstrum
CD  Colour Difference
CDF  Characteristic Defect Frequency
CESA  Continuous Energy Separation Algorithm
CSR  Continuous Speech Recognition
DCT  Discrete Cosine Transform
DESA  Discrete-time Energy Separation Algorithm
DFT  Discrete Fourier Transform
DSP  Digital Signal Processing
DTFT  Discrete Time Fourier Transform
DTW  Dynamic Time Warping
EA  Envelope Analysis
EMD  Empirical Mode Decomposition
EPB  Early Printed Books
ESA  Energy Separation Algorithm
ESOL  English for Speakers of Other Languages
FFT  Fast Fourier Transform
FSS  Feature Subset Selection
GLR  Generalized Likelihood Ratio
GMM  Gaussian Mixture Models
HMM  Hidden Markov Model
ICA  Independent Component Analysis
iMARK  interactive Media Annotation Resource Kit
IWR  Isolated Word Recognition
LCR  Luminosity Contrast Ratio
LDA  Linear Discriminant Analysis
LMS  Learning Management System
LPC  Linear Predictive Coefficients
MFCC  Mel-Frequency Cepstral Coefficients
MI  Mutual Information
MIFSS  Mutual Information Feature Subset Selection
NCLB  No Child Left Behind
NDLR  National Digital Learning Resources
NVR  Natural Voice Recording
OBA  Octave Band Analysis
PCA  Principal Component Analysis
PLP  Perceptual Linear Prediction
PSCG  Professional Studio Content Generation
QoE  Quality of Experience
QoS  Quality of Service
LIST OF ACRONYMS

RASTA Relative Spectra
RC Real Cepstrum
REB Rolling Element Bearing
RUL Remaining Useful Lifetime
RMA Rich Media Applications
RWL Reading While Listening
SPL Sound Pressure Level
STFT Short Time Fourier Transform
STFTM Short Time Fourier Transform Magnitude
SVM Support Vector Machine
SVR Synthesized Voice Recording
TBDSI The Book: Discovering Sounds Initiative
TEO Teagers’ Energy Operator
TIMIT Texas Instruments and Massachusetts Institute of Technology
TTS Text To Speech
UGC User Generated Content
VAD Voice Activity Detection
VE Voting Experts
VLE Virtual Learning Environment
VTN Vocal Tract Normalization
WPM Words Per Minute
ZCR Zero Crossing Rate
LIST OF ACRONYMS
Advances in computer processing power and capabilities over the last 20 years fused with the greater availability and access to high speed Internet connections, have led to the widespread introduction of various computer software packages, tools and web-based resources for different applications. These application software programs are developed for some certain purpose, which usually involves enabling the end-user to perform some particular task or function. There are various application software programs available at present which cover areas such as: word processing, spreadsheet tools, presentation programs, business software, simulation software, industrial automation, and advanced communication services mainly via the Internet. Another important area which has received a lot of attention and coverage is educational application software. There have been many programs developed for teaching and self-learning using computers, which cover various subjects and topics targeted towards different age groups. In the area of children's learning most of the programs have been developed towards literacy, numeracy and reference such as children's interactive encyclopedias.

A particular area of interest to this research has involved the exploration of a novel interactive educational program as a valuable resource to assist with teaching literacy and language learning. This system is called ReciTell\textsuperscript{1} which is an interactive learning environment. The technology provides an innovative learning environment through the complementary fusion of visual, auditory and kinaesthetic learning modalities. This system provides synchronized media

\textsuperscript{1}The origin of the name ReciTell has stemmed from the creation of a portmanteau word, which is an effective blend of the two words \textit{Recite} and \textit{Tell}. This name encapsulates the role of the technology and its applications well.
in the form of the spoken audio and the visual text highlighting of words to create a more enriching learning environment for teaching children how to read\(^2\). This is achieved using an approach that emulates traditional methods within an advanced technological environment. This has been rendered in an interactive virtual book format which operates within a rich media Flash environment.

The creation of these interactive educational software applications can be extremely labour intensive and one that requires a specialized skill-set and media-design knowledge. As a direct result of this, these educational applications can prove to be quite complex and expensive to produce in particular for large scale product development, including aspects surrounding content localization for different regions.

In light of this, the research encompassed within this thesis has explored techniques for automating the media production process to enable the streamlined development of these high-quality educational applications in an efficient and cost effective manner. In order to achieve this some important scientific problems and technical challenges needed to be researched and addressed in order to make the proposed automation technology realizable. These problems comprised of: (i) automatic speech segmentation, (ii) timing synchronization and animation effects, and (iii) pedagogic design considerations and issues. The most challenging and difficult problem out of all three of these was the automation of the speech segmentation process. Consequently, a large part of this thesis has been focused towards proposing suitable methodologies and techniques for automatic speech segmentation of natural voice recordings into linguistically meaningful units, that of \textit{words}.

\subsection{1.1 Speech Segmentation}

A general definition for segmentation usually refers to breaking something down into constituent parts that are of interest. Specifically, here the interest lies in breaking the signal down into segmental units that are relevant to the end-application, which in this case is for educational applications. The different approaches to speech segmentation tend to be heavily dependent upon the requirements of the segments’ role within their given end-application. Speech signals are unlike printed text for instance, wherein the words are separated by spaces\(^3\); in the case of spoken audio there is not any reliable acoustic delineation between subsequent words\(^4\) [45]. This may contrast with the immediate impression that we have when listening to speech as our mind automatically performs this task without any real conscious effort. The mind breaks the

\footnote{Also there are important aspects of this which are suitable for language learning. However since literacy was looked at in greater detail, the discussion is therefore focused towards this.}

\footnote{This is not the case for some languages where white spaces do not appear between the words, examples of this being Chinese and Japanese.}

\footnote{Aside: automatic alignment methods are being investigated to align printed text with handwritten manuscripts [338, 160, 258]. This image processing problem is effectively the \textit{spatial-image} counterpart to the \textit{temporal-speech} problem being addressed herein.}
1.1. Speech Segmentation

speech signal down into linguistically and psychologically meaningful units such as words and morphemes in order to assess meaning [49]. In continuous fluent speech the sounds of phonemes, syllables and words flow into one another. This is a result of many complex factors, in particular co-articulation effects as well as fluency, pronunciation, speaking rate and other prosodic factors such as intonation and stress.

The speech segmentation units of interest to this research are the boundary points for the beginning and end of each word in the audio stream. The diagram in Fig. 1.1 illustrates where these units lie within the greater hierarchy of segmenting a spoken audio document.

1.1.1 The Research Problem

The problem can be stated quite concisely, as it involves taking a speech signal $s(n)$ and determining $X$ segment boundaries across the signal. These boundary points $\zeta_x$ comprise of the beginning and end points, denoted $b_x$ and $e_x$ for the $x^{th}$ word; see Eq. (1.1).

$$\zeta_x = (b_1, e_1, b_2, e_2, \ldots, b_x, e_x, \ldots, b_{X-1}, e_{X-1}, b_X, e_X)$$ (1.1)

There are many forms of constraints that can be considered when addressing this segmentation problem. The problem can be text-dependent or text-independent, which depends on whether the text transcript is known to the system or not. The approach can be supervised or unsupervised, the former requires extensive training, which is similar to a modern HMM based speech recognition system, some examples of these being commercial systems such as Microsoft’s Vista and Windows 7, or Nuance’s Dragon Naturally Speaking. Some non-commercial examples are Hidden Markov Model Toolkit\(^5\) and CMU SPHINX\(^6\). This is contrary to what is required here however, as being mindful of the segmentation system’s role and usefulness in the end application, the system itself should not require training and therefore should be an unsupervised method. This requires that the segmentation method itself be sophisticated and robust enough to be speaker independent.

\(^5\)HTK is a portable toolkit for speech recognition. Current version 3.4.1.
\(^6\)CMU SPHINX is an open source toolkit for speech recognition.
1.2 Automated Condition Monitoring of Rotating Machines

In addition to the interactive educational software applications and the speech segmentation problem, another area has been investigated as part of this research. This was on an industrial application in the area of Automated Condition Monitoring. The problem involved automatic estimation of the Remaining Useful Life (RUL) of rotating machines. This was conducted during a research internship at Bell Laboratories Ireland. While the application area is very different when compared to speech processing, there are two fundamental similarities between the two, which are: (i) both problems involve automating a process, and (ii) both are operating on acoustic signals in the range 20-20kHz, acquired by an acoustic transducer device (microphone).

1.3 Thesis Outline

The remainder of this thesis is organized as follows.

Chapter 2: Acoustic Signal Sources and Models

This Chapter describes some of the background theory concerning speech science. Different aspects of speech are considered such as physiological, lexical and acoustical. Prosodic and co-articulation effects are also presented. This is followed by a discussion on modelling speech production. Before the Chapter concludes a different class of speech signal is considered, that of synthesized speech. Synthesized speech has an important function within the proposed segmentation methodology as a reference signal with the DTW procedure.

Chapter 3: Automatic Speech Segmentation: Review of Motivation and Methods

A review of the motivation and methods in the area of automatic speech segmentation is provided in this Chapter. The topic is given a broad introduction by discussing both human cognitive and artificial algorithmic forms of speech segmentation. The literature is then reviewed under the areas of: (i) thematic applications, (ii) segmentation methodologies and feature sets, and (iii) the evaluation measures used to assess segmentation performance. Different aspects relating to speech segmentation approaches are discussed in the methodologies section.

Chapter 4: Speech Analysis and Feature Extraction

This Chapter explores different feature extraction methodologies for characterizing the speech signal for the process of speech segmentation. A discussion is provided on the philosophy of speech features in relation to their important role within speech processing. The Chapter examines different methods of feature extraction and signal analysis techniques. Particular attention has been given towards the more predominant state of the art approaches from the literature. In
addition to these other novel methods of feature extraction for speech segmentation have been explored. The method of Principal Component Analysis (PCA) is described for the purpose of transformation and dimensionality reduction of the features vectors.

Chapter 5: Speech Segmentation Using Dynamic Time Warping

This Chapter reports on the proposed segmentation methodology that has been investigated. This research has explored a methodology that uses Dynamic Programming (DP) principles, more specifically the technique known as Dynamic Time Warping (DTW). The proposed algorithm is described in detail incorporating the various features and signal analysis techniques that have been described previously in Chapter 4. Important areas such as segmentation tolerances are discussed in relation to considerations of the end-applications, as well as the speech corpus that has been used and other experimentation details. The Chapter concludes with a detailed analysis of results for the proposed methodology. These results strongly show the effectiveness of the system as an automatic means of segmenting continuous speech signals at the word-level boundaries.

Chapter 6: Applications: Integration and Implementation

This Chapter discusses some important background theory which has supported the research and development of the technology within the Automated Content Generation System (ACG), as well as the educational applications. Some of these pertinent areas reported upon include: eye movement characteristics, appropriate reading rates, some cognitive aspects of synchronization in multimedia resources followed with a discussion on the hypotheses of learning styles. Leading on from this, a discussion is provided on how the proposed technology fits within the ACG system and the application areas that have been pursued. The three main application areas that have been explored are: (a) ReciTell (b) iMARK and (c) The Book: Discovering Sounds Initiative (TBDSI). These application areas form important contributions to this research. They are discussed under four thematic areas: (i) Background information, (ii) Problems addressed, (iii) Proposed solution, and (iv) Outcomes.

Chapter 7: Automated Condition Monitoring of Rotating Machines

This Chapter describes a different problem that has been considered in the area of Automated Condition Monitoring (ACM). The ability to predict the Remaining Useful Life (RUL) of Rotating Machines is a highly desirable function of ACM systems. Typically, vibration signals are acquired through contact with the machine and then used for monitoring. In this Chapter, a novel implementation of the ubiquitous feature extraction approach Envelope Analysis (EA) is applied to acoustic noise signals (< 25kHz) in order to predict the RUL of a rotating machine. A well known drawback of the EA approach is that the frequency band of interest must be known or pre-estimated. As a result of this, the approach is compared to a Data-Driven approach to
feature extraction which utilizes an Information Theoretic approach to feature selection that does not require any \textit{a-priori} information regarding the frequency band of interest. It is shown that the Data-Driven approach, with an accuracy of 97.7\%, significantly outperforms the EA approach, with an accuracy of 93.7\%. The Chapter shows that the improved performance of the Data-Driven approach is due to new information being uncovered in spectral locations across the entire spectrum from 0 to 25 kHz, and not just within one frequency band typically used by the EA approach.

**Chapter 8: Conclusions**

The final chapter discusses the contributions of this thesis and outlines some directions for future work.

**1.4 Contributions of this thesis**

This thesis has proposed some important contributions, five of which are listed here below:

1. The research has proposed a novel automatic speech processing approach for the segmentation of natural voice recordings at the word level boundaries. This has been integrated into the ACG system. This is an unsupervised text-dependent methodology, which solves a major problem within the technology for automating the production process. This proposed segmentation method incorporates a novel implementation of Principal Component Analysis (PCA) for feature transformation and dimensionality reduction for speech features. The results of the proposed segmentation system show its effectiveness and its ability for fulfillment of its important role within the ACG system.

2. The thesis has explored some state of the art speech features for incorporation into the speech segmentation system. In addition to these features other feature extraction techniques have been proposed for the speech segmentation task. These features have achieved very good performance in comparison to the more predominant approaches in the literature.

3. The research has led to the introduction of a novel automation system to enable the streamlined production of rich media based interactive educational applications and resources.

4. The development of a new innovative learning environment called ReciTell has been explored as part of this research. This technological educational environment provides an engaging interactive learning experience for the pupil through the fusion of visual, auditory and kinaesthetic learning modalities. Additional applications which utilized some or all of this technology have been investigated, including: iMARK and The Book: Discovering Sounds Initiative.
5. An industrial research problem has also been addressed, in the area of automatically estimating the RUL of rotating machines. This work has proposed a novel implementation of the Envelope Analysis technique for this RUL estimation problem. The methods proposed to address this problem are shown to achieve very good performance.

1.5 Publications

Portions of the work described in this thesis have appeared in the following documents and publications:


*Bell Laboratories Alcatel Lucent, Ireland.
†Trinity College Dublin, Ireland.


Trinity College Dublin, Ireland.


*Bell Laboratories Alcatel-Lucent, Ireland.
†Trinity College Dublin, Ireland.


*BITS, Pilani - Hyderabad Campus, India.
†Trinity College Dublin, Ireland.
This Chapter describes some of the background theory that is important for understanding the theory behind speech processing in general, and indeed for the methods proposed and discussed herein. Furthermore, it provides greater knowledge in relation to the nature of the problem that has been addressed. The Chapter begins by introducing some properties of acoustic signals. The topic of discussion moves towards a particular class of acoustic signal, that of speech signals. An overview is given for the different parts of human anatomy that are responsible for speech communication. After this a more detailed look at the theory of speech production is given. The different aspects of speech are considered from both physiological aspects to lexical and acoustic viewpoints also. Spectrogram signal plots are provided for a sample speech utterance of both synthetic and natural speech, for comparison purposes. Prosodic characteristics of speech signals are described along with co-articulation effects. Later in the Chapter, some discussion on modelling speech production is given for both linear and non-linear models. Finally, speech synthesis is discussed, firstly from a historical viewpoint of the technological advances and progress in the area, to secondly giving a description of the particular type of modern speech synthesis that was of interest to this research.

2.1 Acoustic Signals

Firstly, this section begins by defining what is meant by the term sound. A sound is deemed to have occurred where a disturbance propagates through an elastic medium, which causes an alteration in pressure or a displacement of the particles of the medium, which can be detected
aurally by a person or an acoustical instrument [37, 27]. The particular propagating medium of interest to this work is air, which is a gas, and the frequency range of signals concerned is the human audible hearing range, which is in the range 20 Hz to 20kHz [325, 37].

Consider for instance a sinusoidal vibrating wall in gas (air) as illustrated in Fig. 2.1. This vibration accelerates adjacent air particles and compresses that part of the gas nearest to it as it progresses forward [27]. During the forward-going part of the cycle the particles have a forward momentum gained from the wall. These particles in turn collide with their neighbouring particles to the right and through this successive chain of collisions they transfer this forward momentum. Similarly, when the wall reverses direction for the negative-going part of the cycle, rarefaction occurs, which causes particles to be accelerated in the reverse direction [27].

There are three interesting aspects of the Fig. 2.1 that are worth considering, (a) variation in pressure (b) particle displacement and (c) particle velocity. Notice the orthogonal relationship (⊥) that exists such that, when the particle displacement is at a maximum the incremental pressure is zero and, vice versa, when the pressure is at a maximum the displacement is zero, as there is a 90 degree phase difference between them. This relationship is presented in,

\[
\begin{align*}
\text{Pressure} & \quad P(t) = \cos \omega t \\
\text{Particle displacement} & \quad s(t) = \sin \omega t \\
\text{Particle velocity} & \quad \frac{ds}{dt} = \cos \omega t
\end{align*}
\]  

(2.1a) (2.1b) (2.1c)
2.2. Speech Communication

Since particle velocity is the derivative of displacement, it is therefore in phase with pressure. The speed of sound is not instantaneous, it is finite, and is governed by the following equation,

$$\lambda = \frac{c}{f}$$

(2.2)

where \(\lambda\) corresponds to the wavelength, and \(c\) corresponds to the speed of propagation through the medium. Here the discussion is limited to air for which the speed \(c\) is approximately 340 metres/sec \([218, 71]\) and \(f\) denotes frequency in Hertz (\(\equiv\) cycles per second). In relation to the direction of propagation consider the following waves and media as shown in Table 2.1 as discussed in [27]. The vocal tract in speech production, can effectively be thought of as being a lossless tube for modelling purposes. Planar propagation occurs in the vocal tract for sound waves for frequencies roughly \(< 4\text{kHz}\). The propagation only occurs for frequencies whose wavelength is greater than the vocal tract diameter (typically \(\approx 2\text{cm}\)). Therefore, for the case of 4kHz, substituting this into Eq. (2.2) yields:

$$\lambda = \frac{340\text{m/sec}}{4000\text{cycles per second}} = 8.5\text{cm}$$

(2.3)

Since this is much greater than 2cm, the planar wave propagation assumption is valid. This topic of speech propagation through the vocal tract will be discussed again later in this Chapter under models of speech production.

2.2 Speech Communication

First it is important to give an overview of the basic elements in human communication. Human speech essentially allows us to communicate ideas with one another. The communication usually comprises of words and phrases that form sentences, all of which form part of a language. There are various parts of human anatomy that play a part in speech communication. These parts of anatomy are illustrated in Fig. 2.2. The female on the right is speaking, so she is using her brain for motor control and thought processing as well as the organs used in the production of speech sounds; these will be discussed below. The female on the left is listening to what is being

<table>
<thead>
<tr>
<th>Wave type</th>
<th>Medium</th>
<th>Direction</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sound waves</td>
<td>Air</td>
<td>Longitudinal</td>
<td>The waves have the same direction of vibration along their direction of travel.</td>
</tr>
<tr>
<td>Light, Heat, Radio</td>
<td>Free Space</td>
<td>Transverse</td>
<td>The vibrations of electric and magnetic fields are (\perp) to the direction of the wave.</td>
</tr>
<tr>
<td>Vibration Liquids</td>
<td>Water</td>
<td>Circular</td>
<td>The vibration occurs in circles/ellipses but the wave travels horizontally.</td>
</tr>
</tbody>
</table>

Table 2.1: Overview of different types of waves, mediums and directions of propagation.
said and is using her auditory mechanisms in order to hear the speech. The sound enters into the ear, where it is transformed and then analysed by the brain to interpret the meaning and understanding of the message. The figure also shows that facial expressions and gestures play an important role in communication.

2.3 Theory of Speech Production

2.3.1 Physiology of Human Speech Production

In order to gain a better understanding of speech signals it is important to know more about the physiology of human speech production. The main organs responsible for speech production are the lungs, larynx, pharynx, nasal cavity, and the mouth, including the tongue and the lips, as depicted in the cross-sectional diagram in Fig. 2.3, [218, 71, 225]. The lungs provide the necessary source of energy through muscular expulsion of air through the trachea (wind pipe) and into the vocal tract. The air flow gets modulated through various stages in the speech production process which produces the acoustic power in the audio range. The organs that are passed through en route towards the exit point at the mouth (lips) and the nose (nostrils) are responsible for the resonant sound.

These organs modify the sound accordingly [120]. Acoustic resonances within the vocal
system are significant for certain properties of speech sounds. The air flow that passes up from the lungs through the trachea gets modulated by the vibration of the vocal folds [218,71,225]. The vocal folds consist of two folds of tissue stretched across the opening of the larynx. The diagrams illustrate this in Fig. 2.4. The folds are both connected at one end to the thyroid cartilage and at the other end connected to two arytenoid cartilages [233]. The two arytenoids are controlled by muscular movements. They can move far apart to form an opening which has a triangular shape. This is the normal position during breathing. They can also be brought tight together, which is the normal condition when a person holds their breath or during swallowing, thus preventing food from entering the trachea and lungs. These two extreme cases of the vocal folds being abducted and adducted are presented in Fig. 2.4. Also of particular interest is the fact that the vocal folds can also be moved into the position where they are almost touching which effectively forms a narrow slit or opening which is referred to as the glottis [120,243,218,133].

As air from the lungs is delivered through the trachea it is forced through the glottis; the vocal folds then vibrate and this vibration modulates the airflow passing through. This modulation action in the glottis is referred to as phonation. As shown in Fig. 2.5a notice that the point of closure is more abrupt and sharp than the opening phase which is more rounded. This is significant as it means that after the glottis vibration has reached sufficient amplitude, (i.e. after one or two cycles), during the closing cycle, the vocal folds actually make contact thus constricting the airflow completely. This is observed on the glottal flow waveform in Fig. 2.5a with the almost flat regions after the sharp abrupt closing. This abrupt stopping of airflow is significant as in the acoustic signal it contributes to greater energy content in the upper portion

---

1 Vocal folds are sometimes referred to as Vocal Chords.
Acoustic Signal Sources and Models

of the harmonic spectrum [120]. The reasons for this abrupt closing is related to two forces, an elastic restoring force of the vocal fold and a Bernoulli force operating on the glottis [218,233,71].

During phonation this time between openings and closings of the vocal folds is referred to as the fundamental period denoted $T_0$, and the rate of vibration is referred to as the fundamental frequency or pitch denoted $F_0$, where $F_0 = \frac{1}{T_0}$. This frequency is dependent upon size and tension of the vocal folds. As the average size of men's vocal folds tends to be much larger, they therefore yield a longer duration $T_0$, and hence a smaller pitch $F_0$. This is the reason a male's voice has a lower pitch than that of a female's voice. For adult males, their pitch typically lies in the range 50-200Hz and females are usually one octave higher. Each person has their own associated level at which their pitch rests at on average [71]. Because of the associated differences in geometries involved for males and females within the vocal tract the resonance frequencies will also vary considerably as well.

The power spectrum for glottal (voicing) signal is given in Fig. 2.5b. This shows harmonic components at integer multiples of fundamental frequency $F_0$, usually denoted $H_1, H_2, \ldots$. The spectrum for this signal has a roll-off of $-12$ dB/octave. The effects of radiation in the mouth causes a significant gain, which reduces the roll-off to $-6$ dB/octave. This is indicated with a broken line in diagram. The effect of the vocal tract is to act as a filter on this voicing signal as indicated for the schwa vowel /a/ in Fig. 2.5c. This shows the filtering effect that it has on the harmonic components in the spectrum. The vocal tract has a different frequency response for the different sounds.

As a side note, the relationship between harmonics is sometimes considered in voice quality parameters. For instance the difference in magnitude between the first harmonic ($H_1$) and the second harmonic ($H_2$) is known to be highly correlated with an open quotient parameter (OQ)
Figure 2.5: In (a) the time domain plot of the glottal area and glottal flow, and (b) the harmonic spectrum of the glottal flow signal, showing a -12dB/octave roll-off. The broken line shows a -6dB/octave roll-off, this is due to the added radiation effects in the mouth region, and (c) the vocal tract filter response is given on the left and shows the affect that this has on the harmonic spectrum on the right.

An interesting study conducted by Hanson and Chuang [107] suggests that the differences between the first harmonic \(H_1\) and the third harmonic \(H_3\), is 9.6 dB lower for males than for females, which indicates that spectral tilt can be useful for differentiating for speech [125].
2.3.2 Manner and Places of Articulation

From a linguistics point of view, speech can be broadly broken down into vowels and consonants. This can be further classified into other categories [225]. Alternatively, from an acoustic point of view speech can be broken down into categories of voiced, unvoiced, and silence. Consonants of the two can be more easily defined using the characteristics (i) Place of articulation, (ii) Manner of articulation, and (iii) Voicing [218,71,225]. These characteristics are as follows:

Place of articulation is defined as being the location or point where principle constriction of vocal tract occurs [218,225]. Table 2.2 gives a detailed list of the various places of articulation [55], these positions in the vocal production system are indicated in Fig. 2.6. Manner of articulation defines the airflow and the degree of constriction in the vocal tract [218,225]. Voicing refers to whether or not phonation is present during articulation. Accordingly, the consonants that have phonation present are referred to as voiced and the others are termed unvoiced. The resonance mechanism for voiced speech is the vocal tract, starting at the larynx and passing through the pharynx and into the mouth. Some voiced sounds also pass through the nasal cavity as well. The frequencies of these associated resonances and their frequency variation with respect to time are important for determining the speech message. They are referred to as formants and are termed $F_1$, $F_2$, $F_3$, ..., . After sampling there are only usually 3-5 formants present in the Nyquist frequency band. The resonant properties can be viewed or modeled as a filter, as illustrated in Fig. 2.5c for the vowel /a/. This will be mentioned again later in the Chapter, when describing linear speech models.

The different categories of articulation [120,233,71,225] are presented below:
2.3. Theory of Speech Production

<table>
<thead>
<tr>
<th>No.</th>
<th>Place</th>
<th>Detail</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Exo-labial</td>
<td>Outer part of lip</td>
</tr>
<tr>
<td>2.</td>
<td>Endo-labial</td>
<td>Inner part of lip</td>
</tr>
<tr>
<td>3.</td>
<td>Dental</td>
<td>Teeth</td>
</tr>
<tr>
<td>4.</td>
<td>Alveolar</td>
<td>Front part of alveolar ridge</td>
</tr>
<tr>
<td>5.</td>
<td>Post-alveolar</td>
<td>Rear part of alveolar ridge and slightly behind it</td>
</tr>
<tr>
<td>6.</td>
<td>Pre-palatal</td>
<td>Front part of hard palate that arches upward</td>
</tr>
<tr>
<td>7.</td>
<td>Palatal</td>
<td>Hard palate</td>
</tr>
<tr>
<td>8.</td>
<td>Velar</td>
<td>Soft palate</td>
</tr>
<tr>
<td>9.</td>
<td>Uvular</td>
<td>A.k.a. Post-velar; uvula</td>
</tr>
<tr>
<td>10.</td>
<td>Pharyngeal</td>
<td>Pharyngeal wall</td>
</tr>
<tr>
<td>11.</td>
<td>Glottal</td>
<td>A.k.a. Laryngeal; vocal folds</td>
</tr>
<tr>
<td>12.</td>
<td>Epiglottal</td>
<td>Epiglottis</td>
</tr>
<tr>
<td>13.</td>
<td>Radical</td>
<td>Tongue root</td>
</tr>
<tr>
<td>14.</td>
<td>Postero-dorsal</td>
<td>Back of tongue body</td>
</tr>
<tr>
<td>15.</td>
<td>Antero-dorsal</td>
<td>Front of tongue body</td>
</tr>
<tr>
<td>16.</td>
<td>Laminal</td>
<td>Tongue blade</td>
</tr>
<tr>
<td>17.</td>
<td>Apical</td>
<td>Apex or tongue tip</td>
</tr>
<tr>
<td>18.</td>
<td>Sub-laminal</td>
<td>A.k.a. Sub-apical; underside of tongue</td>
</tr>
</tbody>
</table>

Table 2.2: Places of articulation [55] corresponding to the numbers as indicated in Fig. 2.6a.

1. **Plosive**, sometimes referred to as *stops*. These sounds are produced by a build up in pressure behind a total constriction in the vocal tract which is followed by a sudden sharp release.

2. **Affricate** sounds are produced by the transition from a stop to a fricative. Thus they are characterized by an initial closure of the vocal tract with a slow release giving rise to turbulence.

3. **Fricative** sounds are formed by a steady air stream that becomes turbulent at a point of articulation.

4. **Lateral** sounds are produced when the vocal tract is closed at a point but remains open either side.

5. **Nasal** sounds are formed when the vocal tract is closed at a point of articulation where the nasal cavity is open (velum) and the oral cavity is closed.

6. **Trill** sounds occur during periods of vibrations between the articulators and the place of articulation.
7. **Semi-vowels**, these types of sounds occur when the vocal tract is partially open without turbulence. They can be further classified as *liquids*, e.g. /w/ in wet or *glides*, e.g. /r/ in ran. Glides differ from vowels in that they are more transitional as they maintain the target position for much less duration. Liquids are similar to vowels. However they tend to be weaker in magnitude due to the increased constriction in the vocal tract.

8. **Vowels** are not quite as easily defined as consonants as there is no specific place of articulation due to the limited amount of constriction present in the vocal tract during their production. Vowels usually tend to be described in terms of the following: (i) Tongue high or low, (ii) Tongue front or back, (iii) Lip positions and (iv) Nasalization or unnasalized. The diagrams in Fig. 2.6 illustrate this. In Fig. 2.6b the IPA vowel chart is given and Fig. 2.6c shows the tongue positions for the front cardinal vowels, /i/, /e/, /ɛ/, and /a/.

The frequency spectrum for the *schwa* vowel /ə/ is given in Fig. 2.5c.

### 2.3.3 Prosodies

Prosodic aspects of speech include timing, pitch and rhythm. These all contribute to linguistic structure in a language. These features usually span duration longer than phonemes and are thus referred to as *suprasegmental* [71]. They are created through subtle variations or manipulation of the source and the vocal articulators during the production of phonemes in speech. Some notable prosodic cues in a language are **Stress** and **Intonation**. Stress refers to a change in the frequency of the pitch and loudness to signify particular emphasis on different segment syllables, words or phrases [71,225]. Intonation is related to the pitches contour over time. It performs many functions in a language, most notably to signify grammatical structure such as sentence markings, clauses and other boundaries [71].

### 2.3.4 Co-articulation

There are various aspects of variability in the speech signal (inter-speech differences) such as gender, accents, and prosodies. Another important area to consider is the effects relating to co-articulation which will be described now. The *phoneme* can be thought of as being the smallest unit of speech for *ideal* sounds [120,233,71]. The most recent international phonetic alphabet (IPA) 2005, according to the International Phonetic Association, is given in the Appendix Fig. 9.3. This is categorized into various groups and sub-groups of vowels, consonants, suprasegmentals and diacritics. It is important to note that in a language, people do not speak in robotic manner switching between (40-50) ideal phoneme sounds. A phoneme can be thought of as a class of speech sounds that conveys the same meaning. The actual sound that is produced, taking into account the variation as described, is referred to as a *phone* [71].

As described, the production of speech involves movements of the articulators in the glottal source in order to generate the speech signal. Unlike printed text, which has distinct characters, speech signals contain phonemes that overlap in time. At first one might be tempted to think of
the transitions between phonemes as a somewhat contrived mechanical rotary\(^2\) switch that can
instantaneously switch between phonemes. Speech production however is unlike this because
the articulation movements take a certain amount of time to move in and out of position,
as the articulators effectively comprise of tissue, muscles, and cartilages. For this and other
reasons smooth transitions between phonemes usually occur, which is very different to the ideal
mechanical switch concept. Deller et al. \cite{71} describe co-articulation succinctly as being: “the
term used to refer to the change in phoneme, articulation and acoustics caused by the influence of
another sound in the same utterance.” A good example of where co-articulation causes dramatic
changes in the pronunciation of sounds, is for instance the word “stand,” as the last consonant
it is pronounced differently in the contexts of “stand down”, “stand back”, or “stand close” \cite{69}.
Some factors that influence co-articulation are the articulatory positions which are only target
positions and are not always reached as a certain degree of freedom exists during the production
of speech \cite{233,71}. Also important from motor theory is the program for the production of the
sequence of sounds for speech units. The terminology in speech theory refers to left-to-right and
right-to-left co-articulation.

1. **Left to right** is in regard to high-level motor control anticipatory mechanisms in the brain
that have a form of forward planning strategy (anticipatory), where they move articulators
into positions early for the next phoneme.

2. **Right to left** is in regard to low-level articulatory momentum causing some of the previous
phoneme to drift into the subsequent phoneme \cite{120,71,225}.

The topic discussed here in relation to prosodies and co-articulation is what made the manual
operation of early day mechanical speech synthesizers difficult to master. Moreover, it still
remains one of the major associated difficulties of modern day concatenative based synthesizers,
i.e. modelling co-articulation effects between phonemes and for longer segments such as syllables.

The prosodic aspects and co-articulation effects, as aforementioned, are important in relation
to the segmentation algorithm as they will have strong impacts on the feature representations
used to characterize the speech utterances. The topic of feature extraction will be covered in
detail in Chapter 4.

### 2.3.5 Spectrogram

The *Spectrogram* is a popular technique used in speech analysis and processing \cite{120,233,71,
225}. Spectrograms involve computing what is known as Short-time Fourier Transform (STFT)
Analysis over intervals of time with/without frame overlap, incorporating the magnitude content
only\(^3\). In Fig. 2.7 two examples of spectrograms are presented for the same utterance for

\(^2\)The concept of a rotary switch with 1 pole \(N_p\)-way, where \(N_p\) represents the number of phonemes to switch
between.

\(^3\)Sometimes referred to as the STFTM \cite{337,101}.
Figure 2.7: Spectrograms of (a) the Synthetic Voice Recording (SVR) and (b) the Natural Voice Recording (NVR) for the sentence: “Heave on those ropes the boat’s come unstuck”. The word boundaries for the two signals are also given which are indicated with vertical bars.

synthesized and natural speech. The spectrogram contains the amplitude information only and therefore neglects the phase information. While some purists would argue the significance of phase information, in general it is not considered to be a limitation of spectrograms as the vast majority of speech analysis does not regard the phase to be that important [71]. Speech from an acoustic point of view can be broken down into categories of voiced, unvoiced, and silence. These different signal types are present in the spectrogram signal that is given in Fig. 2.7. Voiced speech is indicated with the harmonic energy patterns in lower frequency range, e.g. mid way through the word /heave/ to mid way into the word /those/, unvoiced speech is indicated with bursts of noise like energy in the upper portion of the spectrum, e.g. towards the end of the word /those/ and mid way through the word /ropes/. Silence and pause regions are indicated by periods of very low energy, e.g. at the start and end of the two utterances, and at the end of the word /ropes/ for the synthesized signal.
2.3. Theory of Speech Production

2.3.6 Linear Model

A widely adopted simplified model for speech production is known as the source-filter model [133, 71, 273]. The source for voiced speech is modelled as an impulse train at the rate $T_0$. The source for unvoiced speech is modelled as Gaussian noise. An example of this simplified speech model, is given in Fig. 2.8. The model is in a composite or lumped form which incorporates the effects of the glottal excitation, vocal tract and radiation by a time-varying digital filter [71, 246, 273]. This is given by the all-pole $z$-transform as defined by,

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{2.4}$$

The parameters that define this transfer function $H(z)$ are: the switch between voiced and unvoiced excitation, the pitch period, the gain parameter $G$ and the filter coefficients $a_k$. The voiced signal is represented by an impulse train and unvoiced speech is modelled as a random noise signal, as this signal in many respects resembles white noise [218, 71]. The signal $s(n)$ can be described by a difference equation in relation to $u(n)$ and is given by,

$$s(n) = \sum_{k=1}^{p} a_k s(n - k) + Gu(n) \tag{2.5}$$

This difference equation is useful and will be used again in Chapter 4, in the section describing LPC analysis.

2.3.7 Nonlinear Models

The production and propagation of sound adheres to laws the of physics, such as the conservation of mass, momentum, and energy. The propagation medium, air, follows the laws of thermodynamics and fluid mechanics [243]. While the last section has described a simplified model for
speech production based upon conventional linear theory for planar sound propagation, here in this section some consideration is given to discuss some models and hypotheses for aeroacoustic flow within the vocal tract.

The partial derivative of pressure with respect to direction $x$ has been shown [243] to be given by,

$$- \frac{\partial p}{\partial x} = \rho \frac{dv}{dt} \tag{2.6}$$

The particle velocity $v(x, t)$ is a function of space $x$ and time $t$; the derivative\(^4\) of particle velocity gives the true acceleration of the air particles as,

$$\frac{dv}{dt} = \frac{\partial v}{\partial t} + v \frac{\partial v}{\partial x} \tag{2.7}$$

\(^4\)Using the Chain Rule for partial derivatives [158].
2.3. Theory of Speech Production

Substituting this into Eq. (2.6) gives us:

\[- \frac{\partial p}{\partial x} = \rho \left( \frac{\partial v}{\partial t} + v \frac{\partial v}{\partial x} \right)\]  

(2.8)

This is non-linear as the fluid particle velocity \( v \) multiples the term \( \frac{\partial v}{\partial x} \). This can be linearly approximated as:

\[- \frac{\partial p}{\partial x} = \rho \left( \frac{\partial v}{\partial t} \right)\]  

(2.9)

When there are only small velocity changes due to local compressions and rarefactions, this approximation is relatively accurate\(^5\), such that the term \( \rho v \frac{\partial v}{\partial x} \) is small compared with the term \( \rho \frac{\partial v}{\partial t} \) [243]. However, when the air is moving as a fast sheet-like jet flow, as depicted in Fig. 2.9 and discussed below on the topic of jet flow patterns, this corrective term grows in significance and the linear approximation becomes less accurate. The complete solution for such a non-linear fluid dynamics based model involves the solution of Navier Stokes equations [305]. Whether or not this could be accurate or computationally feasible for a time varying speech model is not yet known. Currently it has been solved for a stationary phoneme [71].

Teagers' work on modelling vowel sounds [306, 305, 304] using hot wire anemometry\(^6\) found unique flow patterns to occur along the vocal tract. Their relatively sophisticated HWA experimental set-up enabled them to measure normal flows as well as radial and axial flows. Important findings came from this original work; in particular, that the air flow in the vocal tract and in the mouth is separated and not isotropic. Effectively, what this means is that instead of the flow being stable and uniform across any cross-section as depicted in the linear model on the left of Fig. 2.9, the Teagers observed unique nonlinear flow patterns throughout the vocal tract and the mouth, notably the patterns termed: Whistle, Smoke Ring Vortices, and Flapping, which are considered below. These findings have been discussed in the literature [243, 71]. The simplified resonants models considered earlier in this Chapter to model the vocal tract have some associated limitations. Teagers' observations for definite three-dimensional tract shapes show flow patterns of jet streams and vortices which have particular instabilities unique to the different speech sounds. Consider again the vocal production mechanism from the vocal cords to the exit at the mouth, where all of the physiological parts (lungs, larynx, pharynx, nose, mouth, teeth, and lips) constitute a coupled system. From our previous discussion on resonances and formants it is fair to say that the parts in this coupled system have associated resonances. Teagers' work extends on this further, such that different parts may also be oscillatory [305].

\(^5\)In relation to the equations given in Eqs. (2.6) to (2.9), for a complete derivation of the wave-equation the reader is referred to the texts of [243, 256].

\(^6\)Hot-Wire Anemometry (HWA) is a relatively sophisticated method often employed for detailed study and examination of turbulent air flows, where rapid velocity fluctuation can occur. The principle of operation is such that a fine (delicate) piece of wire typically tungsten is heated up, as the air flowing past the wire has cooling effect on the wire, and as the resistance of the wire is dependent upon the temperature (similar to a thermistor) therefore a relationship is obtained between the resistance and the flow velocity [46, 305].
1. **Whistle** jet flow pattern is where a jet stream has a sheetlike laminar flow along the side of the wall and re-attaches at the false vocal folds see Fig. 2.9. This causes vortices to be entrapped in the cavities. This particular flow is referred to as the Whistle as strong similarities exist with the jet flow in a policeman’s style whistle [243]. The similarities exist as the jet velocity is modulated by the vortex which expands and contracts within the entrapped cavity.

2. **Smoke Ring Vortices** flow or (shedding) is another observed pattern that occurs when the jet flow has a fast enough velocity as illustrated in Fig. 2.10. This fast moving jet stream causes vortices to be shed downstream between the jet stream and the vocal tract wall.

3. **Flapping** across the cavity walls. This pattern is where the jet flow switches back and forth at approximately the first formant frequency, between the tongue and the palate. This flow pattern can also occur in the vocal tract as well.

All three of these flow patterns strongly indicate that the nonlinear jet flows give rise to additional acoustic sources throughout the vocal tract via the conversion of non-acoustic flow into an acoustic propagation [243]. This whole area is quite complex and not yet fully understood. Teagers’ argument is against a particular attitude where physical effects can only occur for which elegant linear equations are available to describe them. This attitude they have claimed is hampering scientific progress [305].

Furthermore, additional studies have been carried out on the topic from an aero-acoustics perspective by MaGowan [194] and Hirschberg [117], which give further evidence of non-acoustic velocity components giving rise to acoustic sound sources due to interactions with vocal tract boundaries at regions of discontinuities. Also studies by Liljencrants [172] and Thomas [308] on the two-dimensional numerical simulations of Navier Stokes equations, describing fluid flow, suggest that further down the vocal tract after the glottal exit, shedding vortices can exist as...
2.4 Synthesized Voice

discussed in point two *Smoke Ring Vortices* above. Moreover, Maragos et. al. [189] motivated by Teagers' work, have used an AM-FM model to model *speech resonants* or *cavity resonators* which loosely refer to oscillator systems formed by the vocal tract cavities. This model separates the signal into amplitude and frequency components by proposing what is known as an Energy Separation Algorithm (ESA) for both Continuous and Discrete-time, termed (CESA) and (DESA) respectively. The algorithm is particularly noteworthy as it is derived from what is known as Teagers’ Energy Operator [138] (TEO). The TEO is a high-resolution (instantaneous) energy estimator and is considered later on in the text in Chapter 4 in the area of feature extraction.

2.4 Synthesized Voice

2.4.1 Background to Speech Synthesis

Up until this point in the text only human speech has been considered. Another form of speech that needs to be looked at is that of synthesized speech. Speech synthesis refers to speech signals that are generated by a machine. These synthetic speech signals from this point on shall be referred to in the text as Synthetic Voice Recordings (SVR) and human speech recordings shall be referred to as Natural Voice Recordings (NVR). It is important to begin this topic by giving some background to events and key milestones in the theory of speech synthesis. The concept of machines being able to speak has been of interest to humans for some time. To quote Euler “It would be a considerable invention indeed, that of a machine able to mimic speech, with its sounds and articulations. I think it is not impossible.” Leonhard Euler (1761) [80]. The motivation for speech synthesis has evolved with time; some motivational factors are as follows:

1. Novelty: in general the idea of a machine being able to talk seemed quite advanced and futuristic. This goes as far back to Kempelen’s *Acoustic-Mechanical Speech Machine* [318], see Fig. 2.11b.

2. Speech analysis via synthesis: to better understand the human articulatory system.

3. Practical applications, reading machines for the visually impaired, navigation systems, answering machines and information services.

In order to gain a deeper understanding of speech synthesis and the underlying challenges, it is worthwhile to take a look at aspects and events that cover the history of this technology. In 1779 in St Petersburg (Russia), Kratzenstein studied different aspects of vowels sounds: /a/, /e/, /i/, /o/, and /u/, using this knowledge he constructed *Acoustic Resonators* (tubes) to model the human vocal tract, see Fig. 2.11a. The resonators were activated using air to vibrate the reeds. Later in 1791 in Vienna (Austria), Kempelen introduced his *Acoustic-Mechanical Speech Machine* [318], see Fig. 2.11b. This device was advanced for its time as

---

7In a similar fashion to some wind instruments that include the clarinet and saxophone families.
Figure 2.11: Important developments in speech synthesis, (a) Kratzenstein’s *Resonant Tubes* 1779, (b) Kempelen’s *Acoustic-Mechanical Speech Machine* 1791, (c) Wheatstone’s Reconstruction of Kempelen’s Machine 1835, (d) Faber’s *Marvelous Talking Machine* 1845, (e) Dudley’s Voder 1939, and (f) Modern Speech Synthesis Software using AT&T voice engine.

it had the ability to produce single speech sounds as well as some combinations of speech sounds. The device consisted of a pressure chamber similar to bellows\(^8\) to model the lungs and a vibrating reed with a leather tube to model the vocal folds and vocal tract respectively. The operation was as follows: the different vowel sounds were produced via the appropriate manipulation of the leather tube by the operators hand. Consonants were produced using four separate passages to cause a constriction controlled by the operator’s fingers. Plosive sounds were generated by exploiting mechanisms of a hinged tongue and movable lips. Kempelen’s model was significant as it led to a greater understanding of speech production. Before Kempelen’s machine it was generally thought that the larynx was the main source of acoustic articulation, whereas the theory underpinning this machine suggested that the cavity that exists between

\(^8\) A device for blowing air into a fire, the operation involves squeezing two handles to force a stream of air from an airbag.
2.4. Synthesized Voice

the vocal folds and the lips is the main source of acoustic articulation. In 1835 in Dublin (Ireland), Wheatstone demonstrated his reconstruction of Kempelen’s machine at the meeting of the British Association for the Advancement of Sciences [71], see Fig. 2.11c. Alexander Graham Bell later viewed the machine and was greatly inspired by it, so much so that he began building his own machine with his brother Melville [43,71]. This work led to his famous U.S. Patent, the Voice Telephone in 1876 [23]. In 1845 in Philadelphia (United States), Faber exhibited\(^5\) his Marvelous talking machine, see Fig. 2.11d. Faber’s machine was operated by controlling 16 levers, similar to that of a piano. Each key could produce one of 16 elementary sounds, an additional 17th key was required for opening and closing the glottis. The machine effectively modeled the human organs of speech production where levers and strings replaced (mimic) muscles and tendons [257]. In more recent times, research into the area of mechanical synthesis has diminished considerably. The 20th century saw advances in the area of electrical speech synthesis. Dudley at Bell Laboratories produced the first electrical connected speech synthesizers called the Voder [77], this was launched at the World Fair in 1939, see Fig. 2.11e. Later in 1954 Cooley et al. [66] developed the pattern play-back machine, which read stylized spectrograms that were hand-painted onto a moving transparent sheet.

The vast majority of modern day synthesizers are software based, see Fig. 2.11f. However, some purists still argue the validity of the mechanical speech synthesis approach. A good example of contemporary work in this area is the Waesada Talker [90]. Hopefully, this discussion into the technological advances in the area of speech synthesis has gone some way towards reinforcing the sheer complexity and various challenges that have shrouded speech synthesis over the years and highlights the great strides and advances that have been made in the area.

2.4.2 Modern Speech Synthesis Technology

Speech synthesis is now commonly referred to as Text-to-speech (TTS) synthesis. A TTS system generally takes an incoming portion of text and converts it into intelligible and naturally sounding speech. This signal shall be referred to as a Synthesized Voice Recording (SVR). State of the art TTS systems differ quite considerably from the rather crude cut and paste approaches used by some Telecommunication applications that string isolated words and numbers together in some automated voice answering systems. State of the art TTS systems have reached a point now where they are sophisticated enough to incorporate acoustic representations, linguistic analysis and prosody in context [275]. Modern day speech synthesis can be categorized into four different areas: (i) Mechanical Synthesis, (ii) Formant Synthesis, (iii) Articulatory Synthesis by Rule, and (iv) Concatenative Synthesis.

The speech synthesizer that was used extensively in this research was AT&T Natural Voice\(^{TM}\) TTS, which is a concatenative synthesis technology, which is referred to as unit selection. The

\(^5\)Also launched again the following year in London’s Egyptian Hall, which drew endorsement from the Duke of Wellington [173].
basic principle behind concatenative speech synthesis is that short segments of recorded speech
are concatenated together to form the synthetic signal [275, 120]. These short segments are
stored in an inventory of speech sounds and are called upon as required. In earlier concatenative
synthesis systems, units of di-phones\textsuperscript{10} were used extensively. However the use of these units
on their own has some limitations with respect to modelling co-articulation effects only on a
phoneme level. Co-articulation effects can last longer than a phoneme duration. For this reason
larger units are also incorporated such as demi-syllables units (half syllable in duration). In
the case of unit selection the units are based on multiple linguistic levels [275, 120, 80]. The
procedure is as follows: the systems receives a text string and converts this into a string of
phonetic symbols. Then using a set of rules along with a pronunciation dictionary target values
for the pitch ($F_0$), phoneme durations and amplitude levels are estimated. The next stage
assembles the unit segments of speech as defined by the target values set in the preceding stage.
These segments are acquired from a large inventory of stored speech sounds. In the case of
unit selection the choice of speech sounds to choose from is relatively large. The many different
units to choose from are characterized by differences in terms of pitch, duration, and amplitude,
which is a consequence of the context of the linguistic message present when they were spoken.
The most appropriate one is chosen based on the target values set out. These units are then
concatenated together in the final stage with minor modifications to form natural sounding
speech.

2.5 Conclusion

This Chapter has taken a look at some of the theory surrounding speech science. Firstly, some
general properties of acoustic signals were described. Then a closer look at theoretical aspects
of speech production was provided. These were described in terms of the physiology of the
human speech production mechanism, as well as lexical and acoustic properties surrounding
the different speech sounds. The topic of prosodic features and co-articulation effects was also
discussed. Before the Chapter concluded a different type of speech signal was considered, that
of \textit{synthesized speech}. This type of speech signal was also of interest to this work, as it formed
an important part in the proposed speech segmentation algorithm. This was first presented in
terms of the advances in the area of speech synthesis and then moved on to give an overview of
a modern concatenative speech synthesis method, that was used in this research. The particular
type that was used was \textit{unit selection concatenative} speech synthesis generated by a modern
AT&T\textsuperscript{TM} speech engine.

\textsuperscript{10}A di-phone is a segment of speech from the centre of one phoneme to the centre of the next phoneme. As the
centre of phones tend to be the most stable region, therefore these points are more appropriate segment points
for rejoining afterwards than the phoneme boundaries [275].
This Chapter provides a review of motivation and methods in the area of automatic speech segmentation. Firstly, a broad introduction on the topic of speech segmentation is given in terms of human cognitive and artificial algorithmic processes. A review of the literature is then provided under the areas of: (i) thematic applications, (ii) segmentation methodologies and feature sets, and (iii) the evaluation measures used to assess segmentation performance. In the literature review section under methodologies, different topics in relation to segmentation approaches have been discussed.

3.1 Review of the Literature

The definition of the term speech segmentation is the process of determining the boundaries between words, and sub-word units (phonemes), in spoken natural languages. It is important to appreciate that the term applies to both the cognitive segmentation processes used in the brain [69, 280, 62, 49, 262, 260, 208], that is from a perception and psychophysics viewpoint, and to the artificial algorithmic segmentation processes of natural language speech processing from pattern recognition and machine learning [104, 204, 129, 184, 310, 32, 72, 230]. While the latter forms the main focus of this work, valuable hypotheses, models and insights can be gained by careful consideration of the former. Broadly speaking, it is often the case that both biological and
organizational examples of intelligence can provide inspiration for building intelligent systems [322, 236]. In terms of the cognitive segmentation, this is effectively regarded as the process of dividing up the continuous speech stream into linguistically and psychologically significant units in order to access the meaning from the intended message [49]. Studies by Saffran et al. [261, 260], have shown that both adults and 8 month old infants have the ability to segment continuous spoken speech into words. The studies used an artificial language generated by a speech synthesizer comprising of words where there were no audible breaks between the words and where each syllable was spoken with the same emphasis.

These studies were interesting in that they found that 8 month old infants can segment fluent speech based solely on the relationship between adjacent sounds [260]. Their results support a particular hypothesis that final syllable lengthening facilitates word learning and speech segmentation [262]. Other additional experiments by Hauser et al. [110] on Cotton-top Tamarins produced similar results which suggest that perhaps this ability could in fact be innate [115, 204]. Being able to formally characterize this ability would be extremely advantageous for both cognition based theoretical purposes and for the suitable development of methods and models for machine learning [203]. A related aspect to all of this was discussed by O’Shaughnessy [218], stating that people use rhythmic sound sequences where the amplitude regularly rises and falls to perform perceptual segmentation and it is these transitional periods which effectively allow speech to be perceived at rates of 40 to 50 phonemes per second [97] although 10 to 12 phonemes per sec [218, 249], is a more typical rate [218]. The typical duration of phonemes

---

1 Ant colonies are a good example of this where their collective intelligence as a colony can act more intelligently than that of an individual ant.

2 Also known as the Pinche Tamarin, is a small New World monkey found in tropical forests.
3.2 Motivations and Thematic Applications

usually lie in the min-max range of 30 to 500 milli-seconds [186]. The duration of words can also vary considerably which depends heavily on the content and application, for example in audio books 150 to 160 Words Per Minute (WPM) is commonly used which gives an approximate average word duration in the region of 2.5 seconds (neglecting silences and pauses). This has been found to be an appropriate rate as it satisfies two important criteria: (i) the rate at which the narrator can comfortably vocalize the words and (ii) the rate at which the listener can adequately hear and comprehend the text [326] being spoken aloud. This topic, with regard to speech rates, will be discussed again in greater detail in Chapter 6 where more emphasis will be placed on the duration of word highlighting and eye fixations for the purpose of the development of instructional learning resources.

The problem formulation of speech segmentation within a pattern recognition framework can be considered as the most general setting of interpretation [316]. The diagram in Fig. 3.1 encapsulates, at a high-level, some of the core elements involved in the process of speech segmentation. Clearly, the diagram shows that the system has a certain amount of knowledge. This knowledge must be deductively supplied in an appropriate form while other knowledge components may be inductively acquired through learning [316]. Note the feedback branch from the output of the system to the learning block, this effectively indicates that some systems can incorporate some form of corrective feedback or further refinement of initial segmentation estimates; examples of this are commonly found in the literature in [223,5,156,192,310,72,278] and will subsequently be discussed later in this Chapter.

This section reviews the literature for speech segmentation in terms of the state of the art methodological approaches and techniques along with the critical points of current knowledge, understanding and discourse. A rich literature exists covering the broad topic of speech segmentation with many different proposed features, methodologies, motivations and applications. Accordingly, the literature has been discussed in terms of three important areas, (i) thematic applications, (ii) the segmentation methodologies and features sets, and (iii) the performance and evaluation metrics that are used in practice, see Fig. 3.2.

3.2 Motivations and Thematic Applications

Speech segmentation is seldom an end unto itself, that is of course apart from the case of pure perceptual modelling for theoretical purposes. Rather it usually tends to form a valuable pre-processing stage to some subsequent process belonging to a larger system. There are various applications present in the literature, of which there are three prominent areas where segmentation acts as an important pre-processor. These are (i) concatenative corpus based speech synthesis [129,310,184,79,230], (ii) training in modern speech recognition systems [184,310,72], and (iii) variable speech coding using Voice Activity Detection [157,288,221,299]. The particular application explored in this research is its role within the context of digital technologies for producing educational resources and instructional materials. The specifications of the segmentation
depend heavily on the nature and requirements of the given application. In general the methodologies and hypotheses explored in the literature tend to be heavily based around solutions that are customized towards meeting the requirements and expectations of the end-application or model. Much of the early segmentation approaches reported in the literature were based around end-point detection for the purpose of speech recognition (isolated word recognition) [135,71,162] and speech coding, Voice Activity Detection (VAD) preprocessing stages [71,165,221,299]. For early speech recognition systems, that worked on an isolated word basis, it was important to be able to accurately determine the onset and offset of a word prior to the template matching and decision rule stages. Similarly speech coding systems should have accurate estimates of voiced, unvoiced and silence regions in the speech signal for efficient coding of signals. The area is generally referred to as Voice Activity Detection VAD [165,221,299]. More recently the literature on speech segmentation has been dominated by approaches for phonetic (sub-word) segmentation. The two major motivations for much of this research are corpora-based speech synthesizers (TTS) and large vocabulary continuous speech recognition (LVCSR) systems. Both of these two applications require accurate segmentation typically at the phoneme level. Such systems have a strong desire to have large corpora of accurately segmented and labelled (transcribed) speech signals. Some tools have been proposed and made available specifically for the purpose of manual segmentation, an example of this is the Transcriber tool [20] by French Délegation Générale pour l'Armement (DGA). More recently, an audio editor XTrans was launched at the Inter-speech conference in 2009 by the Linguistic Data Consortium LDC, which is a specifically designed editing tool for the manual transcription and annotation of audio recordings [94]. Manual segmentation at the phonetic boundaries is considered to be the most accurate compared to automated means [310,192] and many tools and resources are available to perform manual segmentation. It still, however, remains a rather time consuming, tedious and arduous task and one that is prone to human error over large data sets [242,59,310,185,192]. For large scale corpus

---

Audio software editing suites include: Adobe Audition, Sony Sound-forge and Audacity.
3.2. Motivations and Thematic Applications

development it can prove to be extremely costly and one which usually requires the specialist expertise of phoneticians to perform the segmentation and labeling (transcription) accurately. Automation of this procedure offers the opportunity to deliver a fast and cost effective means of segmentation of speech signals requiring little human control or intervention. Furthermore, in some cases the approaches need automatic systems because it is not possible or practically feasible to perform the task manually.

In relation to this research a strong motivation exists for the automation of the speech segmentation process to form part of an automated content generation system. For this application there are two main motivational factors which are as follows. Firstly, with regard to the Professional Studio based Content Generation (PSCG) system it is important in relation to increasing the quality of the audio recording and in order to minimize the time spent in the recording studio. The reason why this is the case is that any mistakes and timing synchronization issues will be picked up immediately, i.e. within minutes after making the recording, thus avoiding the need to make additional retakes on subsequent days which often results in noticeable changes in the voice of the actors which effectively reduces the Quality of Service (QoS) of the audio narration (composition) for the end-user. Secondly, with regard to the role of the technology within the User Generated Content (UGC) tool for use within next generation Virtual Learning Environments (VLEs). In regards to the UGC tool, the automation of both the segmentation and synchronization of the media content is really the only viable option as the alternative would be to have the recordings sent away to be segmented and synchronized at some hub or content media development centre/hub. Although this is possible it would not be practical or cost effective and would be logistically difficult to implement in practice. Thus there is a clear requirement to have a robust and easy to use automated means of segmentation. These areas in relation to automated content generation for both PSCG and UCG will be further expanded upon and discussed in Chapter 6.

There are many motivations for automating the process for speech segmentation, some of the more notable areas have already been alluded to in the text thus far. Some of the major applications and motivations are discussed below:

Speech Recognition

Speech recognition is the process of analyzing spoken information to identify the linguistic message\(^4\) or content and act appropriately upon that information [307], [71]. Effectively the process can be thought of as a mapping from a continuous time signal to a sequence of discrete entities, phonemes, words and sentences [182]. Broadly, speech recognition can be divided up into two main types termed Isolated Word Recognition (IWR) and Continuous Speech Recognition (CSR). Both of these are discussed as well as highlighting the requirements imposed on the

---

\(^4\)It is important to stress linguistic message here as sometimes the terms speech recognition and voice recognition are incorrectly interchanged. The two tasks in fact are very different as voice recognition tries to identify a particular individual's voice and therefore it belongs to a class of biometric based problems.
segmentation pre-processing stage.

**Isolated Word Recognition (IWR)**

Early speech recognition systems were of the Isolated Word Recognition (IWR) kind, [71,135,71,162,244]. This type of recognition requires segmentation of the word boundaries for the template matching stage which was usually Dynamic Time Warped (DTW) based. This type of segmentation was usually referred to as end-point detection, some examples include [135,269,162,245], more recently, [331,171,281]. A lot of emphasis was placed on making algorithms and systems robust against background noise and accordingly in this regard a lot of robust approaches have been reported in the literature [135,162]. These IWR systems were typically of a limited vocabulary due to complexity, processing and memory capabilities associated with storing and processing templates of feature profiles for each word. However, despite the small vocabulary limitation many of these systems have been successfully employed in systems where they can improve the efficiency of entering information into a machine. These systems are particularly well suited to applications where the vocabulary and message content can be heavily constrained or restricted. Some examples of this include manufacturing environments (sorting tasks), assisting people with disabilities, surgery where hands are not available to control equipment, in dark environments, cockpits, optimization of phone dialing and information exchange services.

**Continuous Speech Recognition (CSR)**

Much of the recent research efforts have focused on the ambitious and challenging task of Continuous Speech Recognition (CSR) for large vocabularies. Sometimes these systems are referred to as Large Vocabulary Continuous Speech Recognition (LVCSR) [238] or Automatic Speech Recognition (ASR) systems [205]. While the early IWR used template matching framework employing dynamic programming principles, for the more modern speech recognition systems the predominant methodology is Hidden Markov Model (HMM), which is sometimes used in conjunction with other models forming hybrid systems, e.g. HMM/GMM, HMM/ANN, HMM/SVM. These systems differ considerably in their segmentation requirements, as for the IWR systems it was sufficient to determine the word boundaries, whereas HMM approaches require segmentation and transcription of the speech signal at the phonetic level. This is an essential requirement for the training of the models in the ASR systems. A good degree of research effort has been reported in the literature specifically targeting this problem of phonetic segmentation for ASR training. There is motivation for automatic segmentation machines in this area, being that the greater the accuracy of the segments coupled with the benefits gained of automation provides for more efficient development of training data for training the ASR models. In the past manual transcription of speech signals has been used which proves to be extremely time consuming and laborious and thus imposes a major limitation. There are many examples present in the literature of speech segmentation approaches which are specifically aimed towards improving the performance of speech recognition systems. Some of these include: [242,310,187,19,185], and
3.2. Motivations and Thematic Applications

for early Isolated Word Recognition systems: [331, 171, 281, 135, 269, 162, 245].

Concatenated Speech Synthesis (Unit Selection)

Another important application which many speech segmentation approaches are focused towards is speech synthesis, and more specifically concatenated speech synthesis which uses short segments of speech sounds from an inventory to generate the speech. This topic has been mentioned already in Chapter 2. In order to improve speech synthesis systems it is necessary to have large phonetically and prosodically labeled speech databases [185, 310, 192]. A major limitation of modern speech synthesizers for naturalness is based around prosody generation problems. The majority of the techniques for prosody generation are based around having large prosodically and phonetically labeled corpora available [185]. For Unit Selection based synthesis, they require having a selection of di-phone sounds to choose from based on lexical context [275]. As speech synthesis systems are all the time growing more and more sophisticated, the size and quality of these corpora are important for the performance of such systems. There are many segmentation systems in literature aimed towards speech synthesis [185, 310, 192].

Psychology and Cognitive Modeling

An active area of research, for some time now, is that of modelling speech segmentation in order to gain a better understanding of how humans perform the segmentation task. In fluent speech there are many words that are not marked by clearly delineated acoustic cues or acoustic gaps, such as brief pause periods [132, 49, 208]. A study mentioned in [208] estimates that only 40% of word boundaries in English contain a stop consonant or fricative. Segmentation requires a listener to divide the continuous speech signal into linguistically and psychologically meaningful units [132, 49]. Cairns likens this inter-relationship between the two processes of recognition and segmentation as constituting that of a chicken or egg causality dilemma. Stating that segmentation in speech into meaningful units requires recognition of the meaningful units, but so too recognition requires segmentation in the first place. There is much theory and discourse present in the literature on this topic from many stances and viewpoints, including literature from psychological modelling [69, 280, 132, 62, 49, 262, 260, 208, 287] and from artificial intelligence models [204, 203, 42, 21, 62, 61, 286]. This topic will be mentioned again in the section of top-down or bottom-up segmentation topologies later in this Chapter.

Speech Coding using VAD

Speech classification and segmentation is important for the purpose of speech coding algorithms which perform Voice Activity Detection. Some VAD systems wish to segment the signals

The chicken or the egg causality dilemma refers to the philosophical question: “which came first, the chicken or the egg?”
into regions of Speech/Non-speech/Silence and then into Voiced/Unvoiced and Stationary/Non-stationary noise:

1. (a) Speech further into (i) Voiced (ii) Unvoiced.

2. (b) Non Speech, further into (i) Stationary Noise and (ii) Non-stationary Noise.

VAD systems from the literature include: [157,288,221,299].

Content Indexing and Retrieval

Spoken document processing is a broad application area, however many of the problems and challenges that are encountered are quite similar. Although many advances have been made in the area of speech and language processing, there are still many underlying technological challenges that encumber further progress in the area of spoken document processing [220]. In order for applications to fully embrace spoken documents instead of its written text counterpart, significant advances need to be made in the area. These particular problems surround major fundamental differences that characterize the two media, such as that for spoken languages they are not delineated with explicit punctuation and formatting that is present in text documents. Hence one of the major challenges for spoken document processing is speech segmentation for various levels of segmental granularity: word-level [285,72], phrase-level/sentence-level [176,155,296], dialog-level [154,104], story-level boundary detection [255,254], topic and subtopic-levels segmentation [127]. The problem is compounded by the variations across different genres; for example conversational speech will have dis-fluencies and discourse markers [220] for managing turns taken by different participant speakers that form part of the interaction (dialog). This differs considerably from news broadcast style which tends to be read monologue from a script. A good review of this topic, on aspects relating to segmentation for spoken document processing, is presented by Ostendorf et al. in [220].

Limited Resources for Languages

Another important motivational factor for developing automated segmentation systems is language related. Due to limited resources available for some languages a strong motivation exists to develop automatic systems to aid with efficiently producing phonetically segmented and labeled corpora of speech signals [315,213,152,192]. Sourcing audio information for the purpose of speech research is not a problem as there is generally a wide profusion of audio recordings to draw from, such as content archives, radio (interviews, plays), television shows (soap operas, sitcoms, documentaries), news (read monologue), interviews (conversational dialog), audio books and so forth. However, in order for a speech corpus to be really useful it needs to be segmented and labeled accordingly, therefore it should contain information about the content and about the time alignment between the labels and the segments [310,185]. The importance of developing such automatic segmentation methods is a critical point in the value chain for the
3.2. Motivations and Thematic Applications

development of many corpus based technologies such as multilingual synthesis and advanced speech recognition systems [184].

Segmentation approaches have been proposed for use and tested on various different languages (corpora). Some examples listed here, are the languages that were used by some of the methods that formed part of this literature review.

1. **American English** TIMIT Corpus [184,109,278], ICSI Meeting Corpus [13], TIDIGITS [297].

2. **French** [184,186,185] data corpora: BDSONS [53] and BREF-80 [163].


4. **Castilian Spanish** (VESLIM [310,309].

5. **Dutch** [184,72] COGEN Corpus Gesproken Nederlands⁷, Spoken Dutch Corpus [96].

6. **Swedish** GROG Project, read-aloud monologue (professional published children’s audio books) and spontaneous dialog [285,284].

7. **Czech** [192,213].

8. **Polish** [152].

9. **Korean** [146].

10. **Afrikaans, isiZulu, Setswana**, [315].

**English Language**

The English language as previously mentioned has good coverage in terms of the range speech corpora available. Examples of some corpora include: TIMIT - a large corpus of North American Speech, CHILDES [180], CMU Kids [82], PF-STAR [38] matched recordings of read British English spoken by British and non-British children, YOHO [50] a speech corpus created to support text-dependent speaker authentication research compiled for biometric applications, and TIDIGITS [168] created for the purpose of designing and evaluating algorithms for speaker-independent recognition of connected digit sequences. The English language was the main focus of this research and the widely used speech corpus TIMIT has been used extensively for experimentation and testing purposes.

⁶Distributed by LDC [184] see: http://www.ldc.upenn.edu/.

⁷Produced by the Katholieke Universiteit Leuven (KUL) and the University of Ghent (UGent).
3.3 Segmentation Methodologies

Firstly, this section discusses some topics that are related to the methodologies surrounding automatic speech segmentation. Some of these aspects are as follows: (i) supervised and unsupervised methods, (ii) text-dependent and text-independent methods, (iii) top-down and bottom-up hypotheses, and (iv) discusses phonemes as segmentation units. Following this some of the more predominant approaches from the literature are discussed along with an overview of the different types of feature sets that have been proposed for the speech segmentation task.

3.3.1 Supervised and Unsupervised

Segmentation methodologies can be broadly grouped in terms of supervised and unsupervised pattern recognition. Approaches that exploit a priori known information for training are referred to as supervised pattern recognition systems [307]. Another important type of problem exists where training data of known class labels are not available to the system. In this setting, the system is presented with a set of feature vectors and the task is to uncover underlying similarities in an effort to cluster (i.e. form groups) the vectors, consequently this is referred to as unsupervised pattern recognition or clustering [307]. As expected, the problem of unsupervised speech segmentation tends to be more difficult to achieve the same accuracy as that of supervised systems. However, unsupervised algorithms afford the important benefit of not requiring extensive training data and complex phonetic statistical modelling that is the case for supervised based approaches. This in many respects gives the unsupervised methods a significant advantage over supervised systems, for example where laborious training is not feasible or practical for use in the end-application. This is true of the educational based applications focused on in this thesis.

3.3.2 Text-dependent and Text-independent Approaches

Text-dependent speech segmentation is a special or constrained type of speech segmentation problem, where the word strings in the speech utterances are known to the system a-priori; in some cases even the phonetic strings are known. This information can be deduced using a phonetizer\textsuperscript{8} which is similar to the front-end stage of a speech synthesis system [275, 79], as

\textsuperscript{8}This is sometimes referred to as Letter-to-Sound (LTS), phonetizers come in two main forms (i) dictionary based and (ii) rule based using grapheme-to-phoneme rules [79].
3.3. Segmentation Methodologies

Figure 3.3: This diagram gives an overview of the different hierarchical topologies of (i) Top-down (ii) Bottom-up segmentation. The phoneme symbols are represented in ARPA-bet format [71].

was described in Chapter 2. In the case of Text-independent speech segmentation, no prior information about the word string is known, which is a more difficult challenge. This can be further exacerbated for the case of unsupervised segmentation in the text-independent case. This is due to not having any a-priori information, that is (i) no training data is available (ii) no a-priori information is known about the content (word string) to be segmented and (iii) therefore the number of segment boundaries is not known. Thus the problem is compounded by the fact that not only does the system have to find the temporal location of the segment boundaries, it also has to estimate how many boundaries there are to begin with. This particular class of problem is often referred to as blind speech segmentation, some examples include: [317,200,279].

3.3.3 Top-down and Bottom-up Segmentation

The topic of cognitive speech segmentation is often debated in the literature from two stand points bottom-up and top-down, [69,49]. The different hypotheses (interactive and modular) for human speech segmentation are based around the role of bottom-up and top-down cuing information in the segmentation process [49]. A simplified model of top-down and bottom-up segmentation is presented in Fig. 3.3. This example is for an utterance of speech, which is segmented at different levels, i.e. sentence, words, phonemes.

An interesting speech segmentation approach was motivated by how humans actually perform manual segmentation\(^9\) using audio editing suites and tools [309]. Under certain conditions this can be considered to be a definite form of top-down based segmentation process, that is starting with a view of the entire waveform and horizontally zooming into the various regions of interest.

\(^9\)This manual segmentation is for the purpose of transcription and labelling of speech corpora.
Usually the procedure involves starting with rough initial estimates for segmentation markers which are then further refined and adjusted accordingly.

This segmentation concept has been explored in other fields, such as image processing. There are systems that use both bottom-up and top-down cuing information for the purpose of efficient segmentation [161, 40, 41] as well as studies in experimental psychology for visual search tasks [328]. The segmentation approach proposed in this thesis can be interpreted as an effective coupling of the two topologies, as both top-down and bottom-up information is used within the segmentation algorithm, i.e. word-level and sentence-level information.

3.3.4 Phoneme Segments

The segmentation approach taken in this work focuses specifically on estimating the word boundary information to perform the segmentation task. However one might argue, why not focus on the lower level speech units such as the phoneme? the premise being that once these smaller segments are obtained, the task reduces to a straightforward consolidation of these segment units together to form the larger segments of the desired word boundaries. However, it is thought that adopting such a strategy in trying to estimate initial phonemes would be counterproductive, as it would necessitate the advanced and complex modeling where phoneme inclusions and deletions occur as the system cannot guarantee all the phonemes to occur. This is due to co-articulation, pronunciation variation based on context, talking styles, fluencies, and other related effects. Effectively, in this regard it is thought that such a strategy would be unnecessarily over complicating the solution to the problem.

Ostendorf [219] in a related discussion for speech processing, describes this judicially as “the case against the phoneme”, wherein she presents some reasons why it might be unfavourable to focus most of our research efforts and attention around the phoneme speech unit. It is often the case that speech signals are modelled as consisting of a straightforward concatenation of phoneme units. This simplified model has been sometimes termed the \textit{beads on a string} model\textsuperscript{10} [103]. However, it is not beneficial to simplify the speech signal in this regard, as the speech signal is a lot more complex in nature, there are various other factors that need to be accounted for, such as the ones previously described, e.g. co-articulation, fluencies, pronunciation and so forth. As a consequence of this, other units which are described based on acoustically-derived or data-driven bases have been proposed in [219], hence moving away from the simple pronunciation model and one which also needs to incorporate and account for acoustic modeling.

On the other hand, the beads on the string model is perhaps more applicable to larger segments, such as \textit{words} where they form the beads with additional beads for silence and pause periods. As proper narration of the text transcript is present, we can say that with a degree of certainty all of the words will be present in the audio stream otherwise errors would be present. Granted sometimes for informal conversational speech the pronunciation at word level can be

\textsuperscript{10}Individual beads on a string correspond to particular phonemes in a speech signal.
3.3. **Segmentation Methodologies**

affected by syntax, for example saying /going to/ can often become a single word /gonna/ [219]. As is usual when dealing with conversational speech it is necessary to allow for differences in pronunciation and speech containing fragments of words, interruptions, incomplete sentences, fillers and discourse markers and dis-fluency factors [202]. This, however, is not a major issue within this work as the nature of the spoken audio will mostly be read in monologue. Although conversation dialog will sometimes play a part, for instance character’s voices in a children’s story could form more dialog than narration, the dialog should still be spoken with proper elocution, pronunciation and articulation of each of the words for the educational purposes. The reason for this is heavily motivated by the end-application being educational resources and instructional materials.

3.3.5 **Automated Segmentation Approaches**

The most predominant approach in the literature in the area of automatic speech segmentation is based on using stochastic Hidden Markov Model (HMM) [129, 223, 5, 156, 310, 184, 192, 72, 84]. Another approach that exists uses Dynamic Time Warping (DTW). However this approach has been explored to a much lesser extent [184, 227, 315]. These two approaches have similarities such that they are both heavily based on using dynamic programming principles. In fact, the HMM approach can be considered to be a generalization of DTW [71]. Both of these methodologies are well suited towards the specific class of segmentation under investigation herein, that of text-dependent speech segmentation. Many of the other approaches from the literature which are not of these two kinds are usually aimed towards a different class of speech segmentation problem, that of text-independent speech segmentation. Although that being said, recently a text-independent segmentation method, that was initially proposed 28 years earlier by Brandt in [44] was re-examined by Jarifi et al. [129] and incorporated into a HMM based segmentation system. Brandt’s method is known as the Generalized Likelihood Ratio (GLR) method, which works off the basis of detecting discontinuities in the speech signal.

HMM based approaches are a form of supervised statistical based machine learning algorithm. They are quite complex in nature and require a good degree of training data to work in an effective manner. A popular method for speech segmentation, particularly at the phonetic boundaries but also at the word boundaries [285, 72], is based on a technique called **Forced Alignment** [136, 129, 223, 5, 156, 310, 184, 192, 84]. This approach uses a speech recognition engine based on statistical methods using Hidden Markov Models (HMMs). HMMs are the predominant methodology employed by most modern speech recognition systems. An overview of the Forced Alignment HMM methodology is presented in Fig. 3.4.

The recognition system is used in such a manner where the known *a-priori* word string is exploited to provide segmentation. The Viterbi algorithm is commonly used in the HMM procedure, and hence sometimes the technique is referred to as **Viterbi Alignment** [136]. This a-priori knowledge results in the Viterbi decoding procedure getting simplified to the task of
Viterbi training. In this training procedure the Viterbi algorithm is Forced to pass through certain words as defined by the known a-priori transcript, hence the name Forced Alignment. As no recognition is actually performed, since it does not require determining the correct words in the sequence, it requires only to determine the correct state (sub-phone) sequence. The model identifies the corresponding temporal regions in the speech signal where the phonemes in transcript occur and this gives the segmentation timing information.

The model parameters are trained based on a collection of speech data. The underlying framework effectively can model each phone unit by content dependent or independent basis [223]. Although the HMM approach can be achieved using standard techniques such as the Baum-Welch and Viterbi algorithms, this can be generally regarded as the baseline system. The accuracy for the baseline system is often not satisfactory to be fully useful for the target end-applications. Accordingly, many improvements and further refinements stages are proposed in the literature to improve upon the baseline system [223, 5, 156, 310, 192, 72]. These are some examples of refinements of the baseline system:

Some approaches proposed statistically benchmarking the discrepancies between the automatic and manual segmentations. Using this information they provide enhancements to the automatic approach; examples of this are: Matousék et al. [192] and Adell et al. [5] propose refinement through a Boundary Specific Correction method (BSC). Toledano [310] proposes several approaches to refinement using: Fuzzy Logic, Neural Networks (NN), and Gaussian Mixture Models (GMM).

Other approaches have been proposed that use boundaries obtained from multiple segmentation methods, and these multiple segmentation estimates are combined to give a better estimate. Examples of this, using straightforward averaging of the segmentation estimates, are Kominek and Black [156], and Demuynck [72]. More sophisticated averaging methods have been proposed.
3.3. Segmentation Methodologies

by Park and Kim to compensate for context-dependent factors, which are a significant source of errors for each system and which compute the weighted sum of the bias-removed estimates to give final segmentation estimates [223]. Demuynck also proposes a technique based on confidence intervals derived from the forward and backward algorithm [72].

Refinements have also been proposed, with additional stages, that incorporate other acoustic features, such as the approach by Flammia et al. using the Spectral Variation Function (SVF) [84] and by Saito using the pitch contour [263].

Jarifi et al. [129] discuss some of the limitations of HMM model based approaches, commenting on them as being able to model steady state areas well in speech. However, they have certain shortcomings in terms of detecting the transitions between phonemes (transient events in the signal). Moreover, Matoushek et al. [192] and Toledano et al. [309] attribute some of the limitations of the approach to the principle of the HMM-based approach itself; that is HMMs within the context of speech recognition are built to identify the phonetic segments, not to produce accurate segment boundaries. The framework itself is not optimized for the specific task of segmentation, rather it is aimed towards a different task, that of recognition.

The DTW technique was originally used to solve an important problem in relation to template matching for early IWR speech recognition. This technique solved the problem of temporal alignment of reference utterances with a query utterance (unknown word). This process is referred to as time warping and was carried out on feature representations for the two signals, where they are expanded or compressed in order to time align them with each other. This template matching process was applied to incoming words against reference words stored in an inventory word bank. The reference word that produced the best template match, based on computing the distance between them, was considered to be the word spoken. The basic idea behind using DTW, for the purpose of automating the speech segmentation process, is similar in some respects to the recognition case. In this case, however, the reference utterances are generated using a speech synthesizer. Using the speech synthesizer in this regard is advantageous as the phonetic boundary information is known for the synthetic signal. This is because the signal is produced by a machine, therefore a transcript of temporal boundaries can be produced automatically. The two signals are time aligned similar to the template matching procedure mentioned. Then the required segment points can be estimated by transferring the synthetic reference signal segment boundary values over onto the natural speech signal. Some examples from the literature where the DTW methodology has been explored are Malfrere [184], Paulo [227] and Niekerk [315].

There have been other approaches reported in the literature incorporating Neural Networks, [266,179,184,58,300]. Furthermore, some hybrid approaches have also been proposed HMM with ANN [184,309].

There are many blind speech segmentation methods reported in the literature, two noteworthy ones are Brandts [44] Generalized Likelihood Ratio (GLR) method, [129,128] and more recently the Cohens Voting Experts Algorithm [65], applied to Audio Segmentation by Miller
et al. [204, 203]. Brandts algorithm effectively aims at determining discontinuities in the speech signal. This approach is linguistically unconstrained and as a result it makes many insertions and omissions. Recently Miller et al. have proposed a blind speech segmentation approach based on the Voting Experts (VE) algorithm proposed by Cohen and Adams [64]. The approach was inspired by studies by Saffran et al. [261, 260], which have already been mentioned earlier in this Chapter. These studies have strongly suggested that human speech segmentation of words is performed in an unsupervised manner without requiring any external cues [204]. The underlying idea behind the VE algorithm is such that natural chunks or segments exhibit two informational theoretic properties; they have (i) low internal entropy and (ii) high boundary entropy, i.e. at the edges. The VE algorithm was originally targeted towards text sequences, and Miller et al. have developed the idea further, extending it to the audio signal domain [204, 203].

Other blind segmentation approaches have been proposed using techniques of Non-negative Matrix Factorization/Deconvolution [297, 178], Support Vector Machines [179, 178], Wavelets [152], N-gram Language Modeling [296], Continuous Multi-resolution Entropy/Divergence [59]; Convex Hull Method [279, 201], Feature based methods [268, 14, 146, 321, 19, 109, 1, 286, 245].

3.3.6 Speech Features for Segmentation

There are many features that have been proposed for the task of characterizing a speech signal for the purpose of segmentation. Some of these features are novel and have been specifically formulated for the task itself, while others have been adopted from other speech processing systems in particular speech recognition and speech coding. Some of the more noteworthy features taken from the literature on speech segmentation are listed below:

- **Perceptual Linear Prediction** Features with RASTA [184], Mel Frequency Cepstral Coefficients [315, 129, 179, 192], Cepstral [213, 184], Information Theoretic [242, 203, 59, 298, 109], Kullback-Leibler Divergence [59, 42], Energy [321, 184, 185, 186, 245], Zero Crossing Rate (ZCR) [268, 184, 321, 185, 186, 245], Autocorrelation [14], Onsets [286], Time-Frequency [146], Prosodic Features including Pitch $F_0$ [259, 13, 231, 321, 177, 296], Spectral Variation [279, 316].

3.4 Segmentation Performance Metrics

Some early speech recognition systems in the literature [135, 162], have incorporated the subsequent recognition process to measure their systems segmentation accuracy and performance. This is done by measuring the performance of the speech recognition system itself, i.e. correct number of words identified. The rationale or the premise behind this was such that the more accurate the segmentation system is, the greater the accuracy of recognition. The central goal of speech segmentation is to estimate, as accurately as possible, important events or transi-
tion points in speech. These segmentation points can be linguistically or acoustically motivated units. As a task, this has an inherent degree of subjectivity, as the ground truth on which we measure the performance is generally based on manual segmentation estimates. For professionally transcribed corpora (American English TIMIT corpus [83]), (French BDSONS corpus [53]), (Spanish LATINO-40\textsuperscript{11}), these are segmented and transcribed by phoneticians (experts) using sound editing suites and tools. However, often there is some debate amongst these experts as to where the exact locations of the transition points occur as a relative amount of subjective based judgment is present. Hence for the recognition systems aforementioned, measuring performance based on recognition results proved advantageous as it was an objective measure on which to base the performance assessment.

There have been various evaluation approaches proposed, some of which are not suitable for the type of segmentation that this work addresses. Some of these unsuitable measures are aimed towards text-independent segmentation. These segmentation approaches attempt to uncover the temporal location in the speech signal for segments, as well as estimating how many segments are present. This is not the case in this work as it is focused on text-dependent segmentation, since the number of segments are known a-priori as a transcript is present.

Text-independent segmentation performance measures include, Recall, Precision, Hit-rate, False Alarm, [296], Widow-diff [255, 232], and R-value, [249]. Text-dependent measures are, Overlap Rate [212, 212, 228], Mean Deviation, Standard Deviation [192], and Segmentation Accuracy for Tolerances [310, 129, 212, 192, 184, 72, 185].

The most common measure from the literature for the text-dependent case has been found to be segmentation accuracy for different tolerance values. This measure accounts for the percentage of segmentation estimates that lie within the given tolerance (usually in msec) from the ground-truth. This measure has been used in this research for tolerances of 50, \ldots, 150 in msec. Also measures of sample mean \(\bar{x}\) and standard deviation \(S\) for segmentation error have been used.

3.5 Conclusion

In this chapter the requirements and motivation for developing a new segmentation system have been presented and the limitations of known methods have been discussed. The particular end-application of this research is in the area of automating the production process for generating high quality rich-media educational content. This requires an automated segmentation method that is robust enough to not require any manual intervention and accurate enough to be able to find the word boundary points in the continuous natural spoken voice recordings. This demands high segmentation accuracy for tolerances in the region of ±90msec. An unsupervised algorithm is required as unlike supervised approaches from the literature, this affords the important benefit of not requiring extensive training data and complex phonetic statistical modelling. This flexi-

\textsuperscript{11}Distributed by LDC [184]: http://www.ldc.upenn.edu/.

bility is extremely important in relation to the end-application in the area of educational content generation as it enables the segmentation system to be more adaptable and user-friendly for technology integration purposes. The approach should also be *text-dependent* as the information is available, i.e. the word strings in the speech utterances are known to the segmentation system *a-priori*. The objective of having a *text-independent* approach would unnecessarily compound the segmentation problem; by the fact that not only does the system need to find the temporal location of the segment boundaries it also has to estimate how many segment boundaries there are present.

This Chapter began by discussing the problem of speech segmentation and related this to other works and discourse in the field. The literature has been reviewed and discussed in terms of three main areas: (i) Thematic Applications and Motivations (ii) Segmentation Methodologies and Speech Features and (iii) Performance and Evaluation Metrics. The next Chapter will report upon the signal analysis and feature extraction methods explored which form an integral part of the proposed automated segmentation algorithm.
Speech Analysis and Feature Extraction

This Chapter explores different methodologies for extracting pertinent feature profiles from speech signals. The features are reported and discussed relative to the speech segmentation problem. The Chapter begins with a discussion on the philosophy of features and their importance for addressing the challenges of speech segmentation. The work has investigated many different approaches to feature extraction, some being univariate features and others being multivariate (vector based) features. Particular attention has been given to the more predominant approaches from the literature, most notably short-time spectrum based features: Filter banks, Spectral: Flux, Centroid and Roll-off; Cepstrum features such as Mel-Frequency Cepstral Coefficients (MFCCs); Linear Predictive Analysis features such as Perceptual Linear Prediction (PLP) features; and Dynamic feature extensions such as first order derivative (delta or $\Delta$), second order derivative (delta-delta or $\Delta\Delta$) and Relative Spectra: termed RASTA processing.

The Chapter also proposes some other features for the purpose of speech segmentation. These features are spectrum features such as the Power Spectrum and the Teager\(^1\) Energy Spectrum. Such features tend to have quite a high frequency resolution ($\frac{f_s}{N}$), in terms of STFT spectral bins, as $N$ is usually relatively large. Take for example a 512 point Fast Fourier Transform (FFT) which gives 256 spectral bins as possible feature points to utilize. Additionally, in contrast to these rather fine spectral detail features, novel features are proposed for the segmentation task that incorporate a degree of spectral smoothing in non-equally spaced spectral bands which are termed energy based Octave Band Analysis (OBA). These proposed features are not derived

\(^1\)Teager's Energy Operator (TEO) is an interesting non-linear energy operator that was first introduced by Kaiser in [138].
from any perceptual effects, in contrast to PLP and MFCC. The rationale and motivation behind exploring these features was to try to uncover whether or not any additional gains could be achieved for speech segmentation. Also, they provide a good comparison against the more widely adopted State of the Art features from the literature namely, (MFCCs) [243,336,68] and (PLPs) features [209,113,111]. Throughout this Chapter some visual plots of the different feature representations are given which will enable comparison between features. This will provide greater insight into understanding the different features which are applied to speech segmentation. Additionally, it will aid in comparison between the two signal types of Synthesized Voice Recording (SVR) and Natural Voice Recording (NVR). Before the Chapter concludes, the technique of Principal Component Analysis (PCA) is discussed as a valuable method for feature transformation and dimensionality reduction of the feature vectors.

4.1 Signal Representations: The Philosophy of Features

Firstly, it must be appreciated that the characteristics and requirements of features depend heavily on the intended end-application or their role within a subsequent stage belonging to a larger system. Consider for example three applications, where feature extraction or signal characterization plays an important pre-processing role: (i) speech coding (ii) speaker recognition and (iii) speech recognition.

The case of speech coding is concerned with the efficient transmission and storage of speech signals. In this case, the aim is to represent the signal in the least number of possible bits [290]. Thus the feature extraction stage requires pertinent signal characteristics that enable the signal to be adequately reconstructed while preserving the quality and the intelligibility of the signal. Suitable features in this case include Linear Prediction Coefficients (LPC) [243,290,71,225].

Speaker identification concerns itself with the verification and identification of speaker’s voices which belong to a specific class of biometric problems. The problem requires extracting what are known as speaker dependent characteristics from the speech signal [121,51]. Anatomical and behavioral differences between candidates cause speaker-dependent characteristics and attributes to be manifested in the signal due to the speaker’s vocal tract configuration [121]. Therefore extraction techniques and features that best characterize the vocal tract \(v(t)\) see Eq. (4.1)) offer a suitable way of discriminating amongst speakers for the purpose of the speaker recognition task. A good overview of some techniques for this task has been presented by Campbell in [51]. Features that have proved popular for this task are the MFCC and LPCC, which is interesting considering these features have been principally used for other tasks such as speech recognition and speech coding. Ideally, features for speaker identification differ to those required for the speech recognition or speech coding tasks. In this particular case fine detail is needed to discriminate associated talkers from one another independent of the linguistic message. These features MFCC and LPCC effectively characterize the vocal tract so well, that they also characterize the subtle speaker-dependent differences which is undesirable. However, there are different
4.1. Signal Representations: The Philosophy of Features

Techniques to reduce these adverse effects such as Vocal Tract Normalization (VTN), [235].

Similar to the case of speech recognition the features used in speech segmentation should ideally characterize the linguistic message of the spoken audio well, rather than the speaker-dependent aspects. This is so that the two feature profiles for the SVR and NVR signals have similar characteristics for the time alignment procedure. A simple model of the speech signal is given by:

\[ s(t) = g(t) \ast v(t) \]  

where \( s(t) \) represents the speech signal in time, \( g(t) \) represents the excitation signal, which can be modeled as periodic excitation for voiced speech or white noise for unvoiced speech, and \( v(t) \) denotes the vocal tract impulse response. The term \( v(t) \) is a time varying filter which changes in order to produce different speech sounds [121].

It is reasonable to question why different representations of the speech signal are needed, or why the raw speech signal is not used directly as one composite feature to represent the speech signal. Although in other areas, such as biomedical engineering which sometimes performs processing on raw ECG signals, this direct approach can be adopted due to the type of signal (ECG) concerned. However, for the purpose of characterizing the speech signal well a similar strategy is not viable due to the high variability present in the speech signal. The speech signal for different sounds exhibits quasi-stationary properties as well as periods of non-stationarity and non-linearity. For instance the speech signal can vary between voiced speech, which is quasi-periodic, (e.g. /i/, /e/, /e/, and /a/ vowels), to unvoiced speech, which is highly non-stationary, (e.g. /sh/ fricative), to mixed (voiced and unvoiced e.g. /z/ of azure), to short silence regions (e.g. the pause before the burst of plosive sounds) having low energy.

The variabilities can be defined in terms of three distinct categories (i) linguistic variabilities comprising of effects of phonetics, phonemics, syntax, semantics, (ii) speaker inter/intra-variabilities includes effects of co-articulation, pronunciation, fluencies, age, gender, accent, emotion, speaking rates, styles and other related idiosyncrasies, (iii) channel variabilities include background noise, signal acquisition effects, reverberation and microphones distortions, bandwidth, sampling and quantization [182,71,225]. While the latter is important, it is not as significant to this research as the former two variabilities; linguistic and speaker. These two are of more immediate interest to this work where the characteristics of each must be carefully considered in accordance with the task at hand.

Ideally for the purpose of speech segmentation these features should be neutral in terms of speaker-dependent characteristics so their effects are minimized (negligible). In a similar way speech recognition requires the features to characterize well the message from both an acoustic and linguistic point of view, that is to define the speech in terms of phonemes, syllables, words, phrases and so on. For speech recognition the task can be thought of as trying to uncover what

---

Aspects of non-linearity in speech production has been described in greater detail Chapter 2.
has been said, whereas for this work, that of segmentation, the word string is already known a-priori. The challenge therefore lies in determining when exactly the word boundaries, occur temporally in the given utterance.

In order to overcome all of these variability factors within a pattern recognition context, multiple feature representations are often employed [182]. This process is called feature extraction or sometimes feature generation [307]. The goal is to generate features that exhibit high information-packing properties [307] in order to best describe the speech signal over time. Typically, multivariate features are used in the pre-processing stages for most speech processing systems [334, 182, 166], although some authors in the past have proposed methods using single features [14, 286, 244, 245]. In general, in speech processing systems using one feature (parameter) representation alone in the system has been found to be more complex, and difficult to achieve good performance [135, 162, 211]. Equally, this was found to be the case in this research as using a univariate feature extraction stage was found to give rise to singularities in the feature warping stage which hampered the algorithm performance considerably, therein giving rise to gross errors in the segmentation process. This topic of singularities will be taken up again in the results discussion section of Chapter 5.

It is usually relatively easy to compare univariate feature representations for two signals, e.g. SVR and NVR, whereas for multivariate feature representations this is not always the case. Often it can be difficult to notice underlying patterns or similarities when comparing the multivariate features (time series) from two signals. Although no patterns may be obvious upon visual observation, these signal representations can often perform well with respect to the time alignment procedure. This can be somewhat deceiving and almost counterintuitive when first observed or encountered. Some reasons as to why this is the case can be attributed to time differences between the two time series, as well as differences in magnitudes (scaling) and other signal decomposition based factors in the feature profiles. There are often underlying statistical structure details and patterns present within the data, that the template matching algorithm has the ability to uncover using vector based distance metrics. These may not necessarily be obvious in a visual sense when comparing similarities for feature representations, such as by viewing them in a two dimensional plot format. This is something to be mindful of when viewing and comparing feature profiles in the preceding sections of this Chapter.

4.1.1 Pre-emphasis

The glottal signal in analysis is often modeled as a two-real-pole filter where the poles lie near \( z = 1 \), and the lip radiation effects can be modelled as adding a zero near \( z = 1 \) effectively cancelling the spectral effects of one of the glottal poles. It is common in many speech processing

---

3The definition of what is meant by exactly is left open here and shall be quantified later in Chapter 5, under the topic of tolerances. For the moment it is perfectly fine to assume accuracy in the region of tens of milli-seconds.

4The term singularity refers to when the DTW procedure is not well-behaved, which produces undesirable warps [145].
systems to perform pre-emphasis filtering on the speech signal. The reasoning behind this pre-emphasis stage is to compensate for the -6dB/octave spectral slope of the speech signal [233]. This was topic discussed in Chapter 2 and was summarized in Fig. 2.5. The filter emphasizes the higher frequency components which effectively removes some of the spectral tilt in the spectrum. This is beneficial as it eliminates the spectral contributions of the larynx and the lips and the analysis can therefore be concentrated towards seeking parameters corresponding to the vocal tract only [71]. In particular, it is an important technique for LPC analysis and has been adopted for use as a pre-processor in speech recognition systems such as CMU Sphinx [166] and the HTK recognition system [334]. A popular choice of pre-emphasis filter is the simple filter given by,

$$y(n) = s(n) - \mu s(n - 1),$$

where $y(n)$ is the output from the filter, and $s(n)$ corresponds to the input speech signal. The $z$-transform of this filter is defined as,

$$P(z) = \frac{Y(z)}{S(z)} = 1 - \mu z^{-1}.$$  

Where the value of $\mu$ can vary over the range $[0, 1]$, where $\mu = 0$ corresponds to no filtering at all. Appropriately chosen values of $\mu$ usually lie in the range $[0.8, 1]$ [166, 71], a popular choice being $\mu = 0.97$. This filter introduces a zero near $w = 0$, and a 6dB per octave gain. The addition of this zero near $z = 1$, tends to cancel the remaining effects in vocal tract as described above.

Further, studies [100, 183] have proposed an optimal value in a Mean Squared Error (MSE) sense for $\mu$, determined from using short-term autocorrelation analysis. This is related to another practical reason for applying pre-emphasis filtering which is to overcome numerical instability in LP analysis [71]. If the spectrum is dominated by low frequencies (high spectral tilt/slope) this could result in an ill-conditioned auto-correlation matrix. As already mentioned the effects of pre-emphasis stops this by effectively whitening the spectrum. Ideally, pre-emphasis filtering should only be performed on voiced speech. This technique can be thought of as emphasizing the upper portion of the spectrum and therefore giving the higher formants in the spectrum more influence on the outcome. This concept is indirectly incorporated into a non-linear energy spectrum transformation method explored later in this Chapter.

### 4.2 Short-term Processing of Signals

Speech is a non-stationary signal as its statistical properties vary with time [215]. This is clearly evident from a straightforward observation of a spectrogram plot of an utterance of speech, for example refer back to Fig. 2.7 in Chapter 2. However, despite this, the signal characteristics over durations in the region of 30 to 40 milliseconds can be assumed to remain constant [215, 71]. This assumption forms the basis for short-term signal analysis of speech. The procedure of short-term signal analysis involves dividing the discrete time speech signal into successive frames of
duration $N$ samples. Accordingly, over these short frames the signal can be assumed to be quasi-stationary [112, 169, 237]. The mathematical formulation of segmenting the signal into these frames is equivalent to multiplying the signal by a rectangular window $w(m)$ as shown in Eqs. (4.4) and (4.5).

Windowing is applied to the speech signal as given in the following,

$$f_n(m) = x(m)w(n - m)$$

(4.4)

where $f_n(m)$ is a short-time section of $x(m)$ at time $n$, the window $w(m)$ is defined as:

$$w(m) = \begin{cases} 1, & m = 0, 1, ..., N - 1, \\ 0, & \text{otherwise}, \end{cases}$$

(4.5)

Alternative windowing profiles can be used such as the Hamming window [108], which is defined as,

$$w(m) = \begin{cases} 0.54 - 0.46\cos(2\pi m/N - 1), & m = 0, 1, ..., N - 1, \\ 0, & \text{otherwise}. \end{cases}$$

(4.6)

Other types of windows commonly used in signal processing include Hann (or Hanning) and Triangular, which are shown in Fig. 4.1. Furthermore, some other examples are presented in the Appendix; see Fig. 9.4. However the most predominate used window in speech processing is the Hamming window function [12, 307]. The choice of window has a significant effect on the smoothness of the spectrum. It is fair to say that a wide range of analysis windows have been proposed in the literature. Harris gives a detailed account of this topic in [108]. The standard windows are real valued and symmetric, and they have a frequency spectrum which has a sinc-like shaped response $\left(\frac{\sin(x)}{x}\right)$ [277], which can be observed in Fig. 4.1.

The length $N$ is usually discussed in terms of the approximate number of pitch periods that the window embraces. A typical frame length of $N$ samples, corresponding to a duration of 20 milliseconds, is often used [234, 111, 166], but in general the duration of window length varies roughly between 10 to 40 milliseconds. A time-frequency resolution trade-off exists such that if the window is too long the signal properties may change too much, to the point that the stationarity assumption no longer holds true. If the window is too short in duration, narrow-band signal components will not be resolved in the spectrum. Thus, short windows give good time resolution and a smoothing of spectral harmonics into wider frequency formants which tends to be better for formant analysis. Whereas for features that require harmonic analysis and $F_0$ (pitch) detection, the resolution of individual harmonics is important. In this case longer windows are more appropriate [218]. Furthermore, studies have found that sounds with onsets faster than 20 milliseconds are heard as abrupt plucks and not gliding bows. This is because the maximum rise time for a stimulus onset which causes an overshoot in the auditory neural firings is approximately 20 milliseconds [218]. Taking all of these factors into account, the appropriate integration time for speech processing is usually around 20 milliseconds.
4.2. Short-term Processing of Signals

Figure 4.1: Three different window functions: Hann, Hamming and Triangular are shown here in the two domains; the time domain on the left and the frequency domain on the right. The curves show the windows’ characteristic shape in the time domain and subsequent frequency domain response, i.e. the magnitude and width of the side lobes.

One of the practical considerations of the STFT is that the use of the FFT requires the length of the short-term signal to be a power of two [277]. This can be achieved by either using the correct length or by zero padding the signal to the correct length. The term zero padding refers to adding zero’s to the signal to make it the appropriate length. In the frequency domain this corresponds to interpolation, which is favourable as it results in a smoother spectrum. To increase the temporal resolution frame overlapping is usually incorporated into the procedure. This is often quantified by a term hop size which refers to how much each frame advances along the signal, i.e. the amount of shift between consecutive frames. In general, the more overlap that exists the greater the resolution will be across time. Accordingly, a trade off exists as the smaller the hop size the larger the increase in processing. This is due to the increase in the number of frames to be processed. A criterion for estimating the hop size is the window length divided by the main lobe width in terms of the number of frequency bins [10,11]. For the particular case of the popular Hamming window, a good rule of thumb is to let the hop size equal to one fourth of the window size which corresponds to a window overlap of 75%. Commonly adopted overlap sizes are 75% and 50%. [108].
4.2.1 Short-time Fourier Transform

In general the STFT is one of the most widely used analysis techniques in speech processing. The discrete-time STFT is given by

\[ X(n, \omega) = \sum_{m=-\infty}^{\infty} x(m)w(n - m)e^{-j\omega m} \]  

(4.7)

The discrete STFT is obtained by sampling the discrete-time STFT with a frequency sampling interval of \( \frac{2\pi}{N} \) [243]. This relation is defined as,

\[ X(n, k) = X(n, \omega) \big|_{\omega = \frac{2\pi}{N} k} \]  

(4.8)

such that

\[ X(n, k) = \sum_{m=-\infty}^{\infty} x(m)w(n - m)e^{-j\frac{2\pi}{N} km} \]  

(4.9)

By changing the limits of summation to correspond to a window length \( N \), Eq. (4.9) becomes,

\[ X(n, k) = \sum_{m=0}^{N-1} x(m)w(n - m)e^{-j\frac{2\pi}{N} km} \]  

(4.10)

The spectrogram was first presented in Chapter 2, where two examples of utterances of speech, SVR and NVR respectively, were given in Fig. 2.7. STFT with regard to speech processing is generally referred to as Spectrogram analysis, as most speech processing systems tend to operate on the magnitude information only, i.e. the \((\text{STFT}_M)\). There is also another technique known as a multi-resolution spectrum which is described in [54].

4.2.1.1 STFT derived Features

It is fair to say that STFT analysis is the most predominant tool used in speech processing. There have been many features and algorithms derived which use the STFT. Some examples of univariate features derived from the STFT are the Spectral Centroid, Spectral Roll-off and Spectral Flux. These have successfully been used in speech processing front-ends and have been covered in [121,307,313].

Spectral Centroid gives an indicator of spectral shape or a measure of the \textit{brightness} of the acoustic structure present. High values of spectral centroid indicate more energy in the higher frequencies [121,307]. The equation for computing Spectral Centroid is given by:

\[ C(i) = \frac{\sum_{m=0}^{N-1} m \left| X_i(m) \right|}{\sum_{m=0}^{N-1} \left| X_i(m) \right|} \]  

(4.11)

where \( i \) indexes the speech frames.

Spectral Roll-Off is based around a percentage or threshold point \( c\% \), where \( c \) usually lies in the range 85-90\%. It is defined as the frequency sample \( m_c(i) \) below the point where \( c\% \) of the magnitude distribution of the STFT is concentrated [307]. This is satisfied by the relation,
4.2. Short-term Processing of Signals

\[ \sum_{m=0}^{\text{mc}(i)} |X_i(m)| < \frac{c}{100} \sum_{m=0}^{N-1} |X_i(m)|. \]  
\((4.12)\)

such that \(\max[m_c(i)]\) is required.

Spectral Flux is a measure of local spectral change between successive frames, \((i)\) and \((i-1)\), as given by,

\[ F(i) = \sum_{m=0}^{N-1} (X'_i(m) - X'_{i-1}(m))^2 \]  
\((4.13)\)

\(X'\) is used here, to denote normalization as the values are normalized by the maximum value. An example which shows each of these three features across time for a speech utterance is given in Fig. 4.2. In addition, other information theoretic based spectrum features have been proposed for use in front-end processing of speech signals. A notable one is the Spectral Entropy feature examples of which are presented in [7,329,206,131].

Another more straightforward method for feature extraction from the STFT is to use the spectral bins directly. Usually the Power spectrum is used rather than using the STFTM directly. Using these STFT features in this regard has been explored in this research as simple features derived from the spectrum. In addition to using the Power Spectrum the STFTM has been transformed into a Teager Energy Spectrum\(^5\) for the purpose of feature extraction. The energy operators shown below in Eqs. (4.14) and (4.15) were developed by Teager during his work on modeling speech production and were first introduced by Kaiser [138,139]. Hence the operator is known as Teager's Energy Operator (TEO).

\[ \Psi_e[x(t)] \triangleq (\frac{dx}{dt})^2 - x(t) \frac{d^2x}{dt^2}(t) = [\dot{x}(t)]^2 - x(t)\ddot{x}(t), \]  
\((4.14)\)

where \(\dot{x} = \frac{dx}{dt}\), and its discrete counterpart:

\[ \Psi_d[x(n)] \triangleq x^2(n) - x(n-1)x(n+1), \]  
\((4.15)\)

for discrete-time signals \(x(n)\), for \(n = \{0,1,2,\ldots,N\}\)

Before proceeding there is a clear prerequisite to distinguish what is meant by the term: the energy of a signal. Traditionally in signal processing literature [71,138] the energy of a discrete time signal is given by,

\[ E_x \triangleq \sum_{n=-\infty}^{\infty} |x(n)|^2 \]  
\((4.16)\)

and \(x(n)\) is called an energy signal if \(0 < E_x < \infty\).

\(^5\)A practical consideration - since the TEO can only be applied for \(\omega\) in the range \((0 \rightarrow \frac{\pi}{2})\), not the range \((0 \rightarrow \frac{\pi}{4})\), therefore in order to have the same frequency bands to compare with the Power Spectrum, this is accounted for by up-sampling the signal by a factor of 2 and subsequently increasing the FFT size by a factor of 2, e.g. from 1024 to 2048 points.
Speech Analysis and Feature Extraction

Figure 4.2: On the left hand side is the SVR and on the right hand side is the NVR. The speech utterance is “Chief interest in radiosterilization resides in the military services,” from the TIMIT Corpus and shown in (a & b). The feature profiles below the speech utterances are as follows: (c & d) Spectral Centroid, (e & f) Spectral Flux, and (g & h) Spectral Roll-off.

The non-linear energy operator $\Psi_{cld}$ provides an instantaneous estimate of the energy required to generate a signal [138]. A derivation is presented in [138] that uses basic physics of motion for a simple spring and mass oscillator to show that the total energy in the system
4.2. Short-term Processing of Signals

is proportional to both the amplitude and the frequency. This is very different to the energy measure in Eq. (4.16), where, from Parseval's relation all frequencies are treated uniformly. The operator \( \Psi_d \) is considered to be a high resolution energy estimator [106]. This property has provided motivation for investigating its use for obtaining spectral energy features. Pre-emphasis filtering, as discussed at the start of this chapter, effectively tries to emphasize the higher portion of the spectrum. The TEO has a similar effect on the spectrum giving the higher frequency content more weight. This can be observed from the three spectral plots as presented in Fig. 4.3. The absolute magnitude of the short-time spectrum (STFTM) is provided to reference the other two spectra. There is a noticeable change in the spectra for both the Power and the Teager Energy Spectrum. In the case of the Power Spectrum, the spectrum is emphasized in terms of the magnitude only whereas in the Teager case the spectrum is emphasized according to the magnitude squared as well as frequency \( \sin^2(\omega_k) \). The effects of emphasizing the upper portion of the spectrum like this effectively incorporates more of the pitch information and reduces the influence of the spectral roll-off considerably. There have been various techniques proposed for the purpose of flattening the spectrum; some have been motivated for pitch extraction purposes and others for LPC signal analysis. Some of these techniques for flattening (pre-emphasizing) the spectrum include: Adaptive filtering. Filter-banks with automatic gain control proposed by Sondhi et al. [289], LPC-derived inverse filter proposed by Markel et al. [190]. Non-linear Processing. Non-linear processing in the time domain is another approach, as it is understood that the low amplitude segments of the signal contains most of the formant information and the high amplitude segments contain most of the pitch information. Two examples that exploit this property are, (i) clipping the signal using thresholds [76, 289], and (ii) a cubing method which cubes the signal (i.e. \( s^3(n) \)) as proposed by Atal [18].

4.2.2 Energy based Octave Band Analysis

The spectral bins for the STFT tend to be quite discrete/granular in nature such that they contain a lot of spectral detail. Therefore it is often the case that these spectral bins are divided up into frequency bands, which can be considered as a form of spectral smoothing. There have been various approaches proposed for dividing the spectrum up into bands. Critical bands are an approach often used [114, 111, 148], . These bands can be based on warping the frequency axis into non-linear scales such as the Mel-scale (MFCC) [68] or the Bark scale (PLP) [111]. These types will be explored later in this Chapter. In general the bands tend to be non-linearly distributed with more bands being concentrated in the lower portion of the spectrum, which for speech is where the majority of the important message related information is contained. One such scheme for dividing the spectrum up into bands is known as Octave Band Analysis. The formula to compute these bands is given by,
Figure 4.3: On the left hand side is the SVR and on the right hand side is the NVR. The speech utterance is “Chief interest in radiosterilization resides in the military services,” from the TIMIT Corpus and shown in (a & b). The feature profiles presented are as follows: (c & d) Power Spectrum, (e & f) Power Spectrum as 3D magnitude plot, (g & h) Teager Energy Spectrum (TEO Spectrum), and (i & j) Teager Energy Spectrum as 3D magnitude plot.
4.2. Short-term Processing of Signals

\[ f_2 = 2(f_1) \quad \text{1 Octave Band} \quad (4.17a) \]
\[ f_2 = \sqrt[3]{2} f_1 \quad \frac{1}{3} \text{ Octave Band} \quad (4.17b) \]
\[ f_c = \sqrt{f_1 \cdot f_2} \quad \text{Geometric Centre Frequency} \quad (4.17c) \]

for the case of Octave and \(\frac{1}{3}\) Octave bands. The particular bands used in this work are given in Table 4.1, and were originally proposed much earlier in the literature by Beranek [27].

The procedure involves dividing up the STFT spectrum into separate bands. Then the energy for each band is computed in Parseval's sense.\(^6\) Parseval's theorem is represented below in Eq. (4.18) and in terms of the discrete counterparts, the Discrete-Time Fourier Transform (DTFT) Eq. 4.19a and Discrete Fourier Transform (DFT) Eq. 4.19a:

\[ \int_{-\infty}^{\infty} \left| x(t) \right|^2 dt = \int_{-\infty}^{\infty} \left| X(\omega) \right|^2 d\omega \quad (4.18) \]

where \(X(\omega) = \mathcal{F}[x(t)]\) is the continuous Fourier Transform.

\[ \sum_{n=-\infty}^{\infty} \left| x(n) \right|^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left| X(e^{j\omega}) \right|^2 d\omega \quad (4.19a) \]
\[ \sum_{n=0}^{N-1} \left| x(n) \right|^2 = \frac{1}{N} \sum_{k=0}^{N-1} \left| X(k) \right|^2 \quad (4.19b) \]

where \(n\) denotes the sample in time, and \(k\) denotes the frequency bin. Here we are defining the energy contained in each frequency band as:

\[ E_b \triangleq \frac{1}{N_b} \sum_{k=b_l}^{b_u} \left| X(k) \right|^2 \quad (4.20) \]

where \(b_l\) denotes the lower band, \(b_u\) the upper band and \(N_b\) the number of frequency bins in the particular band. Therefore the energy for each band, \(b\), in a Parseval sense is given by Eq. (4.20). The OBA and \(\frac{1}{3}\) OBA are given in Fig. 4.4 for a sample speech utterance.

4.2.3 Cepstrum

Cepstral analysis performs a deconvolution operation by transforming the product of two spectra into a linear summation using a logarithmic operation. As these two signals are sufficiently different spectrally, that is the excitation signal and the vocal tract response, they can be separated using linear filtering [218]. Cepstrum analysis is a form of non-parametric homomorphic

\(^6\)Defined by Parseval in [224], signal energy can be stated as the sum or integral of the square of a function is equal to the sum or integral of the square of its transform.
Speech Analysis and Feature Extraction

Figure 4.4: On the left hand side is the SVR and on the right hand side is the NVR. The speech utterance is "Chief interest in radiosterilization resides in the military services," from the TIMIT Corpus and shown in (a & b). The feature profiles below the speech utterances are as follows: (c & d) Octave Band Analysis (OBA) (e & f) 1/3 Octave Band Analysis (1/3 OBA).

filtering. An alternative parametric approach to this deconvolution process is called linear prediction analysis and will be described in the next section. There are two main variants on the Cepstrum methods which are the Real Cepstrum (RC) and Complex Cepstrum (CC). The RC is the most widely used method. In an effort to aid comprehension and clarity Bogart et al. [39], the proposers of the technique introduced some new terminology\footnote{Their motivation for the coinage of this rather odd terminology was succinctly summarized by Bogert et al, "In general, we find ourselves operating on the frequency side in ways customary on the time side and vice versa." [39].} such as [sic]: Cepstrum, quefrency, and liftering; see the table of terms in the Appendix, Table 9.1. In particular the three terms mentioned have been widely adopted by the signal processing community and hence have entered the Digital Signal Processing (DSP) lexicon and industry parlance [216, 71, 307, 225].
Table 4.1: Octave and 1/3 Octave band filter banks. Note: the centre frequencies are in terms of the geometric centre as defined in Eq. (4.17c).

<table>
<thead>
<tr>
<th>Lower cutoff frequency (Hz)</th>
<th>Centre frequency (Hz)</th>
<th>Upper cutoff frequency (Hz)</th>
<th>Lower cutoff frequency (Hz)</th>
<th>Centre frequency (Hz)</th>
<th>Upper cutoff frequency (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>37.5</td>
<td>53</td>
<td>75</td>
<td>20</td>
<td>30</td>
<td>45</td>
</tr>
<tr>
<td>75</td>
<td>106.1</td>
<td>150</td>
<td>71</td>
<td>79.9</td>
<td>90</td>
</tr>
<tr>
<td>150</td>
<td>212.1</td>
<td>300</td>
<td>142</td>
<td>159.9</td>
<td>180</td>
</tr>
<tr>
<td>300</td>
<td>424.3</td>
<td>600</td>
<td>284</td>
<td>319.7</td>
<td>360</td>
</tr>
<tr>
<td>600</td>
<td>848.5</td>
<td>1200</td>
<td>568</td>
<td>639.5</td>
<td>720</td>
</tr>
<tr>
<td>1200</td>
<td>1697.1</td>
<td>2400</td>
<td>1136</td>
<td>1279.0</td>
<td>1440</td>
</tr>
<tr>
<td>2400</td>
<td>3394.1</td>
<td>4800</td>
<td>2272</td>
<td>2558.0</td>
<td>2880</td>
</tr>
<tr>
<td>4800</td>
<td>6928.2</td>
<td>10000</td>
<td>4544</td>
<td>5116.0</td>
<td>5760</td>
</tr>
</tbody>
</table>

Before proceeding further to define the Cepstrum mathematically, it is important to present some background theory on homomorphic systems which forms the fundamental theory behind the approach. Homomorphic systems are a class of non-linear systems that conform to the general superposition principle\(^8\) [98,234]. They are of particular interest for many speech processing

\(^8\)See Eqs. (9.1), (9.2) and (9.3) in the Appendix for more on the general principle of superposition.
systems [234] as they provide us with a valuable method by which the excitation signal and vocal tract response may be separated. As previously stated the linear production of speech can be modeled as

\[ s(n) = g(n) \ast v(n) \]  

(4.21)

where \((\ast)\) denotes convolution, \(g(n)\) and \(v(n)\) correspond to the excitation signal and the vocal tract response respectively.

The system \(D_\ast[\cdot]\) is referred to as the characteristic system for homomorphic deconvolution. This is given by,

\[ D_\ast[s(n)] = D_\ast[g(n) \ast v(n)] \]

\[ = D_\ast[g(n)] + D_\ast[v(n)] \]

\[ = \hat{g}(n) + \hat{v}(n) = \hat{s}(n). \]  

(4.22)

(4.23)

(4.24)

Further, the z-transform of Eq. (4.21) is given by

\[ S(z) = G(z) \cdot V(z) \]  

(4.25)

Taking the logarithm of \(S(z)\) produces,

\[ \hat{S}(z) = \log[S(z)] = \log[G(z) \cdot V(z)] \]  

(4.26)

\[ \hat{S}(z) = \log[G(z)] + \log[V(z)] = \hat{G}(z) + \hat{V}(z) \]  

(4.27)

Thus the Cepstrum approach effectively capitalizes upon these two mathematical properties:

1. Convolution in the time domain is equal to multiplication in the frequency domain,

2. The logarithm of a product can be broken down into the sum of the logarithms of the individual terms.

This obeys the superposition principle and is a special case of the general class of methods referred to collectively as homomorphic signal processing [71]. It was original work in this area by Bogert et al. that lead to the introduction of Cepstrum analysis [39].

The RC is defined as:

\[ c(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log | X(e^{j\omega}) | e^{jwn} d\omega \]  

(4.28)

which is the inverse transform of the logarithm of the magnitude of Fourier Transform. The RC, \(c(n)\) is defined as the even part of \(\hat{x}(n)\), where:

\[ \hat{X}(e^{j\omega}) = \log(X(e^{j\omega})) = \log | X(e^{j\omega}) | + j \text{arg}[X(e^{j\omega})] \]  

(4.29)
4.2. Short-term Processing of Signals

and since the magnitude is real and even, and the phase is imaginary and odd, the phase information can be discarded [218]. Then expressing \( c(n) \) in terms of the DFT yields:

\[
c_d(n) = \frac{1}{N} \sum_{k=0}^{N-1} \log |X(k)| e^{\frac{j2\pi kn}{N}} \quad \text{for } n = 0, 1, \ldots, N - 1
\]

(4.30)

It is worth noting that historically when the RC was first proposed it was defined on the Power spectrum as opposed to the magnitude spectrum as presented here. Ideally the Cepstrum should represent a transformation on the speech signal with two properties: (i) the convolutionally mixed component signals after transformation will be separated in the Cepstrum, and (ii) the component signals after transformation will be linearly combined in the Cepstrum [71].

4.2.3.1 Mel-Frequency Cepstral Coefficients

The MFCC technique was introduced by Davis and Mermelstein [68], and exploits perceptual based auditory principles as well as the decorrelating property of the Cepstrum [243] through the discrete cosine transform. The stages in the computation of the MFCC features are represented in Fig. 4.5. Accordingly, as can be inferred from their name, these Mel-Frequency Cepstral Coefficients (MFCC) are an extension to the Cepstral analysis technique. This is certainly true to the extent that conventional Cepstral Analysis can be regarded as the forefather to MFCC based approaches. It extends on the approach by taking into account the human non-linear perception of pitch derived from studies in the field of psychoacoustics [71]. The inclusion of this perceptual phenomenon is achieved by non-linearly warping the spectrum into Mel-scale bands based on the human perception of pitch; see the curve shown in Fig. 4.6. This curve is described by the formula given in Eq. (4.31). In the field of psychoacoustics the term *Mel* is a unit of measure of the perceived pitch or frequency of a tone [71]. This *Mel* unit does not correspond linearly with the physical frequency of a tone, as the perceptive manner of pitch by the human auditory system does not have a linear relationship [294,71]. It was originally experimental work by Stevens and Volkman (1940) [294] that led to the derivation of Mel-scale. They designated 1000 Hz as being to equal 1000 Mels. They used this as their basis or reference point to perform a mapping between the real frequency scale in (Hz) and the non-linear perceived scale (Mels). This was achieved empirically through experimentation. They used this initial reference point to begin with and requested listeners to increase the frequency until they reached the point where they perceived it as having doubled in frequency, then trebled, and 10 times and so on. Accordingly, they labeled these pitches 2000 mels, 3000 mels and 10000 mels, respectively. This Hz-to-Mel relationship is illustrated in the curve in Fig. 4.6. A formula that has been approximated [247,234] to describe this relationship is given below in Eq. (4.31).

\[
f_{\text{mel}} = 2595 \log 10(1 + \frac{f}{700})
\]

(4.31)

\(^9\)The term *Mel* is derived from the musical word *Melody* [264].
The procedure for computing the MFCC is as follows: firstly, the speech signal is windowed by an analysis window $w(m)$, similar to the one previously described in the short-time analysis section. The discrete STFT spectrum is computed from this window of speech, denoted $X(n, \omega_k)$ as defined by,

$$X(n, \omega_k) = \sum_{m=-\infty}^{\infty} x(m)w(n - m)e^{-j\omega_km}$$  \hspace{1cm} (4.32)

where the notation has been simplified to aid clarity such that $\omega_k = \frac{2\pi}{N}k$ with $N$ the STFT length. The absolute value of the STFT is computed, as $|X(n, \omega_k)|$. This signal is then weighted by a series of triangular filter banks where the centre frequencies and bandwidths effectively approximate (model) the auditory critical band filters. These filters follow the non-linear Mel-scale which was defined in Eq. (4.31) and shown in Fig. 4.6.

As depicted in Fig. 4.7 the filters are triangular in shape and are normalized in accordance with their associated bandwidths. The next stage in the process computes the energy of each of the mel-scale bands; see Eq. (4.33). The $l^{th}$ filter response is denoted $\psi_l(\omega_k)$. 
4.2. Short-term Processing of Signals

\[ E_n(n, l) = \frac{1}{A_l} \sum_{k=L_1}^{L_l} | \phi_l(\omega_k)X(n, \omega_k) |^2 \]

(4.33)

where the term \( A_l \) is a normalization factor to account for the varying bandwidths thus to give an equal energy distribution for a flat spectrum [243].

The RC associated with \( E_n \) is computed as follows:

\[ C_{mel}(n, m) = \frac{1}{R} \sum_{l=0}^{R-1} \log[|E_n(n, l)|] \cos\left(\frac{2\pi lm}{R}\right) \]

(4.34)

where \( R \) denotes the number of filters. Typically around 24 bands are considered. An example of MFCC features for a speech utterance is given in Fig. 4.8, for both SVR (reference) and NVR (query) signals. Over the last two decades there has been wide spread adoption of these MFCC features within many speech processing systems, in particular speech recognition. This is attributed to their robustness and to the performance gains that are achieved when they are incorporated into the front-end processing stages. There are various subtle variations on the approach of some of the implementation details of computing the MFCC features. A good review is provided by Zheng in [336]. More recently Chakroborty et al. have proposed a somewhat counter-intuitive variation on the technique, the Inverted-MFCC [57], where the non-linear filter bank structure is configured in the reverse order, i.e. from sparse to dense frequency bands.

4.2.4 Linear Predictive Analysis

The basic principles of Linear Predictive (LP) Analysis are presented first, before moving on to describing the Perceptual Linear Prediction (PLP) features. LP analysis uses the principle that a speech sample can be predicted by a linear combination of some \( p \) past speech samples. The parameter coefficients are the weights which define the linear combination [246]. LP analysis has found many applications in speech processing where it can provide an accurate and economical representation of speech parameter [243, 218, 234, 246]. There are many applications where LP
Speech Analysis and Feature Extraction

Figure 4.8: On the left hand side is the SVR and on the right hand side is the NVR. The speech utterance is “Chief interest in radiosterilization resides in the military services,” from the TIMIT Corpus and shown in (a & b). The feature profiles presented in (c & d) are for Mel-Frequency Cepstral Coefficients (MFCC).

analysis is useful in areas of speech processing, three such areas include: (i) speech coding to reduce the transmission and storage of data, (ii) speech recognition to increase the accuracy, reduce computational complexity and increase robustness, and (iii) speech synthesis for increasing the efficiency and quality of synthetic speech.

Recalling the simplified model of speech production that was presented in Chapter 2, Section 2.3.6 in Fig. 2.8, this gave the all-pole z-transform as,

\[ H(z) = \frac{S(z)}{U(z)} = \frac{G}{1 - \sum_{k=1}^{P} a_k z^{-k}} \]  \hspace{1cm} (4.35)

This transfer function \( H(z) \) was defined by the following parameters: the switch between voiced and unvoiced excitation, the pitch period, the gain parameter \( G \) and the filter coefficients \( a_k \). The voiced signal is represented by an impulse train and the unvoiced signal is modelled as a random noise generator [218,71,246]. The signal \( s(n) \) is described by a difference equation in relation to \( u(n) \) which is given as,

\[ s(n) = \sum_{k=1}^{P} a_k s(n - k) + Gu(n) \]  \hspace{1cm} (4.36)
A linear predictor can be defined as:

\[ s(n) = \sum_{k=1}^{p} \alpha_k s(n - k) \quad (4.37) \]

and the system function for a \( p \)th order linear predictor is represented by the polynomial:

\[ P(z) = \sum_{k=1}^{p} \alpha_k z^{-k} \quad (4.38) \]

The term prediction effectively equates to an estimation and therefore there will be some associated prediction error \( e(n) \), which is defined as,

\[ e(n) = s(n) - s(n) = s(n) - \sum_{k=1}^{p} \alpha_k s(n - k) \quad (4.39) \]

From Eq. (4.39) it can be deduced that the error signal can be thought of as being the output from a transfer function \( A(z) \) as given in Eq. (4.40) with input \( S(z) \)

\[ A(z) = 1 - \sum_{k=1}^{p} \alpha_k z^{-k} \quad (4.40) \]

Noting the similarities of the term on the right hand side of Eq. (4.39) and in Eq. (4.36), if \( a_k \) and \( \alpha_k \) are equal then the error signal will reduce to \( e(n) = Gu(n) \). Therefore the transfer function \( A(z) \) can effectively be thought of as being an inverse filter for the original transfer function that was defined back in Eq. (4.35). Hence we can represent \( H(z) \) in terms of \( A(z) \):

\[ H(z) = \frac{S(z)}{U(z)} = \frac{G}{A(z)} = \frac{G}{1 - \sum_{k=1}^{p} \alpha_k z^{-k}} \quad (4.41) \]

The challenge of LP analysis is to find predictor coefficients \( \{\alpha_k\} \) for \( k = 1, 2, \cdots, p \) that allow the system \( H(z) \), as defined in Eq. (4.41), to be modelled. The approaches for computing the predictor coefficients try to minimize the mean square error for a short-time frame of speech. There have been many approaches proposed for computing the LP coefficients. Some of the major ones are mentioned in [246] and these include: (a) the covariance method (b) the autocorrelation formulation (c) the lattice method (d) the inverse filter formulation (e) the spectral estimation formulation (f) the maximum likelihood formulation and (g) the inner product formulation. These methods for computing LP coefficients are described in many text books on speech signal processing. An examination of each of these techniques is beyond the scope of this discussion. The interested reader is referred to [243, 218, 71, 225, 246]. Before proceeding something that has been left out of the discussion is the effect of the order \( p \) on the LP analysis. Essentially if \( p \) is high enough the model can provide good representations of most speech sounds. The diagram in Fig. 4.9 gives a good indication of how the model order \( p \) affects the resolution of the spectral detail in the spectrum. As the order of \( p \) increases it is observed that
Figure 4.9: The plots illustrated above are for a segment of a stationary vowel sound (voiced speech). They show the effects the model order $p$ has on the spectrum. In the plots: (a) $p$ is equal to 6, (b) $p$ is equal to 14, (c) $p$ is equal to 24 and in (d) $p$ is equal to 128. Notice the effects that increasing the order $p$ has on the spectrum. After Quatieri [243].

the underlying fine harmonic structure becomes more pronounced and evident in the spectrum. This spectral smoothing property is often exploited in preprocessing feature extraction stages to effectively smooth out the harmonic structure details in the spectrum. One such feature extraction technique is known as PLP which uses a relatively low order autoregressive all-pole model for the purpose of smoothing the spectrum. The PLP feature extraction technique shall be described now.

4.2.4.1 Perceptual Linear Prediction Features

Perceptual Linear Prediction (PLP) is a speech signal analysis approach that was originally proposed by Hermansky [111] in 1990. Similar to MFCC, the approach has found widespread use in many speech processing systems in particular speech recognition [209,113]. These features
are similar to MFCC, as they also take into account many perceptual auditory effects; see Fig. 4.10. The procedure for computing the PLP features is as follows:

The speech signal is windowed and the Short-time Fourier Transform is obtained. The power spectrum is obtained by taking the square of $S(\omega)$ given by,

$$P(\omega) = |S(\omega)|^2 = a(\omega)^2 + b(\omega)^2$$

(4.42)

where $a(\omega)$ and $b(\omega)$ are the real and imaginary parts of $S(\omega)$.

The rather fine spectral detail present in the power spectrum$^{10}$ $P(\omega)$ is reduced down into a much lower dimension, which is referred to as the critical-band resolution [209,114,111]. This is achieved as follows. Firstly, the frequency axis of the power spectrum $P(\omega)$ is warped into the non-linear Bark-scale frequency [274], $\Omega$ as given by,

$$\Omega(\omega) = 6 \ln\left(\frac{\omega}{1200\pi} + \left[\left(\frac{\omega}{1200\pi}\right)^2 + 1\right]^{0.5}\right)$$

(4.43)

This non-linear warped power spectrum is then convolved with the masking curve of the critical-band from Fletcher [85]. This critical band has been approximated [111] to give the piece-wise shape described by,

$$\Psi(\Omega) = \begin{cases} 0, & \text{for } \Omega < -1.3, \\ 10^{2.5(\Omega+0.5)}, & \text{for } -1.3 \leq \Omega \leq -0.5, \\ 1, & \text{for } -0.5 < \Omega < 0.5, \\ 10^{-1.0(\Omega-0.5)}, & \text{for } 0.5 < \Omega < 2.5, \\ 0, & \text{for } \Omega > 2.5. \end{cases}$$

(4.44)

These critical bands are presented Fig. 4.11. This convolution is given in Eq. (4.45), which yields the critical band power spectrum.

$$\Theta(\Omega) = \sum_{\Omega=\Omega - 2.5}^{\Omega=\Omega - 1.3} P(\Omega - \Omega_i) \Psi(\Omega)$$

(4.45)

$^{10}$Referring to the fine detail of the discrete STFT spectral bins $k$, for $k(0: \frac{N}{2})$. Where $N$ is usually relatively large.
The next stage performs equal loudness pre-emphasis. This is to compensate for the non-linear sensitivity of human hearing over the audible frequency range [86], see Fig. 9.1 in the Appendix. This is achieved by taking the product of the sampled warped power spectrum $\Theta[\Omega(\omega)]$ with the equalization transfer function $E(\omega)$. This is denoted by $\Xi[\Omega(\omega)]$, and given by,

$$\Xi[\Omega(\omega)] = E(\omega)\Theta[\Omega(\omega)]$$

where

$$E(\omega) = \frac{[(\omega^2 + 56.8 \times 10^6)\omega^4]}{[(\omega^2 + 6.3 \times 10^6)^2(\omega^2 + 0.38 \times 10^9)(\omega^6 + 9.58 \times 10^{26})]}$$

The transfer function $E(\omega)$ as shown in Eq. (4.47) is an approximation [111] empirically designed to counteract these non-linear effects of human loudness perception, thus, effectively aiming towards equalizing the magnitude of the power spectrum in terms of loudness. For frequency plots of $E(\omega)$, see Fig. 9.2 in the Appendix.

This stage incorporates aspects of Steven's power law of hearing [293]. Cubic-root amplitude compression is a simple but effective measure to take the power law into consideration. This is defined as,

$$\Phi(\Omega) = \Xi(\Omega)^{1/3}$$

This compression stage coupled with the preceding stage of equal-loudness pre-emphasis works well together to address human perceptual phenomena. Moreover, they reduce the spectral magnitude variation, hence affording a low-model order in the all-pole modeling stage [114,111]. The inverse DFT is computed for the signal to give the autocorrelation function dual to $\Phi(\Omega)$.

---

Footnote 11: For cases where the Nyquist frequency is $< 5kHz$ the last term on the denominator, $(\omega^6 + 9.58 \times 10^{26})$, of Eq. (4.47) may be omitted.
4.2. Short-term Processing of Signals

Figure 4.12: On the left hand side is the SVR and on the right hand side is the NVR. The speech utterance is “Chief interest in radiosterilization resides in the military services,” from the TIMIT Corpus and shown in (a & b). The feature profiles presented in (c & d) are for Perceptual Linear Prediction (PLP) Features.

An autoregressive model is computed from the solution of linear (Yule-Walker) equations for a $p^{th}$ order all pole model. This operation results in a similar smoothing operation as previously described in the LP analysis section and illustrated in Fig. 4.9. Here, however, in the PLP technique instead of smoothing the STFT spectrum the technique is used to smooth the critical band spectrum. PLP features for a sample speech utterance are presented in Fig. 4.12. The two feature extraction techniques, MFCC and PLP are compared in Fig. 4.13. This block diagram highlights the commonality between the two approaches.

4.2.5 Dynamic Features

Temporal dynamic characteristics have also been proposed for use in the feature extraction process. These dynamic factors try to include some information about the temporal dependence amongst feature vectors [209]. This allows for information to be included into the features from longer time spans than the typical short-term frames which are usually in the region of 20msec. The two techniques explored as part of this research are delta features (differentiation based) and in relation to PLP feature extraction a Relative Spectra (RASTA) processing technique proposed by Hermansky [114].
### Figure 4.13: Comparing the stages of the two perceptual feature extraction techniques, PLP and MFCC, respectively.

#### 4.2.5.1 Delta features

Delta features, first proposed by Furui [91], effectively try to describe the time trajectories of speech parameters between frames (short-term segment), i.e. from frame to frame [209]. There
are different approaches to computing the first order time derivative of a signal [188]. Two delta approaches are given by,

\[ B \Delta_i'(n) = \frac{\partial}{\partial t} x_i(n) = x_i(n) - x_i(n - 1) \quad (4.49) \]

and

\[ F \Delta_i'(n) = \frac{\partial}{\partial t} x_i(n) = x_i(n + 1) - x_i(n) \quad (4.50) \]

these known as the **backward** and **forward** differences, respectively. Another approach is defined as,

\[ S \Delta_i'(n) = \frac{\partial}{\partial t} x_i(n) = \sum_{w=-N}^{N} w x_i(n + w) \quad (4.51) \]

This has more favourable properties as it smooths by averaging over ±N samples. The effect of this smoothing is desirable as it alleviates the inherent noise due to the sharp discontinuities generated during the differentiation process [188]. A similar approach is presented by Lei et al. [167].

In a similar way, these equations are reapplied a second time to the output from the first order derivative to approximate the second order derivatives. These features are an important extension on straight-forward instantaneous features as they uncover important information relative to the surrounding frames (short-term segments) [209]. The significance of this extra information becomes evident from the increases in performance benefited by systems that incorporate dynamic features. Hence the reason why they form an integral part of most state of the art speech processing systems, [209,71].

### 4.2.5.2 Relative Spectra: termed RASTA

The RASTA technique was proposed by Hermansky in [114]. The approach is similar in some respects to the delta processing method previously described in that it filters the time trajectories of the feature representations. The process involves applying a non-linear operation, then filtering the signal, which is followed by applying the inverse non-linear operation. This non-linear operation is typically logarithmic with the inverse being the exponential [209,113,114]. The bandpass filter typically has a pass-band at around 1 – 12Hz. The RASTA technique effectively performs a reduction in the effects of spectral colouration with near-constant or slow varying properties. Furthermore, it has also been found to have some relationships with temporal properties of human hearing [209,113].

### 4.2.6 Time Domain Features

The literature contains many time domain based features that effectively characterize the speech signal across time. The two notable features from the literature are Energy and Zero Crossing Rate (ZCR).
4.2.6.1 Energy Signals

Energy measures are a popular speech feature. Their variants include the square, absolute value, and another variation is by taking the logarithm [241]. The energy measure given here is of the (square type), which is the summation of each signal value squared within the given frame duration. The formula to compute the energy is given in Eq. (4.52). This measure has higher values for voiced speech than unvoiced speech due to speech being characterized as having more energy for the case of voiced speech.

\[ E_i = \frac{1}{N} \sum_{n=0}^{N-1} |x_i(n)|^2 \]  

(4.52)

4.2.6.2 Zero Crossing Rate: ZCR

The zero crossing rate (ZCR) can be thought of as a measure of the noisiness of the signal [307]. As unvoiced speech signals tend to be characterized as being more noise-like, they usually have a higher ZCR, the reason being that voice speech has most of its energy distributed below 3kHz, whereas with unvoiced speech most of the energy is concentrated in the higher frequencies [246]. The formula to compute the ZCR is given by.

\[ Z(i) = \frac{1}{2N} \sum_{n=0}^{N-1} | \text{sgn}[x_i(n)] - \text{sgn}[x_i(n-1)] | \]  

(4.53)

with

\[ \text{sgn}[x_i(n)] = \begin{cases} 1, & x_i(n) \geq 0 \\ -1, & x_i(n) < 0. \end{cases} \]  

(4.54)

where the signal is usually normalized to remove the mean before calculating the zero crossing rate [33].

Some example Energy and ZCR feature profiles are shown in Fig. 4.14 for sample speech utterances (SVR and NVR). Often these two measures of Energy and ZCR have been used in conjunction with one another for the purpose of voiced/unvoiced speech classification or endpoint detection. A classic example of this was presented by Rabiner and Sambur in [245].

4.2.6.3 Teager Energy Features

Boland and Kavanagh have proposed two forms of univariate features [140]. These features have been derived from Teager’s Energy Operator, which was described previously. The features are known as the TEO Weighted Harmonic Product and the TEO Weighted Harmonic Sum, termed TEO-WHP and TEO-WHS. They provide a simultaneous characterization of both the pitch and formant content of a speech signal. This produces features which effectively extract the level of periodicity and aperiodicity content present in a particular frame using non-linear energy estimates. The features were proposed for classifying speech into harmonic (strong stable
Figure 4.14: On the left hand side is the SVR and on the right hand side is the NVR. The speech utterance is “Chief interest in radiosterilization resides in the military services,” from the TIMIT Corpus and shown in (a & b). The speech features shown are (c & d) Energy, and (e & f) Zero Crossing Rate (ZCR).
Speech Analysis and Feature Extraction

The features have demonstrated themselves to perform well for this specific harmonic classification task. These features have not been considered for use within the speech segmentation system as their formulation does not lend them to be appropriate for this specific task. An example of the features for the classification of speech into harmonic and non-harmonic speech segments is presented in Fig. 4.15 with Table 4.16 as a reference.

Figure 4.15: The different features of TEO-WHP and TEO-WHS are presented here. The speech signal is indicated in blue in panel 1 with the associated manners of articulation indicated below in textures w.r.t. Fig. 4.16. The classification for harmonic speech is given in the second panel for the different features with additional features Short Time Spectrum (Sts) Sts-WHP and Sts-WHS, which are equivalent features without performing the Teager energy spectrum transformation. The feature profiles are presented in panels 3 and 4.

<table>
<thead>
<tr>
<th>Speech manner</th>
<th>Texture marker</th>
<th>Phonetic Symbols (for TIMIT)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stops</td>
<td>b, d, g, p, t, k, dx, q</td>
<td></td>
</tr>
<tr>
<td>Closure</td>
<td>bel, del, gel, kel, pel, tel</td>
<td></td>
</tr>
<tr>
<td>Affricates</td>
<td>b, ch</td>
<td></td>
</tr>
<tr>
<td>Fricatives</td>
<td>s, sh, z, zh, f, th, v, ch</td>
<td></td>
</tr>
<tr>
<td>Nasals</td>
<td>m, n, ng, en, em, e, mg, nx</td>
<td></td>
</tr>
<tr>
<td>Semivowels and Glides</td>
<td>l, r, w, y, th, hw, el</td>
<td></td>
</tr>
<tr>
<td>Vowels</td>
<td>iy, ï, eh, ey, ay, as, aw, ay, ah, ao, oy, ow, uh, uw, ux, er, æ, ax, iæ, axh</td>
<td></td>
</tr>
<tr>
<td>Others</td>
<td>pau, api, h#</td>
<td></td>
</tr>
</tbody>
</table>

Figure 4.16: The classes for different manners of articulation.
4.3 Principal Component Analysis

Data analysis techniques can provide us with useful ways of uncovering hidden information from the data. In particular, they are extremely important for analyzing high dimensional data sets. Broadly speaking, they offer ways of extracting more relevant or significant information that may be concealed within the data and facilitate feature dimensionality reduction. All of this is achieved by finding significant statistical patterns that are present in the available data. Three predominant data analysis techniques are Linear Discriminant Analysis (LDA), Principal Component Analysis (PCA) and Independent Component Analysis (ICA). These techniques are interrelated as they try to find linear combinations that best describe the data. LDA is concerned with dealing directly with discrimination amongst classes, whereas PCA derives principal components for the whole data, without taking into account the underlying class structure [191]. The ICA technique was developed initially for use in blind source separation. The technique involves finding directions in the distribution of data that are less Gaussian, under the supposition that classification is easier to perform in these directions [153]. Specifically for the purpose of this research problem, that is the segmentation of speech signals, PCA presents itself as the more appropriate data analysis technique to consider. The reasons why this is the case will become more evident from understanding PCA and from also understanding the nature of the segmentation algorithm described in Chapter 5.

PCA is concerned with analyzing data which comprises of observations described by multiple dependent variables. These variables are generally cross-correlated hence the objective is to extract the important information that is present within the data and to re-express the data in terms of new orthogonal vectors called principal components [3]. PCA has widespread use in the greater field of signal processing, however the technique is rarely used by researchers in speech and audio domains [153]. One exception to this is where the approach was used within an audio-visualization tool called TimbreGrams which was proposed by Tzanetakis and Cook [314]. This approach used the first three principal components to characterize the audio signal. These three components were then used to form the basis for a mapping of the signal to the colour spaces, e.g. RGB or HSV, respectively.

Although many foundations of the PCA approach were laid down by his predecessors\textsuperscript{12}, the method was first formalized by Hotelling in 1933 [122], who first coined the terminology Principal Components [3]. The theory involves computing the major directions of variance of the data points in the n-dimensional space [22]. This is achieved by performing eigen-decomposition of the covariance matrix to give the eigenvector span and corresponding eigenvalue magnitude [36]. An underlying assumption exists here, to the effect that the greater the variance is in a particular direction the more relevant or significant the information that it carries or possesses. The first principal component represents the direction of the greatest variance and similarly the

\textsuperscript{12}This includes early work by Cauchy [56,99], Jordan [134], Pearson [229], also important contributions of Cayley, Silverster and Hamilton are referenced by Stewart [295].
second component corresponds to the direction of the second greatest variance, and so on. It is important to reiterate the property that the principal components are mutually orthogonal to one another [3,22]. This is significant as they provide the basis to perform the feature transformation which yields a more optimal representation from a pattern recognition point of view.

4.3.1 Motivation and Goals behind PCA

The motivation here is to perform a suitable transformation of the data such that the transform domain features exhibit high information packing properties compared to the original features [307]. This means that the more significant or relevant information is compressed into a smaller number of features, affording a desirable reduction in the dimension of the feature space, $\mathbb{R}^j \to \mathbb{R}^f$ for $f \leq j$. This acts to effectively remove any information redundancies that are usually present within the features. This is particularly true for spectral speech features which inherently contain high degrees of correlation amongst features. This is clearly evident from a quick observation of a speech spectrogram. Referring back to Chapter 2, Fig. 2.7, we can see the many relationships, interdependencies and inter-correlation amongst frequency bins against time. Consequently, it is desirable to generate speech features that are mutually uncorrelated $E[y(\alpha)y(\beta)] = 0$, for $\alpha \neq \beta$ [307,22]. This is so as to avoid information redundancies which would effectively encumber the overall performance of the segmentation algorithm.

The goals of PCA can be succinctly described as follows,

1. **Extract** from the data the most significant and important information,
2. **Reduce** information redundancies to simplify the description of the data,
3. **Analyze** the structure of the features over the observations,
4. **Reduce** the feature space dimensionality $\mathbb{R}^j \to \mathbb{R}^f$ for $f \leq j$. PCA finds a linear subspace of dimension or order $f$ such that the data linearly projected onto it have maximum variance, [3,36,22,301].

Also it is important to be aware of some important assumptions that PCA makes on the data in order to better understand its limitations. Some of these assumptions include:

1. **Linearity.** PCA assumes a straightforward linear transformation via a change of basis. However, other approaches exist such as Kernel PCA [301,74], where the input data is assumed to have a non-linear structure and is therefore first transformed into a higher-dimensional feature space which has a linear structure where it is then proceeded by a PCA stage.
2. **Gaussian distribution.** Such that the mean and variance are the statistics that fully describe the data.
3. The importance of the dynamics of large variances effectively assumes that the data has a high signal to noise ratio (SNR). Accordingly, the larger the variances the greater the significance or importance from an information carrying point of view and likewise the lower the variance the more noise like it is [283].

4.3.2 The PCA Procedure

Let $\mathbf{X}$ denote the feature matrix consisting of observation frames $i$, where $(1 \leq i \leq I)$, which are described by dependent feature vectors $\mathbf{v}_j$, where $(1 \leq j \leq J)$. Thus producing a $I \times J$ matrix with rank $L$ where $L \leq \min(I, J)$. The generic elements of $\mathbf{X}$ are simply referenced by $x_{i,j}$; see Eqs. (4.55a) and (4.55b).

$$
\mathbf{v}_j^T = [v_{j}(1), v_{j}(2), \cdots, v_{j}(I)]
$$

$$
\mathbf{X}_{I,J} = (\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \cdots, \mathbf{v}_J) = \begin{pmatrix}
x_{1,1} & x_{1,2} & \cdots & x_{1,J} \\
x_{2,1} & x_{2,2} & \cdots & x_{2,J} \\
\vdots & \vdots & \ddots & \vdots \\
x_{I,1} & x_{I,2} & \cdots & x_{I,J}
\end{pmatrix}
$$

The elements of $\mathbf{X}$ are mean adjusted (centred) so that the mean of each column (feature vector) is equal to zero. Therefore $\mathbf{X}^T \mathbf{1} = \mathbf{0}$, where $\mathbf{1}$ is a $(I \times 1)$ vector of ones and $\mathbf{0}$ is a $(J \times 1)$ of zeros. For example, mean adjusting the vector $\mathbf{v}_1$, is as follows $\bar{x}_{i,1} = x_{i,1} - \bar{x}_1$, for $(1 \leq i \leq I)$. The estimated covariance matrix $\mathbf{C}_X$ is obtained as shown in Eq. (4.56).

$$
\mathbf{C}_X = \frac{1}{(I-1)} \mathbf{X}^T \mathbf{X}
$$

The covariance matrices are positive semi-definite - the eigen-decomposition for these matrices always exists [2]. The eigenvectors, $\mathbf{u}$, and eigenvalues, $\lambda$, for $\mathbf{C}_X$ are obtained. The relationships between these are described as follows.

$$
\mathbf{C}_X \mathbf{u} = \lambda \mathbf{u}
$$

$$
(\mathbf{C}_X - \lambda \mathbf{I}) \mathbf{u} = \mathbf{0},
$$

where $\mathbf{I}$ is the identity matrix.

Furthermore, eigenvectors with corresponding different eigenvalues are orthonormal and can therefore be stored in an orthonormal matrix. Recalling that the matrix is orthogonal when the product of this matrix by its transpose is a diagonal matrix [3] given as,

$$
\mathbf{U}^{-1} = \mathbf{U}^T,
$$

$$
\mathbf{C}_X = \mathbf{U} \Lambda \mathbf{U}^T \text{ with } \mathbf{U}^T \mathbf{U} = \mathbf{I}
$$
Speech Analysis and Feature Extraction

The diagonal matrix denoted $\Lambda$ contains the eigenvalues $\lambda$ across the diagonal and the matrix $U$ contains the normalized eigenvectors $u$. The eigen-pairs of $u$ and $\lambda$, $[(u_1, \lambda_1), \cdots, (u_j, \lambda_j)]$, are ordered in terms of their eigenvalue magnitude, $\lambda_1 \geq \cdots \geq \lambda_j$. The transformation matrix $G$ containing $f$ eigenvectors with the corresponding largest eigenvalues (i.e., the first $f$ eigenvectors in the ordered set), described as,

$$G = [u_1, \cdots, u_f],$$

$$X' = G^T X.$$  \hspace{1cm} (4.59a) \hspace{1cm} (4.59b)

The linear orthogonal transformation is obtained by taking the inner product of the transformation and data matrices, $G$ and $X$ respectively, see Eq. (4.59b). This produces a transform matrix $X'$ of $f$-dimension. A definitive criterion does not exist for determining a suitable threshold for the cut-off point for the number of dimensions $f$. However, many of the approaches are quite similar, which offers to a certain extent a degree of flexibility. One such approach is given by Kocsor [153] which is based around finding the appropriate $f$ value for the inequality defined in Eq. (4.60), for $\mathbb{R}^i \rightarrow \mathbb{R}^f$ where $f \leq j$.

$$\frac{\lambda_1 + \cdots + \lambda_f}{\lambda_1 + \cdots + \lambda_j} > 0.99$$ \hspace{1cm} (4.60)

In order to be able to visualize and gain a better appreciation of the process involved, two examples of the application of PCA on speech signals are provided in Fig. 4.17 for Synthesized Voice Recording (SVR) and in Fig. 4.18 for Natural Voice Recording (NVR), for the same utterance. The two speech utterances, SVR and NVR, are for the same sentence to enable comparison. The features used are the PLP feature Set which comprises of 9 PLP features with dynamic feature extensions of delta $\Delta$ and delta-delta $\Delta \Delta$ included, thus comprising of $j = 27$ features in total. These are presented in both figures in the top panel of Fig. 4.17, and Fig. 4.18. A dramatic roll-off in eigenvalue magnitude is observed for the two cases, see Fig. 4.17c and Fig. 4.18c. The eigenvectors and their corresponding eigenvalues are given in the centre panel. Using the ratio as given by Eq. (4.60) the features representation was reduced from $\mathbb{R}^{27} \rightarrow \mathbb{R}^{16}$. The resulting transformed and dimensionally reduced features are then given in the bottom panels.

### 4.4 Conclusion

This Chapter has introduced and critically appraised a number of different feature extraction and signal analysis techniques which have been investigated in this research. The Chapter began by discussing a key component of this, the rationale behind performing feature extraction, effectively answering the motivational question, as to why feature representations are needed
Figure 4.17: An example of Principal Component Analysis (PCA) applied to a PLP Set of 27 features as shown in plot (a) which are subsequently transformed and reduced to 16 features as shown in (d). The utterance is a Synthesized Voice Recording (SVR) of “Heave on those ropes the boat’s come unstuck”.
Figure 4.18: An example of Principal Component Analysis (PCA) applied to a PLP Set of 27 features as shown in plot (a) which are subsequently transformed and reduced to 16 features as shown in (d). The utterance is a Natural Voice Recording (NVR) of “Heave on those ropes the boat's come unstuck”, where the speech sample is from the TIMIT corpus.
4.4. Conclusion

to characterize speech signals. Various different signal analysis methods were then described, most of these being Short-time Fourier Transform based. The univariate features that have been evaluated are introduced, including Energy, ZCR, Spectral Centroid, Spectral Roll-off and Spectral Flux. The multivariate features that were investigated are Power Spectrum, Teager Energy Spectrum, Energy based Octave Band Analysis, MFCC and PLP features. The theory behind these features have been presented and examples of these features for different speech utterances have been provided. These feature representations are an integral pre-processing component for the segmentation algorithm to be introduced in Chapter 5. The Chapter has presented the feature dimensionality reduction technique known as PCA. The theory and the motivation behind PCA on the multivariate feature representations has been reported along with examples of the technique being applied to speech signals. This technique will be revisited in Chapter 5.
In this Chapter a methodology for speech segmentation is proposed which uses the Dynamic Programming (DP) Principle known as Dynamic Time Warping (DTW). The proposed algorithm is described in detail incorporating the various features and signal analysis techniques that were examined in Chapter 4. Important topics such as segmentation tolerances are discussed in relation to the end-applications as well as other experimentation details including the speech corpus that was used to test the algorithm. Finally, the results of the proposed methodology are presented, which demonstrate the effectiveness of the system as an automatic means of segmenting continuous speech signals at the word-level boundaries.

5.1 Proposed Methodology

An overview of the segmentation algorithm is firstly presented here in order to gain a better understanding of how the various blocks integrate. Later in the Chapter, a more detailed analysis is given for the processing stages. The proposed algorithm is presented in Fig. 5.1 with the various stages and signals indicated with blocks and signal plots. As shown in the diagram, the approach begins by pre-processing both a Synthesized Voice Recording (SVR), which forms the reference signal and a Natural Voice Recording (NVR), which forms the Query signal. The SVR signal is termed the reference signal as this speech signal is generated solely for the purpose of providing reference template profiles to assist in the segmentation process. This synthesized signal is generated from the text transcript (known a-priori) using state of the art AT&T voice
Speech Segmentation Using Dynamic Time Warping

engines\(^1\). In addition to the synthesized signal, the temporal boundary information at the word level is provided by the voice engine, i.e. the beginning and end points of each word in the audio stream. The NVR signal is termed the query signal as it is the actual speech signal that the algorithm performs the segmentation on.

While the SVR signal is acquired or generated using a speech synthesizer, the NVR is acquired either from a microphone or loaded from a pre-recorded audio file. The two signals are presented to the algorithm in digital format. In general, the two signals have the same sampling frequency which for the purpose of this research was fixed at 16kHz at 16 bits per sample. The next stage involves performing feature extraction on both the SVR (reference signal) and the NVR (query signal) independently. The different types of feature extraction and signal analysis techniques for speech signals have been described in detail in Chapter 4. These signal representations include the univariate features: Energy, ZCR, Spectral Centroid, Spectral Roll-off, Spectral Flux and multivariate (vector based) features: STFT, OBA, PLP, MFCC as well as dynamic feature extensions such as first order derivative (delta or \(\Delta\)), second order derivative (delta delta or \(\Delta\Delta\)) and RASTA filtering. The feature extraction stage produces feature representations that aim at best characterizing both the SVR and NVR signals across time. These features then proceed to the next stage where they undergo feature transformation and dimensionality reduction. This involves using the feature representations for both SVR (reference signal) and NVR (query signal) to perform a mapping to a new orthonormal basis set using Principal Component Analysis (PCA).

The methodology behind PCA has been covered in greater detail in the previous Chapter. The PCA procedure was previously described within the context of a more general implementation, operating on one data set. However, in this case it is described operating on two data sets simultaneously. The proposed approach necessitates two feature data sets to be used within the DTW methodology. This requires the role of the PCA procedure to integrate into the DTW process to reduce the feature dimensionality in order to best characterize the two signal types of the SVR and NVR with respect to one another. Effectively, the role of the PCA stage is to efficiently represent the two signals types in order to time-align them. This involves transformation of the feature representations onto new basis sets. The two signals SVR and NVR can be regarded as being similar in terms of the linguistic message, since the SVR signal contains the same spoken text as the NVR signal. They do however differ in terms of the temporal locations of where the words occur (either leading or trailing each other), the duration of the words, also in terms of the voice qualities such as pitch, prosody, intonation and stress. The next part in the procedure involves performing time warping on these features to time-align the two signals, which is achieved using DTW.

\(^1\)AT&T Natural Voices™ (voice engine) is a commercial Text-To-Speech Synthesizer, for the creation of audible speech from computer readable text [17].
5.1. Proposed Methodology

Figure 5.1: This diagram gives an overview of the proposed segmentation method.
The first stage of DTW involves computing a distance matrix using the feature representations from SVR and NVR. The distance between each point in the feature space of the SVR and NVR representations is computed to form the distance matrix. From this distance matrix a cumulative distance matrix is derived using a recursive relation. The optimal path is computed through the cumulative distance matrix using different criteria and constraints. This path essentially provides the warping information to time align the two signals together. As the boundary information for the SVR signal is known a-priori, therefore these points can be mapped across over onto the NVR signal. As these values are in feature space (frames), they are therefore subsequently mapped to the samples (time) to arrive at the segment boundary estimates for the sentence/utterance.

5.1.1 Defining the Problem

The proposed method uses a dynamic programming approach to time align two multivariate time series. These two time series are denoted in matrix form as $R$ and $Q$ of order $L \times I$ and $L \times J$, see Eq. (5.1). These matrices are composed of feature vectors, $r_i$ and $q_j$, for instances of time in frames, referenced by $i$ and $j$ respectively. The dimension $L$ denotes the maximum dimension of the feature vectors which corresponds to the total number of feature representations or profiles. It was assumed that the same number of $L$ features are extracted from the SVR and NVR to create time series.

Given this, $R$ and $Q$ are defined as,

$$R_{L,I} = (r_1, r_2, \ldots, r_i, \ldots, r_I)$$

$$Q_{L,J} = (q_1, q_2, \ldots, q_j, \ldots, q_J)$$

where individual feature vectors are $r_i = (r_1, r_2, \ldots, r_i, \ldots, r_L)^T$; and $q_j = (q_1, q_2, \ldots, q_j, \ldots, q_L)^T$, respectively.

The two times series are computed to derive an $I$-by-$J$ distance matrix or plane, where each of the points $d_{(i,j)}$ on the plane $D_{I,J}$ represents the multi-dimensional distance between the corresponding (pair-wise) feature vectors in $R$ and $Q$. The goal is to find an optimal time alignment between the two signals (times series) such that the distance between them is minimized. This is obtained by finding a warp path, through the plane $D_{I,J}$. The warp path $W$ is defined as,

$$W = (w_1, w_2, \ldots, w_k, \ldots, w_K)$$

where $k$ represents an individual path point and $K$ represents the total number of points in the warp path. The $k^{th}$ point within $W$ has indices of $(i, j)$, where the points are referenced by
5.1. Proposed Methodology

$w_k = (i_k, j_k)$ for $i_k \in \{1, 2, \ldots, I\}$ and $j_k \in \{1, 2, \ldots, J\}$ as shown in Eq. (5.2). This warp path is computed using the DTW procedure which is described below.

5.1.2 Dynamic Time Warping Approach

DTW is a feature matching technique that uses dynamic programming principles from the field of applied mathematics [71]. It is a technique that gained much popularity in early work in Isolated Word Recognition [244]. Although the term Time forms part of the technique’s name this is really only legacy, as it was initially applied to time series data. In many cases it applies to other domains such as for instance the spatial domain. The technique has found a role in many different areas in signal processing; some interesting recent applications include: Gas chromatography: for the alignment of time series data [63]; Clinical applications: for motor rehabilitation for post-stroke neurological patients [311]; and Video processing: a DTW based technique for redundancy removal for video summarization [149].

As previously stated the goal is to time warp the two signals (SVR and NVR) together to yield the best possible template match in time. This reduces to finding the optimal path through the cumulative distance matrix for the feature representations of the two signals, SVR and NVR. One naïve based approach is to search all the possible paths and, once complete, adopt the path which has the minimum cumulative distance. Although this approach is perfectly valid and works in practice, as a consequence of its naïve (brute-force) nature it is extremely computationally expensive, thus making it unfeasible to use in practice. In order to overcome these shortcomings, Dynamic Programming (DP) algorithms often exploit what is known as Bellman’s Optimality Principle\(^2\) to efficiently compute the optimal path. The application of Bellman’s optimality principle greatly reduces the computational load when finding the optimal path. The importance of this principle is fundamental to the theory of DP and subsequently to DTW. Accordingly, a brief description of Bellman’s optimality principle is provided here to better understand DTW as used in the proposed segmentation algorithm.

Let the nodes in the path $W$ be denoted as follows:

$$(i_s, j_t), (i_1, j_1), (i_2, j_2), \ldots, (i_u, j_v)$$

where the path start and end points correspond to:

$$\begin{align*}
(i_s, j_t) & \Rightarrow W(1, 1) \\
(i_u, j_v) & \Rightarrow W(M, N)
\end{align*}$$

5.1.2.1 Bellman’s Optimality Principle

The optimal path between the starting node $(i_s, j_t)$ and the final node $(i_u, j_v)$ is given as:

\(^2\)It was original work by Bellman [24] in the 1950’s which led to his Principle of Optimality, which he then applied to find solutions to problems in various different fields [26, 25, 24].
Speech Segmentation Using Dynamic Time Warping

\[ (i_s, j_t) \overset{\text{opt}}{\rightarrow} (i_u, j_v) \]

which passes through a point \((i_w, j_x)\)

\[ (i_s, j_t) \overset{\text{opt}}{\rightarrow} (i_u, j_v) \]

Bellman's principle\(^3\) specifies that the optimal path between nodes \((i_s, j_t)\) and \((i_u, j_v)\) is simply the optimal path between nodes \((i_s, j_t)\) and \((i_w, j_x)\) concatenated with the optimal path between nodes \((i_w, j_x)\) and \((i_u, j_v)\). This can be expressed mathematically as,

\[ (i_s, j_t) \overset{\text{opt}}{\rightarrow} (i_u, j_v) = (i_s, j_t) \overset{\text{opt}}{\rightarrow} (i_w, j_x) \oplus (i_u, j_v) \]  \hspace{1cm} (5.3)

where \(\oplus\) denotes a path concatenation operation.

The significance of this is that if the optimal path between nodes \((i_s, j_t)\) and \((i_w, j_x)\) is known, then determining the optimal path between nodes \((i_s, j_t)\) and \((i_u, j_v)\) only requires searching onwards from node \((i_w, j_x)\). Therefore there is no need to re-examine the other intermediary paths and it is a straightforward concatenation \((\oplus)\) of the two path segments as defined in Eq. (5.3).

5.1.2.2 Distance between Vectors

Obtaining the distance measure between vectors involves measuring how far apart (distance) two \(L\)-dimensional vectors are in a multidimensional space, denoted \(\mathbb{R}^N\). It involves the computation of a distance matrix for the SVR (reference) signal and NVR (query) signal. The distance \(d(\cdot, \cdot)\) for each point in the distance matrix \(D\) are real-valued. The distance function is governed by three important properties that define the measure of distance [71]. These are as follows (a) \(d(x, y) \geq 0\); (b) \(d(x, y) = 0\) if and only if \(x = y\) (c) \(d(x, y) \leq d(x, z) + d(z, y)\). The latter is referred to as the triangle inequality [324], when considering the notion of distance between points in a \(\mathbb{R}^N\) real Cartesian space these properties are quite intuitive. For a more in depth discussion on this topic the interested reader is referred to [47].

In this work three distances are of particular interest, these are (a) City block\(^4\), (b) Euclidean and (c) Chebyshev metrics. These three metrics are actually special cases of the Minkowski metric of order \(s\), also termed the \(l_s\) metric [71], see Eq. (5.4). In general the Euclidean distance metric is the most commonly used for engineering based problems. The reason for this is that it corresponds well with our intuitive notion of physical distance [71]. For two vectors in a Euclidean space, the distance between them is referred to as the natural distance. However if

\(^3\)Principle of Optimality. An optimal policy has the property that whatever the initial state and initial decisions are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decisions. [25].

\(^4\)The City block distance is also referred to as the Manhattan distance [324,116] and the Taxicab distance [70].
the vectors are represented in some non-Euclidean space\(^5\) the distance between them are referred to as *unnatural*. This is an important point which has significance in a section described later on PCA (data transformation).

Other distance metrics are present in the literature which include the Mahalanobis distance [335, 198, 181], Itakura (which is popular when using Linear Prediction (LP) Analysis) [126], Cosine distance (this measure relates to the angle between two vectors) [335], Correlation distance [6].

In this work however we define the distance measure \(d_s(x, y)\) as,

\[
d_s(x, y) = \sqrt[4]{\sum_{k=1}^{N} |x_k - y_k|^s}
\]

(5.4)

to define three special cases of the Minkowski metric where \(s = 1, s = 2\) and \(s = \infty\). The three different metrics are defined as follows,

- **City block** \(d_1(x, y) = \sum_{k=1}^{N} |x_k - y_k|\), case of \(s = 1\) (5.5a)
- **Euclidean** \(d_2(x, y) = \sqrt{\sum_{k=1}^{N} |x_k - y_k|^2} = \sqrt{(x - y)^T(x - y)}\), case of \(s = 2\) (5.5b)
- **Chebyshev** \(d_\infty(x, y) = \max_k |x_k - y_k|\), case of \(s \to \infty\) (5.5c)

Examples of these distance metrics, in the form of distance matrices \(D_{I,J}\) is given in Fig. 5.2 for the case of MFCC speech features. For the purpose of this study only the Euclidean distance will be used.

### 5.1.2.3 Cumulated Distance Matrix

Using the distance matrix \(D_{I,J}\) a cumulative distance matrix \(\Gamma_{I,J}\) is produced. This matrix is comprised of cumulative distances for each point denoted \(\gamma_{i,j}\). The use of cumulative distance points removes the need to re-calculate partial path distances, as the optimal path is found from tracing back through the matrix \(\Gamma_{I,J}\) [31].

The process to compute the cumulative distance matrix requires initial starting points along the top row and on the first column on the left hand side of \(\Gamma_{I,J}\). These specific points are computed as,

\[
\gamma_{i,1} = d_{i,1} + \gamma_{i-1,1}
\]

(5.6a)
\[
\gamma_{1,j} = d_{1,j} + \gamma_{1,j-1}
\]

(5.6b)

\(^5\)Note this is also the case for orthogonal basis vectors that are not orthonormal [71].
where Eqs. (5.6a) and (5.6b) refer to points along the first column and the first row of $\Gamma_{I,J}$, respectively. The cumulative distances at the other points $\gamma_{i,j}$ are given by the recursive relation:

$$\gamma_{i,j} = d_{i,j} + \min[\gamma_{i-1,j}, \gamma_{i-1,j-1}, \gamma_{i,j-1}]$$

(5.7)

where $d_{i,j}$ is the distance between the current elements at that specified point and second term $\min[\cdot]$ defines the minimum cumulative distance of the neighbouring points. Once the full matrix has been acquired the optimal path can be found by process of tracing backwards through the matrix.

5.1.2.4 Path Constraints

1. **Monotonicity** The points defined along the path $W$ should be monotonic, this requires that $i_{k-1} \leq i_k$ and $j_{k-1} \leq j_k$. Visually, this means that the path as will be defined
by the local constraints below is not permitted to go East or South-wards in direction. Monotonically ordering the points with respect to time prevents the warping path from reaching a point in the plane earlier in time that was already matched.

2. **Boundary Constraints** These boundary constraints require that the complete optimal path $W$ must start at the bottom right corner of the matrix, i.e. $\Gamma_{I,J}$, and finish at the opposite diagonal corner at the top left of the matrix, i.e. $\Gamma_{1,1}$.

3. **Global Constraints** Global constraints limit the number of elements within the cumulative distance matrix $\Gamma_{i,j}$ that need to be searched when finding the optimal path. Two commonly used constraints are presented in Fig. 5.3a. They are the Sakoe and Chiba band and the Itakura Parallelogram. For the case of the popular Itakura global constraint the maximum compression/expansion is by a factor of two. Furthermore if the two signals to be warped are relatively the same length, i.e. $I \approx J$, then the number of points to be searched is reduced by approximately one-third [307, 71, 265].

4. **Local Constraints** These local constraints define or govern the allowable predecessor points that the optimal path may take. Thus they effectively impose limits on the transition points at a local level and satisfy the condition of monotonicity. There have been many different local constraints defined in the literature most notably the Itakura and Sakoe and Chiba [307, 71, 265]. The local constraints used within this proposed segmentation algorithm employ one of Sakoe and Chiba's symmetric constraints as illustrated in Fig. 5.3b, which is defined below in Eq. (5.8).

### 5.1.2.5 Optimal Path

The optimal path $W$ needs to satisfy the constraints as outlined above. Recalling $W$ was defined earlier in Eq. (5.2). The $k^{th}$ point within $W$ has indices of $(i, j)$, where the points are referenced by $w_{(i_k,j_k)}$ for $i_k \in (1 \ldots I)$ and $j_k \in (1 \ldots J)$. As stated previously the optimal path is obtained by tracing back through the cumulative distance matrix $\Gamma_{i,j}$ which has the partial paths already computed. The initial starting point occurs at point $\Gamma_{I,J}$ and finishes at point $\Gamma_{1,1}$ using the recursive relation as given by,

$$w_{(i_k,j_k)} = \min[\gamma_{i-1,j}, \gamma_{i-1,j-1}, \gamma_{i,j-1}] \tag{5.8}$$

It is important to note that the minimization term in Eq. (5.8) does not return a magnitude distance value to $w_{(i_k,j_k)}$: it in fact returns the indices $(i, j)$ of the $k^{th}$ warping path point where the minimum distance value has been found for the local constraints in the matrix $\Gamma_{I,J}$. Aside, the overall distance of the path can be computed either by taking the sum or product, of the
Figure 5.3: These diagrams illustrate different types of constraints, (a) gives two global constraints: the Itakura’s Parallelogram in red and Sakoe and Chiba’s band in blue, (b) gives the local constraints used for the warping path which are defined mathematically in Eq. (5.8).

Cumulative distances in $\Gamma_{l,j}$ that the path passes through respectively as,

$$D = \sum_{k=1}^{K} \gamma_{i_k,j_k}$$  \hspace{1cm} (5.9a)$$

$$D = \prod_{k=1}^{K} \gamma_{i_k,j_k}$$  \hspace{1cm} (5.9b)$$

where the indices $(i_k,j_k)$ of $\gamma$ represent the $k^{th}$ element within the warp path $W$ for $i_k \in \{1,2,\ldots,I\}$ and $j_k \in \{1,2,\ldots,J\}$. Some examples of optimal path traces are presented in Fig. 5.4.

5.1.2.6 Mapping A-priori Boundaries

The optimal path provides important warping information so the two signals SVR and NVR can be time aligned. Using this warping information along with the a-priori knowledge of the word boundaries for the SVR signal the segmentation is achieved. These word boundary values for the SVR are mapped over onto the NVR as the two signals are temporally aligned; this is illustrated in Fig. 5.5. The values are in terms of feature frame indices which are then converted to discrete-time sample points, thus producing the segmentation estimates for the word boundaries on the NVR signal.
5.1. Proposed Methodology

Figure 5.4: These plots show the cumulative distance matrices $\Gamma_{f,J}$ with the optimal path traces $W$ included in black. The Euclidean distance metric with the PLP Set of feature representations was used. The three utterances are (a) “Beg that guard for one gallon of gas.” (b) “Few rural areas are protected by zoning.” (c) “Basketball can be an entertaining sport.” These speech samples are from the TIMIT Corpus.
Figure 5.5: This diagram shows the word boundary values for the SVR being mapped over onto the NVR, wherein the warping path is used to provide the projection information.
5.1. Proposed Methodology

On a practical note, in order to verify that the DTW procedure was working correctly, deterministic signals were used to test its operation. Deterministic signals such as sinusoidal and chirp signals were used for this purpose. These provided suitable test cases for the reference and query signals through making suitable adjustments to: (i) amplitude, (ii) frequency or (iii) phase; with respect to each another. This was valuable for testing as it provided important verification that the algorithm was functioning correctly and effectively before applying more complex signals such as feature vectors extracted from speech signals.

5.1.3 Feature Transformation and Subset Selection

Up until this point, the Feature Transformation and Dimensionality Reduction stage has been omitted from the algorithms procedure. Effectively, the approach has been described in terms of using all of the features in their current formats. In Chapter 4, in the section on the Philosophy of features, reasons why multiple features are needed to represent or characterize the speech signal were discussed.

It is desirable to be able to select the most relevant and significant features and still retain a strong characterization of the speech signal with respect to time. This is desirable as it reduces the computational complexity. Also if there is high mutual correlation amongst the feature representations there is not much extra information gained by the inclusion of additional redundant features. Furthermore as the next stage in the process involves DTW, the transformation of the two signals SVR and NVR and the reduction of redundant features could increase the performance of the DTW algorithm. The technique that has been explored to perform this transformation and dimensionality reduction is Principal Component Analysis (PCA).

PCA is a linear transformation technique that uncovers statistical patterns in data based on variance. The procedure has been described more formally in Chapter 4. In brief, eigen-decomposition is performed on the features to produce orthogonal basis vectors which are termed Principal Components. These components can be ranked in order of significance from the magnitude of their corresponding eigenvalues. The reduction of components is based on some form of thresholding scheme, referring back to Eq. (4.60). The work described here uses feature space representations for the SVR signal and the NVR signal respectively. This gives rise to the challenging problem of how to best perform the PCA on the two groups of feature representations such that they are in an appropriate format for the DTW procedure.

This problem has been encountered before where the same variables are being measured on objects from different groups, and the covariance structure varies from group to group [87]. In this case the two groups considered are, (a) speech synthesizer and (b) human speech. Flury [87] presents some other interesting examples where this might also arise in practice, such as (i) three species of iris (ii) male and female turtles (iii) human and animal bones and (iv) real and forged bank notes. Although the covariance matrices are not exactly identical, it sometimes seems reasonable to assume that the covariance matrices possess some common basic structure [87].
The three different approaches that have been explored in this research are as follows:

1. **Independent Vector Approach.** Perform eigen-decomposition for the two groups of feature representations independently to produce two sets of principal components for the two groups. Then perform transformation and dimensionality reduction independently on each of the two feature representations. Of course now the DTW is being applied to two vector sequences that have vectors whose components are different combinations of the original features.

2. **Signal Concatenation Approach.** Assume that the two feature groups have the same underlying data structure and straightforwardly concatenate the two groups of feature representations together. Then perform the eigen-decomposition on these concatenated feature representations to produce a form of combined set of principal components. These principal components are then used as the basis set for performing transformation and dimensionality reduction for both the SVR and NVR feature representations.

3. **Average Vector Approach.** Perform the two separate eigen-decompositions on the groups SVR and NVR independently, in a similar manner to the Independent Vector Approach, to produce two sets of principal components. Then the strategy involves combining these two groups of principal components to produce joint (average) vectors as principal components. A common approach has been to get an intermediate angle between the eigenvectors to form a new eigenvector. This is called Common Principal Component Analysis, see Krzanowski (1979) [159] and Flury (1984) [87]. This has been more recently extended upon by Yang et al. who proposes an algorithm which measures the similarity between Multivariate Time Series (MTSs) [332] and a feature subset selection technique called CLeVer, [333]. These two approaches work off the basis of manipulating the vector angle of the principal components for different data groups for measuring the similarity or feature subset reduction.

In order to visualize this procedure it is easier to consider the problem in two dimensions as higher dimensional spaces (> three) are harder to visualize and interpret. The two vectors as shown in Fig. 5.6 represent corresponding principal components from the time series of the SVR (reference series) and the NVR (query series). The term corresponding refers in a pairwise sense to the ordering or component ranking from both of the series. The principal components (vectors) are ranked in order of their associated variance, which, as mentioned in Chapter 4, is determined by the magnitude of the their eigenvalues. The objective is to find an intermediate set of vectors (components) denoted \( \tilde{PC}_A i \) where each vector comprises of an average two vectors of \( PC_{Ri} \) and \( PC_{Qi} \).
5.2. Acceptable Tolerances

Before proceeding with the results section it is important to put the desired requirements of the segmentation algorithm into context. In particular, what tolerance values are appropriate for the proposed system? It is important to quantify this so as to put bounds on the segmentation problem. As one would expect the acceptable segmentation tolerance varies depending on the nature of the end-application. For instance, the segmentation of phonemes for the purpose of corpus based concatenated speech synthesis or speech recognition model training, tolerances can vary from of 10 to 80 milliseconds \cite{129,184,310,72,185}. For audio indexing and content based retrieval tolerances in the order of a few seconds would be deemed acceptable. As indexing and retrieval are usually used for the purpose of navigating (audio based scrolling and searching) through large audio files, fine accuracies are therefore not required. Such systems usually only wish to locate the reasonable proximity of the subject content. Some examples of work in this area include \cite{88,177,16}. Another related area is in the segmentation and annotation of melodies in music signals, where an acceptable tolerance of up to 2 seconds is deemed acceptable \cite{210}.

The segmentation required for the proposed educational technologies concerns synchronization of multimedia content. That is the delivery of visual highlighting of content in synchronization with an audio accompaniment. This can be regarded as a form of continuous time-dependent media for end-users. In general, the tolerance values that this system is concerned with are in the order of tens of milliseconds.

A closer look at the various considerations that need to be taken into account will be discussed now. Firstly, there is a prerequisite to gain a greater understanding of cross-modal perception in humans in relation to multimedia synchronization. This area is non-trivial and is in fact one of the oldest questions in cognitive science concerning the perception of simultaneous
events. In the field of experimental psychology it is sometimes referred to as the "Greenwich Observatory Problem" [170]. The problem first formally came to light in 1796 after astronomer Maskelyne performed measurements and recordings of stellar transits at the Greenwich Observatory. Significant discrepancies were present between his own measurements for relating stars position relative to clock ticks⁶ and that of his assistant's measurements, in some cases even differing up to 800 milliseconds [170,267]. Afterwards, this led to theories that time perception and simultaneity of events vary from one person to the next [170].

It is worthwhile mentioning one theory that is discussed in [170]. This theory involves the idea of anticipatory mechanisms being present in the brain. That is, the brain may have evolved an anticipatory strategy such that the arrival of sensory signals through one sensory channel causes a neural module to predict/expect and prepare in advance for the arrival of the signal from the other associated channel. As a related example to this, consider the problem of presenting recorded audio and video footage of a person uttering the plosive sound /p/. This plosive /p/ sound requires the lips to move into place before the burst leaves the mouth. The presented footage therefore should reflect this temporal relation between audio and video for it to appear normal to a viewer. For instance, the presentation of audio before video in this case would appear unnatural. By the very same token, the case considered in this work, of highlighting words in synchronism with the audio, hearing the word before it is highlighted would also have similar undesirable experiences for the end-user.

It is worth considering the speed of light compared to the speed of sound for simultaneous events, which is proportional to the distance. Mindful of the end-application that is synchronous audio-visual content, for the case of 2D content on a computer screen given the close proximity involved, this is negligible. However it becomes significant for the case of interactive whiteboards.

For an observer at distance $d$ from an audio-visual source, the relative time of arrival of the two signals can be determined as,

$$\tau_\Delta = \tau_a - \tau_v = \frac{d}{c_a} - \frac{d}{c_v}$$  \hspace{1cm} (5.10a)

$$\tau_\Delta = d\left(\frac{1}{c_a} - \frac{1}{c_v}\right)$$  \hspace{1cm} (5.10b)

$$\tau_\Delta = d(2.94 \times 10^{-3})$$  \hspace{1cm} (5.10c)

where $d$ is the distance concerned, $c_a$ and $c_v$ are the speed of sound and the speed of light respectively, $\tau_a$, is the audio time delay and $\tau_v$, is the visual time delay. Here it can be seen that the relative time of arrival between audio and video is proportional to the distance from the audio-visual source to the observer.

⁶This is known as the Bradley's Eye-ear Method [170,267], which was subsequently replaced by the electrochronograph [52]. The interested reader is referred to [267] for a discussion on the method and the general topic known as the Personal Equation.
5.2. Acceptable Tolerances

Typical classrooms are configured to support an interactive white-board of 16' wide by 4' high, with an additional 4' in width for the teachers podium to give 20' wide. Normally the centre of the classroom is 20' from the white-board \([4]\). If the audio source is positioned only at the front of the classroom, this would give an approximate \(\tau_\Delta\) of 17.92 msec. For a student at the back of the classroom this becomes more significant as \(\tau_\Delta\) increases to 35.84 msec. Accordingly this can be compensated for by the suitable location of audio sources (speakers) in a more optimal configuration around the room. However, this makes the experience less immersive when compared to having the sound source located at the interactive white-board only.

It is important to be able to quantify suitable levels of tolerance for the end-application. Most speech segmentation systems in the literature are \emph{machine-oriented}, for example corpus based speech synthesis or speech recognition. This work differs substantially as segmentation is directly utilized as a valuable component in a \emph{human-oriented} end-application. Ultimately the desire is to have a high Quality of Experience (QoE) for the end user with no synchronization errors. In the production of multimedia content this involves having an audio segmentation system that is accurate enough to then be able to synchronize this with the visual information such that these two cross-modal events are perceived to be synchronous in time. This is regarded as an \emph{in sync} stream and the presence of out of sync information in the stream is highly undesirable, yielding content that is confusing, appearing artificial, strange and even annoying for the user \([292]\).

Studies and investigations into the area of cross-modal asynchrony have focused on the two modalities of auditory and visual simultaneity. Two important findings from these studies \([170]\) have emerged:

1. Large individual differences exist in perception thresholds, that is significant variations exist from person to person.

2. The perception of cross-modal simultaneity is asymmetric, meaning that the synchronous threshold for the audio preceding the visual is less than that of the synchronous threshold for the audio following the visual.

The simplest approach involves defining tolerances that are equal in both directions, such that

\[
| -tol_{early}| = | +tol_{late}|.
\]

where \(-tol_{early}\) and \(+tol_{late}\) corresponds to segmentation tolerances for early and late segmentation estimates. It is reasonable that these tolerance values should be made on the basis of the \emph{early} threshold as it is more stringent than the \emph{late} threshold. Another approach is to take into account the asymmetric nature of the perceptual thresholds, as presented in Table 5.1. Using this knowledge, define asymmetric tolerances such that,

\[
| -tol_{early}| \neq | +tol_{late}|.
\]
<table>
<thead>
<tr>
<th>Estimates from Studies (msec)</th>
<th>Symmetry</th>
<th>Sound Early</th>
<th>Sound Late</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dixon &amp; Splitz <em>Speech</em> [75]</td>
<td>Asymmetric</td>
<td>-130</td>
<td>250</td>
</tr>
<tr>
<td>Dixon &amp; Splitz <em>Non-speech</em> [75]</td>
<td>Asymmetric</td>
<td>-75</td>
<td>175</td>
</tr>
<tr>
<td>McGrath &amp; Summerfield [195]</td>
<td>Asymmetric</td>
<td>-65</td>
<td>140</td>
</tr>
<tr>
<td>Jaskowski [130]</td>
<td>Asymmetric</td>
<td>-65</td>
<td>165</td>
</tr>
<tr>
<td>Levitin et al. [170]</td>
<td>Asymmetric</td>
<td>-41</td>
<td>45</td>
</tr>
<tr>
<td>Kristofferson &amp; Allan [9]</td>
<td>Symmetric</td>
<td>-90</td>
<td>90</td>
</tr>
<tr>
<td>Ganz [92]</td>
<td>Symmetric</td>
<td>-100</td>
<td>100</td>
</tr>
<tr>
<td>$\bar{x}$</td>
<td>-80.86</td>
<td>137.86</td>
<td></td>
</tr>
<tr>
<td>S</td>
<td>28.84</td>
<td>67.14</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.1: A summary of results from studies and investigations into auditory-visual simultaneity thresholds, for the cases of Kristofferson, Allan and Ganz. The values are symmetric as they show average results for the two scenarios of *leading* and *lagging* sound signals.

The former approach has been used in this work and is the format that is used for presenting the results at the end of this Chapter. Using the threshold information from Table 5.1, some defined tolerances have been derived, these are termed *fine* = ±70msec, *medium* = ±90msec, and *coarse* ±110msec. These particular tolerances will be used extensively in the results section. However, in the interests of completeness the results are presented for a range of tolerance values.

Aside, one might argue that knowledge of the asymmetric nature of cross-modal perception, could in fact be used to the system’s advantage, by adding a positive offset value to all of the segmentation estimates. This would exploit the asymmetric perceptual phenomenon, to improve the overall perceived synchronization of auditory and visual information/events. Effectively, this would shift the statistical mean for speech segmentation estimates from $\bar{x} \approx 0$ to this assigned positive offset value $\bar{x} \approx \text{offset}$. The offset value could be derived from the difference between the early sound thresholds and late sound thresholds, as presented in Table 5.1. This concept is illustrated in Fig. 5.7. This idea is useful and has therefore been incorporated into the proposed automatic content generation system as an option.

### 5.3 Speech Corpus used for Experimentation

The TIMIT Speech Corpus [83] has been used extensively in this work for experimentation. This is a widely used commercial database that was commissioned by the research and development office for the U.S. Department of Defense, DARPA. The corpus is a phonemically and lexically transcribed speech of American English speakers of different sexes and dialects, which was developed for the purpose of training and testing speech recognition systems. The data corpus was worked on by Texas Instruments (TI) and Massachusetts Institution of Technology (MIT), hence the reason it is called TIMIT. Other variations include TIMIT for clean speech,
NTIMIT for standard telephone speech, CTIMIT for cellular telephone speech in various noise conditions and HTIMIT for different telephone headsets used. Although TIMIT was specifically designed for speech recognition purposes, the speech samples and transcripts information have been adopted for use in other speech processing applications. Using a common database, such as TIMIT, is advantageous as it allows for more direct comparisons with related work. TIMIT was used extensively for experimentation and for the performance assessments of the proposed segmentation method.

The speech samples that were used for experimentation came from the core TIMIT Corpus [83]. This corpus includes male and female speakers from the 8 major dialect regions of the United States.

These dialect regions are:

2. Northern.
4. South Midland.
5. Southern.
7. Western.
8. Army Brat - moved around.
The type of sentences used in the corpus are (i) *Dialect sentences* labelled “SA sentences,” these expose the dialectal variants of the speakers, (ii) *Phonetically-compact sentences* labelled “SX sentences,” these provide a good coverage of pairs of phones, with extra occurrences of phonetic contexts, and (iii) *Phonetically-diverse sentences* labelled “SI sentences,” these sentences were taken from existing text sources. The text sources were, (i) the Brown Corpus\(^7\) which is a general corpus (text collection) in the field of corpus linguistics that was compiled by Kucera and Francis for American English in the 1960’s, and (ii) sentences from the “Playwrights Dialog” Hultzen, et al. [123].

5.4 Results

In this section the results are presented for the proposed segmentation system for the different feature types and approaches that have been investigated in this research. The results are presented and later discussed in terms of:


2. **Baseline Multivariate Features.** For the features, Power Spectrum, Teager Energy Spectrum (denoted \(TEO\ Spec.)\), OBA, \(\frac{1}{2}\)OBA, PLP, and MFCC.

3. **Dynamic Multivariate Features.** For the dynamic feature extensions of delta \(\Delta\), delta-delta \(\Delta\Delta\), also RASTA processing, is considered for the PLP features (subsequently denoted PLPR).


5.4.1 Results Format

The results are presented in different formats to allow performance comparison using different metrics. Firstly, the results for segmentation accuracy for different tolerances \(\pm tol\) msec from \(\pm 50\) through to \(\pm 130\) in increments of 10 msec, are presented in Tables, which are referenced below. These results give the segmentation accuracy in terms of percentage, which is the ratio of the number of segmentation estimates that have fallen within the given tolerance value, \(\pm tol\), and the total number of segments. This is found by referencing the segment estimates, \(\hat{\zeta}_i\) against the *ground-truth* measurements \(\zeta_i\) for the NVR signals to give a segmentation error, \(e_i\), as defined in Eq. (5.11).

\(^7\)The British equivalent to the Brown Corpus is the Lancaster-Oslo-Bergen Corpus (LOB Corpus) compiled later in the 1980's.
As aforementioned, the speech utterances used for experimentation and testing purposes is the TIMIT corpus. The TIMIT corpus has been professionally segmented and transcribed by expert phoneticians and hence they provide a reliable ground truth for evaluation purposes. These segmentation accuracy results are also provided in cumulative histogram plots for various tolerances (±tol in msec) ranging from 10msec to 150msec. This provides a good means of comparing the different approaches. In addition to segmentation accuracy, some sample statistics of mean $\bar{x}$ and standard deviation $S$ of segmentation error $e_i$ are also presented.

$$e_i = \zeta_i - \check{\zeta}_i \quad (5.11)$$

Furthermore, normal probability plots are presented which allow the linearity of segmentation error of estimates to be assessed. If the data comes from a normal distribution the data points would be linear on the plot, whereas for other probability distribution functions (pdf's) there will be a curvature away from the straight red line as shown in Fig. 5.10. Plotting the segmentation error for estimates in this format is valuable as it allows the segmentation errors of the estimates to be viewed individually as points on the plot, i.e. in a non-cumulative format. This format is particularly well suited to viewing and for getting a greater appreciation for any segmentation estimate outliers (statistical abnormalities). The plots also allow the segmentation estimation errors to be viewed temporally in terms of leading or lagging, which is useful. These plots (i.e. Figs. 5.11 5.12, 5.14 and 5.16) are provided to complement the other result plots and tables.

The section on acceptable tolerances in this Chapter outlines some defined tolerances, i.e. fine = ±70msec, medium = ±90msec, and coarse ±110msec. Accordingly, these particular tolerance values are highlighted in the results tables with a circumflex. These tolerances will be referred to in text in the following format, $\hat{S}_{acc}(\text{fine,medium,coarse})\%$.

5.4.2 Univariate Feature Approach

The sample statistics for each of the Univariate Features in terms of the segmentation error ($e_i$) are:

1. Energy ($\bar{x}=0.182$ and Std=0.309) sec.
2. ZCR ($\bar{x}=0.013$ and Std=0.135) sec.
3. S. Centroid ($\bar{x}=-0.029$ and Std=0.149) sec.
4. S. Roll-off ($\bar{x}=0.012$ and Std=0.122) sec.
5. S. Flux ($\bar{x}=-0.997$ and Std=0.974) sec.

The results for the Univariate Features are given in terms of segmentation accuracy for various tolerances (in msec) in Table 5.2. These segmentation accuracies are plotted against
tolerances in Fig. 5.9, using cumulative histograms (note: including additional data for tolerances 10 to 50msec). These results for the five Univariate features that have been considered are for use as single reference feature profile within the segmentation algorithm, which are Energy, ZCR, Spectral Centroid, Spectral Roll-off and Spectral Flux. As expected the plots in Fig. 5.9 show strong trends of increases in accuracy for increases in tolerance size. Spectral Flux has the worst performance which is clearly summarized by the sample statistics of ($\bar{x} = -0.997$ and $S = 0.974$) sec. Similarly, this is observed again from the segmentation accuracy measures at the defined tolerances of (fine, medium, coarse)% giving accuracies of $\hat{S}_{acc}(20.60, 24.25, 25.54)$%. The Energy feature improves upon the Spectral Flux feature with sample statistics of ($\bar{x} = 0.182$ and $S = 0.309$) sec, and segmentation accuracies of $\hat{S}_{acc}(36.05, 40.77, 43.56)$%. Further significant performance improvements are found for the remaining three features, with Spectral Centroid $\hat{S}_{acc}(60.52, 65.24, 71.03)$%, ZCR $\hat{S}_{acc}(62.02, 67.60, 74.03)$% and Spectral Roll-off $\hat{S}_{acc}(68.03, 73.82, 79.83)$%. The Spectral Roll-off feature has performed the best out of all the Univariate features that have been considered, which is indicated with the sample statistics of ($\bar{x} = 0.012$ and $S = 0.122$) sec. Although it must be stated that while the segmentation estimates are approximately centred on zero the standard deviation is still relatively high in comparison with the defined tolerances.

5.4.3 Baseline Multivariate Feature Approach

The sample statistics for each of the Baseline Multivariate Features in terms of the segmentation error ($e_i$) are:

1. Power Spec. ($\bar{x} = -0.001$ and $S = 0.045$) sec.
2. TEO Spec. ($\bar{x} = 0.000$ and $S = 0.042$) sec.
3. OBA ($\bar{x} = 0.000$ and $S = 0.050$) sec.
4. $\frac{1}{2}$OBA ($\bar{x} = 0.000$ and $S = 0.048$) sec.
5. PLP ($\bar{x} = -0.003$ and $S = 0.048$) sec.
6. MFCC ($\bar{x} = -0.003$ and $S = 0.046$) sec.

The results for the Baseline Multivariate Features are given in terms of segmentation accuracy for various tolerances (in msec) in Table 5.3. These segmentation accuracies are plotted against tolerances in Fig. 5.11, using a cumulative histogram. The results show that for the baseline features, the worst performance was achieved by the OBA features, with segmentation accuracies for the predefined tolerances of $\hat{S}_{acc}(83.49, 92.57, 97.70)$% and segmentation error ($\bar{x} = 0.000$ & $S = 0.050$) sec. The features of $\frac{1}{2}$OBA and PLP have a relatively similar performance, which is indicated by having close mean and standard deviations as presented above. The MFCC features outperform these with segmentation accuracies of $\hat{S}_{acc}(87.22, 94.98, 98.33)$%
and segmentation error \( \bar{x} = -0.003 \) & \( S = 0.046 \) sec. The best performance out of all the baseline features was the TEO spectrum, which significantly outperforms all other baseline features including the Power Spec, which had the second best performance overall. The TEO spectrum features achieved segmentation accuracies of \( \hat{S}_{acc}(90.60, 96.87, 99.15)\% \) with a segmentation error of \( (\bar{x} = 0.000 \) and \( S = 0.042 \) sec.

### 5.4.4 Dynamic Multivariate Feature Approach

The sample statistics for a subset of the Dynamic Multivariate Features in terms of the segmentation error \( (e_i) \) are:

1. OBA Set \( (\bar{x} = 0.008 \) and \( S = 0.041 \) sec.
2. \( \frac{1}{3} \) OBA \( \Delta \) \( (\bar{x} = 0.006 \) and \( S = 0.039 \) sec.
3. PLP Set \( (\bar{x} = 0.003 \) and \( S = 0.031 \) sec.
4. PLPR \( \Delta \) \( (\bar{x} = 0.005 \) and \( S = 0.034 \) sec.
5. MFCC Set \( (\bar{x} = 0.007 \) and \( S = 0.035 \) sec.

The results for the Dynamic Multivariate Features are given in terms of segmentation accuracy for various tolerances (in msec) in Table 5.4. These segmentation accuracies are plotted against tolerances in Fig. 5.13, using cumulative histograms. Note: the term Set denotes the grouping of original Baseline Features, with Delta and Delta-delta extension features. First, it is clear, from the results in both Table 5.4 and Fig. 5.13, that the dynamic feature extensions dramatically improve upon the original baseline features, the only exception being the second order derivative (delta-delta) for the MFCC features. The combined feature Sets perform well overall, yielding the best performance for OBA, PLP and MFCC features. The first order delta features perform best for the \( \frac{1}{3} \) OBA and PLPR. The best performing dynamic features for each feature type are then compared in Fig. 5.13f. The two best performing Baseline Multivariate Features are included for comparison purposes. The feature types in order of increasing performance are: OBA Set, \( \frac{1}{3} \) OBA, MFCC, PLPR, PLP Set. The PLP Set having the best performance overall with segmentation accuracies of \( \hat{S}_{acc}(97.30, 99.57, 99.95)\% \) and segmentation error statistics of \( (\bar{x} = 0.003 \) & \( S = 0.031 \) sec.

### 5.4.5 PCA Dynamic Multivariate Feature Approach

The sample statistics for each of the PCA Dynamic Multivariate Features in terms of the segmentation error \( (e_i) \) are given in Table 5.5. The performance of the segmentation algorithm with the inclusion of feature transformation and dimensionality reduction achieved through PCA is presented here. This additional stage has been applied to the Dynamic Multivariate Features Sets for the features of OBA Set, \( \frac{1}{3} \) OBA Set, PLP Set, PLPR Set and MFCC Set, and for the
Baseline Multivariate Features of the Power Spectrum and the TEO Spectrum. The features have been reduced using the criterion presented in Chapter 4, in Eq. (4.60). This criterion is based on the magnitude of the eigenvalues, which is used to determine the eigenvectors that form part of the principal component set (or transformation matrix). The reduction summaries for the features are presented in Table 5.5. These are given in terms of the sample statistics ($\bar{x}, S$) sec. The dimensionality reduction in terms of ratios (new-dim : original-dim), are $\approx 0.6$. This is a considerable amount of reduction, effectively reducing the feature vector dimension used in the algorithm by $\approx 40\%$.

The results for the PCA Dynamic Multivariate Features are given, in terms of segmentation accuracy for various tolerances (in msec), in Table 5.6. These segmentation accuracies are plotted against tolerances in Fig. 5.15, using cumulative histograms. The best PCA approach, of the three that are considered in this study, was the Signal Concatenation Approach (denoted PCA-Con). Although the Average Vector Approach (denoted PCA-Ave) had a very similar performance, almost matching that of the PCA-Con method on several occasions. The exception being for the case of Power Spec and TEO Spec; these two feature types in general did not lend themselves well to the PCA technique. The worst performance for all three of the PCA approaches explored, was Independent Vector Approach (denoted PCA-Ind). All of the features types are compared in Fig. 5.15f, using the best performing PCA method, which in all case was in fact the PCA-Con method. Using the PCA methodology as part of the algorithm the worst performances produced were by the Power Spec and the TEO Spec. The best performance for this approach was achieved by the PLP Set Con achieving segmentation accuracies of $\hat{S}_{acc}(97.25, 99.54, 99.95)\%$ and segmentation error of $(\bar{x} = 0.004 & S = 0.032)$ sec.

5.5 Discussion

The results have shown that the Univariate Feature Approach performed poorly in the proposed segmentation system. The poor performance of this approach is attributed to the difficulties that exist in trying to extract a univariate feature that can sufficiently represent the acoustic variations (with respect to linguistic message) present in the speech signal across time. This topic was discussed in Chapter 4, Section 4.1, under the topic the philosophy of features. Effectively in practice, speech signals require multiple features in order to describe the variations present in the signal. These multiple features form a vector representation of the signal. The univariate features were found to be problematic within the DTW procedure. The time alignment of the SVR (reference) and NVR (query) signals proves to be unsatisfactory due to the deficient qualities of the univariate feature representations. This gives rise to singularities occurring in the time warping procedure, effectively producing undesirable warping paths through the cumulative distance matrix. This effectively refers to large vertical or large horizontal warps occurring at particular points. Examples of this are presented in Fig. 5.8. Out of all the univariate features that were considered, the best performance was produced by the Spectral Roll-off feature with
5.5. Discussion

Figure 5.8: The warping paths $W$ are shown in red through the cumulative distance matrices $\Gamma_{I,J}$, for the TIMIT utterances: (a & d) "His work began just six days after the flood," (b & e) "Only the most accomplished artists obtain popularity," and (c & f) "Grandmother outgrew her upbringing in petticoats." The top panel is for the univariate Energy feature and the bottom panel is for the multivariate PLP Set features. Noting that the PLP Set is far superior in terms of their performance in experimentation results. This is due to the PLP Set (multivariate) features being much better at characterizing the speech signals SVR and NVR with respect to each other. It is observed from warping paths that the Energy feature due to its limitations in characterizing the speech signal produces inappropriate warps.

segmentation accuracies of $\hat{S}_{acc}(68.03, 73.82, 79.83)\%$ and segmentation error of $(\bar{x}=0.012$ and Std=$0.122)$ sec. These results are unsatisfactory, as they strongly indicate that the univariate feature approach is not accurate enough to be used within the final system.

The results for the Baseline Multivariate Feature Approach show a considerable improvement on the Univariate Feature Approach. This is easily shown by comparing the results in terms of segmentation accuracies and segmentation error for the two approaches. It is interesting to note that a similar performance has been achieved by the proposed $\frac{1}{2}$OBA features in comparison with the state of the art MFCC and PLP features. These state of the art features take into account the perceptual auditory effects, whereas the OBA features are much less sophisticated. They are derived from the STFT through smoothing the spectrum into octave bands and computing the energy for each band. A better performance was achieved by using straightforward spectral features such as the Power Spectrum and Teager Energy Spectrum (denoted TEO Spec). The
best performance was given by the Teager Energy Spectrum, with segmentation accuracies of \( \hat{S}_{\text{acc}}(90.60, 96.87, 99.15)\% \) and segmentation error of \( (\bar{x} = 0.000 \text{ and } S = 0.042) \) sec. The good performance of both the Power Spectrum and Teager Energy Spectrum is attributed to their fine spectral detail, this has been found to have merit with regard to characterizing the speech signal well for the task of speech segmentation. The Teager Energy Spectrum differs from the power Spectrum by scaling the spectrum non-linearly in terms of frequency. This has the effect of pre-emphasizing the upper portion of the frequency spectrum. From a feature extraction point of view, this is significant as it effectively emphasizes the harmonic structure and pitch information present in the signal, and removes much of the influence of the vocal tract’s frequency response. This is something which is useful as the vocal tract frequency responses can differ quite considerably for the SVR and NVR signals. Hence the removal of the vocal tract’s influence is thought to be favourable for the task of speech segmentation.

While the PLP and MFCC features have performed poorly with respect to the Power Spectrum and the Teager Energy Spectrum in the Baseline Multivariate Feature Approach, these features redeem themselves considerably in the Dynamic Multivariate Feature Approach. Their baseline feature performance is increased considerably with the inclusion of dynamic feature extensions (delta, and delta-delta). The OBA and \( \frac{1}{3}\)OBA features also gained significant increases in performance using these dynamic extensions, although the PLP and MFCC dynamic feature Sets performed the best out of the Dynamic Multivariate Feature Approach. The PLP Set produced the best best performance overall for all of the approaches investigated with segmentation accuracies of \( \hat{S}_{\text{acc}}(97.30, 99.57, 99.95)\% \) and segmentation error statistics of \( (\bar{x} = 0.003 \text{ and } S = 0.031) \) sec. This excellent performance proves the effectiveness of the proposed approach for automatically segmenting speech at the word boundaries using an unsupervised text-dependent methodology. These results show that the proposed method is suitable to perform its important role within the automated production process for the production of educational applications.

Three PCA techniques have been investigated as part of the PCA Dynamic Multivariate Feature Approach, these are:

1. **Independent Vector Approach** (denoted PCA-Ind).

2. **Signal Concatenation Approach** (denoted PCA-Con).

3. **Average Vector Approach** (denoted PCA-Ave).

Of these three techniques, the best performance was consistently achieved by the proposed Signal Concatenation Approach. Although the Average Vector Approach in most cases approaches a similar performance, the Signal Concatenation Approach is however a much simpler and more straightforward technique. The nature of the Signal Concatenation Approach also has an additional merit in that its eigen-decomposition preserves orthonormality between the principal component vectors, which from a strict PCA standpoint is desirable, i.e. still forming an orthonormal basis.
Within the *PCA Dynamic Multivariate Feature Approach* no significant gains have been achieved in terms of increasing the segmentation accuracies or by the same token reducing the segmentation error. However, other gains were achieved in terms of reducing the computational complexity of the algorithm while maintaining the same level of performance. As previously stated the best performance was achieved overall for the PLP Set in the *Dynamic Multivariate Feature Approach*. However within the *PCA Dynamic Multivariate Feature Approach*, using the PLP Set Con technique, has achieved approximately the same performance using a transformed and greatly reduced dimensional PLP feature set ($f \approx 17$ on average), reduced from the original PLP Set with ($j = 27$), which is represented by $\mathbb{R}^{j=27} \rightarrow \mathbb{R}^{f=17}$. This reduction has merit, as it reduces the computational complexity of computing the distance matrix in the DTW procedure.

### 5.6 Conclusion

This Chapter has presented the proposed speech segmentation methodology, which uses the Dynamic Programming (DP) technique known as Dynamic Time Warping (DTW). The segmentation algorithm presented has employed the various features and signal analysis methods that were examined in Chapter 4. The Principal Component Analysis approaches to feature transformation and dimensionality reduction that were investigated have been described. Other areas have been covered, such as defining acceptable tolerances in relation to the end-application (educational technologies), as well as the speech corpus (TIMIT) that was used. The results have been presented and discussed for the various approaches that have been explored. The results have shown that a superior performance has been achieved using the PLP Set within the *Dynamic Multivariate Feature Approach*, with segmentation accuracies of $\hat{S}_{acc}(97.30, 99.57, 99.95)\%$ and segmentation error statistics of $(\bar{x} = 0.003 \& S = 0.031)$ sec. A similar performance has been achieved using a transformed and reduced feature set, i.e. the PLP Set Con within the *PCA Dynamic Multivariate Features*, with segmentation accuracies of $\hat{S}_{acc}(97.25, 99.54, 99.95)\%$ and segmentation error of $(\bar{x} = 0.004 \& S = 0.032)$ sec. These results strongly show the effectiveness and the validity of the proposed automatic speech segmentation methodology. Thus clearly showing that the automatic speech segmentation method is suitable for fulfilling its important role within the automated production process for producing educational applications.
Figure 5.9: Segmentation Accuracy against Tolerances in msec, for Univariate Features.

Table 5.2: Segmentation Accuracy for different Tolerances in msec, for Univariate Features. $\hat{S}_{acc}$ values are indicated with a circumflex.

<table>
<thead>
<tr>
<th>Tol. (m-sec):</th>
<th>50</th>
<th>60</th>
<th>70</th>
<th>80</th>
<th>90</th>
<th>100</th>
<th>110</th>
<th>120</th>
<th>130</th>
</tr>
</thead>
<tbody>
<tr>
<td>Energy</td>
<td>29.83</td>
<td>33.05</td>
<td>36.05</td>
<td>38.84</td>
<td>40.77</td>
<td>42.70</td>
<td>43.56</td>
<td>45.06</td>
<td>48.28</td>
</tr>
<tr>
<td>ZCR</td>
<td>57.30</td>
<td>60.30</td>
<td>62.02</td>
<td>64.81</td>
<td>67.60</td>
<td>70.60</td>
<td>74.03</td>
<td>75.11</td>
<td>76.18</td>
</tr>
<tr>
<td>S. Centroid</td>
<td>52.36</td>
<td>57.51</td>
<td>60.52</td>
<td>63.95</td>
<td>65.24</td>
<td>67.81</td>
<td>71.03</td>
<td>73.18</td>
<td>74.25</td>
</tr>
<tr>
<td>S. Roll-off</td>
<td>59.87</td>
<td>64.81</td>
<td>68.03</td>
<td>71.24</td>
<td>73.82</td>
<td>75.97</td>
<td>79.83</td>
<td>82.40</td>
<td>84.76</td>
</tr>
<tr>
<td>S. Flux</td>
<td>16.95</td>
<td>18.45</td>
<td>20.60</td>
<td>22.75</td>
<td>24.25</td>
<td>24.46</td>
<td>25.54</td>
<td>26.82</td>
<td>27.47</td>
</tr>
</tbody>
</table>
Figure 5.10: Normal distribution plots of segmentation errors, for Univariate Features.
Figure 5.11: Segmentation Accuracy against Tolerances in msec, for Baseline Multivariate Features.

Table 5.3: Segmentation Accuracy for different Tolerances in msec, for Baseline Multivariate Features. $S_{acc}$ values are indicated with a circumflex.

<table>
<thead>
<tr>
<th>Tol. (m-sec):</th>
<th>50</th>
<th>60</th>
<th>70</th>
<th>80</th>
<th>90</th>
<th>100</th>
<th>110</th>
<th>120</th>
<th>130</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power Spec.</td>
<td>73.61</td>
<td>82.00</td>
<td>88.22</td>
<td>92.61</td>
<td>95.57</td>
<td>97.45</td>
<td>98.60</td>
<td>99.27</td>
<td>99.63</td>
</tr>
<tr>
<td>TEO Spec.</td>
<td><strong>76.84</strong></td>
<td><strong>84.88</strong></td>
<td><strong>90.60</strong></td>
<td><strong>94.44</strong></td>
<td><strong>96.87</strong></td>
<td><strong>98.33</strong></td>
<td><strong>99.15</strong></td>
<td><strong>99.59</strong></td>
<td><strong>99.81</strong></td>
</tr>
<tr>
<td>OBA</td>
<td>67.85</td>
<td>76.58</td>
<td>83.49</td>
<td>88.73</td>
<td>92.57</td>
<td>95.26</td>
<td>97.08</td>
<td>98.27</td>
<td>99.01</td>
</tr>
<tr>
<td>1/3 OBA</td>
<td>69.87</td>
<td>78.51</td>
<td>85.21</td>
<td>90.18</td>
<td>93.72</td>
<td>96.13</td>
<td>97.70</td>
<td>98.69</td>
<td>99.28</td>
</tr>
<tr>
<td>PLP</td>
<td>70.03</td>
<td>78.66</td>
<td>85.34</td>
<td>90.29</td>
<td>93.80</td>
<td>96.19</td>
<td>97.75</td>
<td>98.72</td>
<td>99.30</td>
</tr>
<tr>
<td>MFCC</td>
<td>72.33</td>
<td>80.82</td>
<td>87.22</td>
<td>91.82</td>
<td>94.98</td>
<td>97.04</td>
<td>98.33</td>
<td>99.10</td>
<td>99.53</td>
</tr>
</tbody>
</table>
5.6. Conclusion

Figure 5.12: Normal distribution plots of segmentation errors, for Baseline Multivariate Features.
Figure 5.13: Segmentation Accuracy against Tolerances in msec, for Dynamic Multivariate Features.
Table 5.4: Segmentation Accuracy for different Tolerances in msec, for Dynamic Multivariate Features. $\hat{S}_{acc}$ values are indicated with a circumflex.

<table>
<thead>
<tr>
<th>Tol. (m-sec)</th>
<th>50</th>
<th>60</th>
<th>70</th>
<th>80</th>
<th>90</th>
<th>100</th>
<th>110</th>
<th>120</th>
<th>130</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power Spec.</td>
<td>73.61</td>
<td>82.00</td>
<td>88.22</td>
<td>92.61</td>
<td>95.57</td>
<td>97.45</td>
<td>98.60</td>
<td>99.27</td>
<td>99.63</td>
</tr>
<tr>
<td>TEO Spec.</td>
<td>76.84</td>
<td>84.88</td>
<td>90.60</td>
<td>94.44</td>
<td>96.87</td>
<td>98.33</td>
<td>99.15</td>
<td>99.59</td>
<td>99.81</td>
</tr>
<tr>
<td>OBA</td>
<td>67.85</td>
<td>76.58</td>
<td>83.49</td>
<td>88.73</td>
<td>92.57</td>
<td>95.26</td>
<td>97.08</td>
<td>98.27</td>
<td>99.01</td>
</tr>
<tr>
<td>OBA $\Delta$</td>
<td>75.74</td>
<td>83.92</td>
<td>89.82</td>
<td>93.86</td>
<td>96.46</td>
<td>98.06</td>
<td>98.99</td>
<td>99.50</td>
<td>99.76</td>
</tr>
<tr>
<td>OBA $\Delta\Delta$</td>
<td>72.33</td>
<td>80.82</td>
<td>87.23</td>
<td>91.83</td>
<td>94.99</td>
<td>97.05</td>
<td>98.34</td>
<td>99.10</td>
<td>99.54</td>
</tr>
<tr>
<td>OBA Set</td>
<td>76.62</td>
<td>84.69</td>
<td>90.45</td>
<td>94.33</td>
<td>96.80</td>
<td>98.28</td>
<td>99.12</td>
<td>99.58</td>
<td>99.81</td>
</tr>
<tr>
<td>$\frac{1}{3}$OBA</td>
<td>69.87</td>
<td>78.51</td>
<td>85.21</td>
<td>90.18</td>
<td>93.72</td>
<td>96.13</td>
<td>97.70</td>
<td>98.69</td>
<td>99.28</td>
</tr>
<tr>
<td>$\frac{1}{3}$OBA $\Delta$</td>
<td>80.01</td>
<td>87.60</td>
<td>92.73</td>
<td>95.98</td>
<td>97.90</td>
<td>98.97</td>
<td>99.52</td>
<td>99.79</td>
<td>99.91</td>
</tr>
<tr>
<td>$\frac{1}{3}$OBA $\Delta\Delta$</td>
<td>74.01</td>
<td>82.37</td>
<td>88.54</td>
<td>92.86</td>
<td>95.75</td>
<td>97.58</td>
<td>98.69</td>
<td>99.32</td>
<td>99.66</td>
</tr>
<tr>
<td>$\frac{1}{3}$OBA Set</td>
<td>77.65</td>
<td>85.59</td>
<td>91.17</td>
<td>94.86</td>
<td>97.16</td>
<td>98.51</td>
<td>99.26</td>
<td>99.65</td>
<td>99.85</td>
</tr>
<tr>
<td>PLP</td>
<td>70.03</td>
<td>78.66</td>
<td>85.34</td>
<td>90.29</td>
<td>93.80</td>
<td>96.19</td>
<td>97.75</td>
<td>98.72</td>
<td>99.30</td>
</tr>
<tr>
<td>PLP $\Delta$</td>
<td>87.32</td>
<td>93.31</td>
<td>96.74</td>
<td>98.54</td>
<td>99.40</td>
<td>99.77</td>
<td>99.92</td>
<td>99.98</td>
<td>99.99</td>
</tr>
<tr>
<td>PLP $\Delta\Delta$</td>
<td>73.62</td>
<td>82.01</td>
<td>88.24</td>
<td>92.64</td>
<td>95.59</td>
<td>97.47</td>
<td>98.62</td>
<td>99.28</td>
<td>99.64</td>
</tr>
<tr>
<td>PLP Set</td>
<td><strong>88.77</strong></td>
<td><strong>94.33</strong></td>
<td><strong>97.38</strong></td>
<td><strong>98.89</strong></td>
<td><strong>99.57</strong></td>
<td><strong>99.85</strong></td>
<td><strong>99.95</strong></td>
<td><strong>99.99</strong></td>
<td><strong>100.00</strong></td>
</tr>
<tr>
<td>PLPR</td>
<td>41.71</td>
<td>49.01</td>
<td>55.80</td>
<td>62.04</td>
<td>67.71</td>
<td>72.80</td>
<td>77.30</td>
<td>81.25</td>
<td>84.67</td>
</tr>
<tr>
<td>PLPR $\Delta$</td>
<td>85.38</td>
<td>91.88</td>
<td>95.81</td>
<td>97.99</td>
<td>99.11</td>
<td>99.63</td>
<td>99.86</td>
<td>99.95</td>
<td>99.98</td>
</tr>
<tr>
<td>PLPR $\Delta\Delta$</td>
<td>74.71</td>
<td>83.01</td>
<td>89.07</td>
<td>93.28</td>
<td>96.06</td>
<td>97.79</td>
<td>98.82</td>
<td>99.40</td>
<td>99.71</td>
</tr>
<tr>
<td>PLPR Set</td>
<td>69.47</td>
<td>78.14</td>
<td>84.88</td>
<td>89.90</td>
<td>93.50</td>
<td>95.97</td>
<td>97.59</td>
<td>98.61</td>
<td>99.23</td>
</tr>
<tr>
<td>MFCC</td>
<td>72.33</td>
<td>80.82</td>
<td>87.22</td>
<td>91.82</td>
<td>94.98</td>
<td>97.04</td>
<td>98.33</td>
<td>99.10</td>
<td>99.53</td>
</tr>
<tr>
<td>MFCC $\Delta$</td>
<td>78.50</td>
<td>86.33</td>
<td>91.75</td>
<td>95.28</td>
<td>97.44</td>
<td>98.69</td>
<td>99.37</td>
<td>99.71</td>
<td>99.87</td>
</tr>
<tr>
<td>MFCC $\Delta\Delta$</td>
<td>59.92</td>
<td>68.67</td>
<td>76.06</td>
<td>82.13</td>
<td>86.98</td>
<td>90.73</td>
<td>93.57</td>
<td>95.65</td>
<td>97.13</td>
</tr>
<tr>
<td>MFCC Set</td>
<td>84.15</td>
<td>90.94</td>
<td>95.17</td>
<td>97.60</td>
<td>98.89</td>
<td>99.52</td>
<td>99.81</td>
<td>99.93</td>
<td>99.98</td>
</tr>
</tbody>
</table>
Figure 5.14: Normal distribution plots of segmentation errors, for Dynamic Multivariate Features.
Table 5.5: PCA Dynamic Multivariate Features. Dimensionality reduction of feature representations using different PCA approaches. The figures presented below are in terms of sample mean and standard deviation of the resulting remaining dimensions after PCA. Also sample mean and standard deviation is given for the segmentation error, $e_t$, for results comparison.

<table>
<thead>
<tr>
<th>Dim. Reduced to:</th>
<th>$n$-dim.</th>
<th>Estimation Error (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power Spec Con</td>
<td>256</td>
<td>173.063</td>
</tr>
<tr>
<td>Power Spec Ave</td>
<td>256</td>
<td>157.469</td>
</tr>
<tr>
<td>Power Spec Ind</td>
<td>256</td>
<td>157.469</td>
</tr>
<tr>
<td>TEO Spec Con</td>
<td>256</td>
<td>184.813</td>
</tr>
<tr>
<td>TEO Spec Ave</td>
<td>256</td>
<td>172.781</td>
</tr>
<tr>
<td>TEO Spec Ind</td>
<td>256</td>
<td>172.781</td>
</tr>
<tr>
<td>OBA Set Con</td>
<td>24</td>
<td>16.313</td>
</tr>
<tr>
<td>OBA Set Ave</td>
<td>24</td>
<td>15.938</td>
</tr>
<tr>
<td>OBA Set Ind</td>
<td>24</td>
<td>15.938</td>
</tr>
<tr>
<td>$\frac{1}{2}$OBA Set Con</td>
<td>69</td>
<td>42.688</td>
</tr>
<tr>
<td>$\frac{1}{2}$OBA Set Ave</td>
<td>69</td>
<td>41.438</td>
</tr>
<tr>
<td>$\frac{1}{2}$OBA Set Ind</td>
<td>69</td>
<td>41.438</td>
</tr>
<tr>
<td>PLP Set Con</td>
<td>27</td>
<td>16.938</td>
</tr>
<tr>
<td>PLP Set Ave</td>
<td>27</td>
<td>16.781</td>
</tr>
<tr>
<td>PLP Set Ind</td>
<td>27</td>
<td>16.781</td>
</tr>
<tr>
<td>PLPR Set Con</td>
<td>27</td>
<td>17.250</td>
</tr>
<tr>
<td>PLPR Set Ave</td>
<td>27</td>
<td>17.094</td>
</tr>
<tr>
<td>PLPR Set Ind</td>
<td>27</td>
<td>17.094</td>
</tr>
<tr>
<td>MFCC Set Con</td>
<td>60</td>
<td>39.156</td>
</tr>
<tr>
<td>MFCC Set Ave</td>
<td>60</td>
<td>44.781</td>
</tr>
<tr>
<td>MFCC Set Ind</td>
<td>60</td>
<td>44.781</td>
</tr>
</tbody>
</table>
Figure 5.15: Segmentation Accuracy against Tolerances in msec, for PCA Dynamic Multivariate Features.
Table 5.6: Segmentation Accuracy for different Tolerances in msec, for PCA Dynamic Multivariate Features. \( \hat{S}_{\text{acc}} \) values are indicated with a circumflex.

<table>
<thead>
<tr>
<th>Tol. (m-sec)</th>
<th>50</th>
<th>60</th>
<th>70</th>
<th>80</th>
<th>90</th>
<th>100</th>
<th>110</th>
<th>120</th>
<th>130</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power Spec Con</td>
<td>42.34</td>
<td>49.72</td>
<td>56.56</td>
<td>62.84</td>
<td>68.52</td>
<td>73.59</td>
<td>78.07</td>
<td>81.98</td>
<td>85.35</td>
</tr>
<tr>
<td>Power Spec Ave</td>
<td>28.68</td>
<td>34.09</td>
<td>39.32</td>
<td>44.36</td>
<td>49.18</td>
<td>53.78</td>
<td>58.14</td>
<td>62.24</td>
<td>66.09</td>
</tr>
<tr>
<td>Power Spec Ind</td>
<td>18.81</td>
<td>22.47</td>
<td>26.09</td>
<td>29.65</td>
<td>33.15</td>
<td>36.58</td>
<td>39.93</td>
<td>43.20</td>
<td>46.38</td>
</tr>
<tr>
<td>TEO Spec Con</td>
<td>51.50</td>
<td>59.79</td>
<td>67.17</td>
<td>73.61</td>
<td>79.12</td>
<td>83.75</td>
<td>87.55</td>
<td>90.62</td>
<td>93.06</td>
</tr>
<tr>
<td>TEO Spec Ave</td>
<td>27.90</td>
<td>33.17</td>
<td>38.29</td>
<td>43.22</td>
<td>47.96</td>
<td>52.49</td>
<td>56.79</td>
<td>60.86</td>
<td>64.68</td>
</tr>
<tr>
<td>TEO Spec Ind</td>
<td>22.54</td>
<td>26.89</td>
<td>31.15</td>
<td>35.32</td>
<td>39.38</td>
<td>43.32</td>
<td>47.13</td>
<td>50.81</td>
<td>54.35</td>
</tr>
<tr>
<td>OBA Set Con</td>
<td>79.10</td>
<td>86.84</td>
<td>92.14</td>
<td>95.56</td>
<td>97.63</td>
<td>98.80</td>
<td>99.43</td>
<td>99.74</td>
<td>99.89</td>
</tr>
<tr>
<td>OBA Set Ave</td>
<td>77.85</td>
<td>85.76</td>
<td>91.30</td>
<td>94.95</td>
<td>97.22</td>
<td>98.55</td>
<td>99.28</td>
<td>99.67</td>
<td>99.85</td>
</tr>
<tr>
<td>OBA Set Ind</td>
<td>30.26</td>
<td>35.92</td>
<td>41.38</td>
<td>46.61</td>
<td>51.60</td>
<td>56.32</td>
<td>60.76</td>
<td>64.92</td>
<td>68.79</td>
</tr>
<tr>
<td>( \frac{1}{3} )OBA Set Con</td>
<td>78.03</td>
<td>85.92</td>
<td>91.43</td>
<td>95.05</td>
<td>97.29</td>
<td>98.60</td>
<td>99.31</td>
<td>99.68</td>
<td>99.86</td>
</tr>
<tr>
<td>( \frac{1}{3} )OBA Set Ave</td>
<td>74.18</td>
<td>82.52</td>
<td>88.66</td>
<td>92.96</td>
<td>95.82</td>
<td>97.63</td>
<td>98.72</td>
<td>99.34</td>
<td>99.67</td>
</tr>
<tr>
<td>( \frac{1}{3} )OBA Set Ind</td>
<td>32.68</td>
<td>38.72</td>
<td>44.51</td>
<td>50.01</td>
<td>55.22</td>
<td>60.10</td>
<td>64.65</td>
<td>68.85</td>
<td>72.71</td>
</tr>
<tr>
<td>PLP Set Con</td>
<td>88.46</td>
<td>94.11</td>
<td>97.25</td>
<td>98.82</td>
<td>99.54</td>
<td>99.84</td>
<td>99.95</td>
<td>99.98</td>
<td>100.00</td>
</tr>
<tr>
<td>PLP Set Ave</td>
<td>87.82</td>
<td>93.66</td>
<td>96.97</td>
<td>98.67</td>
<td>99.46</td>
<td>99.80</td>
<td>99.93</td>
<td>99.98</td>
<td>99.99</td>
</tr>
<tr>
<td>PLP Set Ind</td>
<td>22.05</td>
<td>26.30</td>
<td>30.49</td>
<td>34.57</td>
<td>38.56</td>
<td>42.44</td>
<td>46.20</td>
<td>49.83</td>
<td>53.33</td>
</tr>
<tr>
<td>PLPR Set Con</td>
<td>69.29</td>
<td>77.96</td>
<td>84.72</td>
<td>89.78</td>
<td>93.40</td>
<td>95.89</td>
<td>97.54</td>
<td>98.58</td>
<td>99.21</td>
</tr>
<tr>
<td>PLPR Set Ave</td>
<td>71.26</td>
<td>79.83</td>
<td>86.37</td>
<td>91.13</td>
<td>94.45</td>
<td>96.67</td>
<td>98.08</td>
<td>98.93</td>
<td>99.43</td>
</tr>
<tr>
<td>PLPR Set Ind</td>
<td>23.61</td>
<td>28.15</td>
<td>32.59</td>
<td>36.92</td>
<td>41.12</td>
<td>45.20</td>
<td>49.13</td>
<td>52.90</td>
<td>56.52</td>
</tr>
<tr>
<td>MFCC Set Con</td>
<td>84.07</td>
<td>90.88</td>
<td>95.13</td>
<td>97.57</td>
<td>98.87</td>
<td>99.51</td>
<td>99.81</td>
<td>99.93</td>
<td>99.98</td>
</tr>
<tr>
<td>MFCC Set Ave</td>
<td>83.81</td>
<td>90.67</td>
<td>94.98</td>
<td>97.48</td>
<td>98.82</td>
<td>99.49</td>
<td>99.79</td>
<td>99.92</td>
<td>99.97</td>
</tr>
<tr>
<td>MFCC Set Ind</td>
<td>34.65</td>
<td>40.99</td>
<td>47.03</td>
<td>52.74</td>
<td>58.09</td>
<td>63.07</td>
<td>67.67</td>
<td>71.87</td>
<td>75.69</td>
</tr>
</tbody>
</table>
Figure 5.16: Normal distribution plots of segmentation errors, for PCA Dynamic Multivariate Features.
This Chapter begins by detailing some important background theory which has supported the research and development of the proposed technology within the Automated Content Generation System (ACG) along with its educational applications. The areas reported upon are: eye movement characteristics, appropriate reading rates, some cognitive aspects of synchronization in multimedia resources, followed with a discussion on hypotheses of learning styles. Following this a description is provided on how the proposed methods integrate into the ACG system and the end-application areas. Later in the Chapter, the three main application areas explored are presented: (a) ReciTell (b) iMARK and (c) The Book: Discovering Sounds Initiative (TBDSI). These application areas form important contributions to this piece of research. They are discussed under four thematic areas: (i) Background information, (ii) Problems addressed, (iii) Proposed solution, and (iv) Outcomes.

These three applications have two key aspects in common, (i) they focus on education, and (ii) they utilize sound as a significant component for improving the interaction, engagement and enrichment of the educational environment to support learning in unique and innovative ways. It appears that little emphasis has been placed on using sound as a major component for digital educational resources. This seems remarkable given the great potential for speech and other sounds for being a valuable medium for creating more engaging and captivating learning environments. Winn [327] stresses the importance of “Human speech” as being “the most powerful and expressive medium the designer has available for use in instructional messages.” There are many potential reasons why sound has often been overlooked in the past; some of these reasons can be legacy, due to technology capabilities, cost, and integration factors. Bishop and Cates [34, 35]
offer some discussion of why this might be the case. One argument is that timing and technology limitations played a part in this, such that much of early influential literature [105,144,8], on the topic was written before the integration of sound was fully realizable.

6.0.0.1 Eye Tracking Fixations and Saccades

Eye movements are an important aspect for the development of synchronized media content, in particular within the context of this research which is aimed towards educational technologies where cross-modal fusion of auditory and visual information plays an integral role for enriching and enhancing the learning environment. Hence consideration is given in this section to various aspects in relation to eye movements when reading. This topic has been covered well in the literature and many studies and experiments have been conducted. A good review on the area has been presented by Rayner in [251]. Some characteristics of eye movements are defined by the terminology *saccades* and *fixations*. Saccades are the movements made during activities such as: reading, looking at a scene or searching for an object, whereas fixations are the periods in between saccades where the eyes remain relatively still somewhere in the region of 200-300msec [312,251]. Eye movements are illustrated in Fig. 6.1, for two different tasks: (a) reading, and (b) memorization. These two tasks are quite different and show the large degree of variation in characteristics. The task on the left hand side involves reading a passage of text for comprehension and the task on the right hand side involves a visual memorization of an urban skyline where the candidate observer was asked to remember as much information as possible about the scene. There are noticeable differences between the eye movement patterns. The reading task is the more organized of the two eye movement patterns, scanning from left to right and then returning to the start of the next line, while eye characteristics for the memorization task have more of a scattered pattern, focusing more on the buildings and less on areas of the sky and water. The velocity of saccades are quite rapid, which can be as high as 500° per second. It is well understood that little information is obtained during saccade movements [312,251]. The reason for this is that during the movement blurring occurs which is referred to as *saccade suppression*. This blurring effect, however, is not perceived during reading, because of the masking effect of the information received before and after the saccade [251].

Table 6.1 gives the eye movement characteristics over different levels of reading proficiency for 1st to 6th Grade Pupils and for Adults. These results are based on the average results for studies conducted on eye movement characteristics by McConkie et al. [193], Rayner [250], Taylor [303], Buswell [48], and have previously been summarized in [251]. This gives important information about the duration of fixations, the number of fixations per 100 words and the frequency of regressions.
When a person is reading a sentence silently, the eye movements show that not every word is fixated. Every once in a while a regression (an eye movement that goes back in the text) is made to re-examine a word that may have not been fully understood the first time. This only happens with about 10% of the fixations, depending on how difficult the text is. The more difficult the higher the likelihood that regressions are made.

Figure 6.1: This figure demonstrates eye movement characteristics acquired using sophisticated eye tracking equipment. The eye fixations are indicated with orange and green dots and the rapid eye movements in between i.e. the saccades are indicated by connecting lines. Credit: images by Rayner and Castelhano [252].

Table 6.1: Advancement in Eye Movement Characteristics

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Grade 1</th>
<th>Grade 2</th>
<th>Grade 3</th>
<th>Grade 4</th>
<th>Grade 5</th>
<th>Grade 6</th>
<th>Adult</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixation duration</td>
<td>355</td>
<td>306</td>
<td>286</td>
<td>266</td>
<td>255</td>
<td>249</td>
<td>233</td>
</tr>
<tr>
<td>Fixations per 100 words</td>
<td>191</td>
<td>151</td>
<td>131</td>
<td>121</td>
<td>117</td>
<td>106</td>
<td>94</td>
</tr>
<tr>
<td>Frequency of regressions (%)</td>
<td>28</td>
<td>26</td>
<td>25</td>
<td>26</td>
<td>26</td>
<td>22</td>
<td>14</td>
</tr>
</tbody>
</table>

6.0.0.2 Reading Rates

As discussed previously, a related aspect to eye movements is defining appropriate reading rates. There is a certain degree of variability in reading rates amongst developing young readers and even among skilled readers. A study [251] conducted on 10 skilled readers found sample mean and standard deviations of WPM to be \( \bar{x} = 309 \) and \( S = 58 \). For younger readers the topic has been discussed in [28, 253, 199], in relation to Reading While Listening (R WL), which is quite different to just oral or listening as it is more of a co-ordinated task.

Three different rates were investigated by McMahon in [199] which were as follows: (i) oral reading rate was found to be 18-50 Words Per Minute (WPM) for first grade and 50-91 WPM for third grade, (ii) read-along reading rate, based on a study that surveyed talking book/kits with audio tapes accompaniment produced by a host of reputable publishers, e.g. Walt Disney, Random House, etc. For first grade the mean and standard deviation for WPM was found to be \( \bar{x} = 112 \) and \( S = 17 \) and for third grade was found to be \( \bar{x} = 141 \) and \( S = 21 \) and (iii) an
intermediate reading rate, which was set to 35% of the oral reading rate, which was found to be an appropriate rate for RWL purposes. The study was quite critical of the rates adopted in commercial read-along book/kits i.e. read-along reading rate, claiming the pace was too fast to be effective for the coordination of RWL. This claim was validated using experimental trials based on a mismatch detection task, where the oral reading and intermediate reading rate was found to be far superior during the trials. McMahon’s supposition of why this was the case was such that the read along reading rates are too fast, i.e. well beyond the target audience’s own reading rate. Therefore a difficulty arises for the early reader to make the connection between the graphemes they see and the phonemes they hear, their own proposed intermediate reading rate being a good trade-off, as the oral rate is perhaps too slow for listening enjoyment and engagement with the content/story [199].

These studies from the last two Sections (i) eye movements studies as presented in Table 6.1 and (ii) RWL rates by McMahon et al., are valuable for considering design aspects surrounding appropriate rates within the proposed applications. Using this knowledge it is possible to derive suitable rates for customizing reading rates for varying degrees of proficiency, in order to best provide for simultaneous listening and reading in coordination within the end-applications with synchronized rich media content. Such a rate is effectively an extension of RWL to incorporate aspects of visual text synchronism. This information about fixation durations, number of fixations, and reading rates, as reported above, provide good supporting evidence upon which to base appropriate levels of WPM rate within the Controlled Recording stage of the Automatic Content Generator (ACG) system, which will be described later in this Chapter.

6.0.1 Cross-Modal Fusion

The human brain processes an immense amount of information every moment from different sensory channels. This information informs the brain about the environment through visual, auditory, somatosensory, olfactory and gustatory cues [102]. Often a single modality provides not enough information to come up with coherent and robust estimates of percepts. The reason for this is that perception is usually a multi-sensory phenomenon [81,330]. The brain analyses and combines information from a stream of sensory information to form a unified interpretation. The different modalities complement one another, which has the effect of increasing the informational content attained and goes towards resolving ambiguity [81]. There are many studies on the topic of cross-modal integration of sensory information, most notably auditory and visual; examples include studies by Ernst and Bülthoff [81], Fujisaki et al. [89], Wuerger et al. [330], Groh and Werner-Reiss [102] and King and Calvert [147]. These theories suggest that the fusion of sensory information from multiple cues (senses) has the effect of producing more accurate information about estimates and increases sensitivity.

Moreover, the coordination and integration from these different sensory channels allows for gaining a unified perpect. The combination of these different sensory channels allows for forming
a unified perception of the environment, which in fact leads to improvements in the detection, localization and discrimination of stimuli received [147]. A typical example of where cross-modal interaction plays an important role is in communication. When we are watching and listening to somebody speak, two sensory information cues integrate to give a unified precept of information. Related to this, there is a perceptual phenomenon which is known as the McGurk effect. The McGurk effect is a term given to a phenomenon that demonstrates the perceptual fusion between auditory and visual (lip-read) information [276]. This was an original observation that was presented by McGurk and McDonald in [196]. The experiment involved the following: in the audio domain the spoken syllable /ba/, and dubbed in the visual domain the lip movements (visemes) for the syllable /ga/. This resulted in the observers reporting that the syllable /da/ was uttered. Thus the McGurk effect strongly demonstrates the perceptual interconnection that exists between the auditory and visual cues.

This is quite intuitive as it can easily be appreciated that enhancements occur during speech communication when people can see each other as opposed to only hearing them. There are significant gains due to informational cue from lip movements (visemes), facial expressions and gestures etc. This topic was mentioned Chapter 2 and illustrated in Fig. 2.2. Relating all of these aspects to the technology and applications of this research is relatively straightforward. The technology that this thesis proposes uses audio in a complementary manner with its visual counterpart. This is presented through a cross-modal fusion of visual and auditory information, as well as the kinaesthetic modality which all form part of the interactive rich media Virtual Learning Environment (VLE). The McGurk effect clearly demonstrates the relationship between coordinating and mapping auditory phonetic sounds with visual visemes (lip movements). This is somewhat similar to this technology, whereby it co-ordinates and associates the auditory speech sounds with visual words in synchronism.

### 6.0.2 Learning Styles and Modalities

The theory on Learning Styles and Modalities effectively suggests that pupils/students have a certain preferred learning modalities, where the term modalities usually refers to perceptual sensory information. Examples of three major learning modalities are: visual, auditory, kinaesthetic. Liu and Ginther define cognitive learning style as “the individuals consistent and characteristic predisposition of perceiving, remembering, organizing, processing, thinking and problem solving” [175]. Keefe provides another definition, “learning styles are characteristic cognitive, affective and physiological behaviours that serve as relative stable indicators to the learning environment” [143]. In general, the theory suggests that pupils have preferred learning modalities. Different examples of the three main modalities are presented in Table 6.2.

There is a vast amount of literature on the topic of learning styles and learner differences. The interested reader is referred to the texts for more information [150, 291, 143, 78, 142]. The names of the different learning styles has evolved over the years for instance visual to global
Table 6.2: Different learning activities and resources for each of the three main learning modalities.

<table>
<thead>
<tr>
<th>Learning Modality</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visual</td>
<td>Demonstrations, Highlighting in text, Flash cards, Diagrams, Video and Animation</td>
</tr>
<tr>
<td>Auditory</td>
<td>Audio tapes, Reading aloud, Oral instructions, Using rhythmic sounds, Poems and rhymes</td>
</tr>
<tr>
<td>Kinesthetic</td>
<td>Physical examples (Props), Games, Plays, Role play, Problem-solving, Field-trips,</td>
</tr>
</tbody>
</table>

and auditory to analytical [291]. As well as these, additional theories suggest other styles to be present and relevant, a review of these has been presented by Stahl et al. [291] listing:

1. Simultaneous / Sequential
2. Correcting / Compartmentalizing
3. Inventing / Reproducing
4. Reflective / Impulsive
5. Field Dependent / Field Independent

Dunn et al. [78] takes this topic further to define a model for learning styles which factors in aspects related to environmental, emotional, sociological, physical and psychological characteristics. The concept of mapping the appropriate learning style to the individual preferred style, in order to enhance their learning capabilities has been advocated by many authors [291,150,78] and referred to as the *meshing hypothesis* [226]. More recently Pashler et al. [226] have questioned the validity of the learning styles approach, in particular the meshing hypothesis. They take issue with incorporating learning style assessments into the general educational practice claiming that there is a lack of supporting evidence to support its utility and effectiveness. The usual scenario being that an individual's learning style is first identified and then the instruction is tailored based upon this particular style [226]. The author appreciates both of these viewpoints in the debate on the learning styles and modalities, and is approaching the area from somewhat of a neutral standpoint. To this effect the proposed educational technologies provide for a complementary fusion of the three major modalities, visual, auditory and kinaesthetic, in such a manner so as to enrich each others analysis, integration and association. This aims towards catering for many different methods and modalities of learning in an engaging, enriching and collaborative learning environment. There are many similarities between what this technology offers and traditional teaching methods. For instance the concept of highlighting text in synchronism with the audio narration in the VLE is similar to the teacher or parent reading
6.1 Content Generation System

The automated content generation methodology depicted in Fig. 6.3 provides for the automation of the audio recording and e-content generation for production of digital educational resources. Automation is of high value for the creation of rich-media educational content as it effectively streamlines the complex audio segmentation and visual synchronization procedures. These tasks are usually performed manually and tend to be extremely laborious, time consuming, and inefficient for large scale production and interactive media resources creation. Hence the proposed system is aimed towards facilitating the efficient generation of high quality, rich-media based educational e-content and resources. The technology comprises two main parts:

1. A methodology for segmenting and synchronizing words with their corresponding visual text counterpart has been developed.

2. A rich media application player that the learning environment runs within. This player is suitable for use on computer desktop and Internet web browsers.

The technology is not application specific and can be fully customized for the production of rich media applications that are both desktop and Internet based. The example platforms investigated include Adobe Flash, HTML, JAVA and AJAX. As alluded to in earlier Chapters, there are two formats for the ACG system deployment, high-end Professional Studio Generated Content (PSGC) and low-end User Generated Content (UCG).
Figure 6.3: The Automated Content Generation Methodology
6.1. Content Generation System

1. Professional Studio Generated Content (PSGC). This system is for high-end use within a sophisticated audio studio recording setting to generate professional (commercial) digital educational content and resources. As a result the system is specifically customized towards use by digital production houses and educational publishers.

2. User Generated Content (UGC). This system has been described previously for use within low-end settings. However the underpinning technology of this system is similar in many regards to that of the PSGC. The term low-end effectively refers to the system being deployed as a desktop application on a user's computer (PC, Laptop, Touch Screen Tablet). In this setting the audio recordings are made using peripheral device based microphone or computer headsets that connect straight into the PC (via audio jack and USB connectors).

Looking again at Fig. 6.3, the process contains six important elements, which are as follows:

1. Voice Recording. The recording stage involves the acquisition or capturing of natural voice recordings from a microphone. As part of this research, it was found that current state of the art speech synthesizers perform poorly when set to the lower WPM rates that are suitable for the purpose of highlighted text within educational applications. The research has found many important benefits of using natural recordings over synthesized speech at slow reading rates, which become apparent when listening to the audio upon playback due to poor performance in pronunciation, prosody, stress and intonation. Although synthesized voice engines are state of the art, they are unable to handle the slow word rates that are appropriate for the purpose of educational resources. The synthesizer sounds unnatural, but it has found uses in other stages within the process, such as a role in the controlled recording process as well as playing an important part in the speech segmentation algorithm.

2. Controlled Recording. This controls the rates at which the narrators (actors) read the text. This should be at a suitable rate for the purpose of highlighting the text. A trade-off exists here, such that the rate at which the text is enunciated during recording should be slow enough to facilitate meaningful word highlighting. Yet it should also be fast enough to maintain a suitable level of speech pronunciation qualities of prosody, intonation, and fluency, so that the text is read in a manner that has meaning and is engaging for the listeners (end-users). Different moderation strategies have been experimented with in this work, which have centred around exploiting a modern speech synthesizer to give appropriate cuing information for controlling the speech rate. This can be achieved in three ways which are described below. Firstly, it is worth pointing out that modern synthesizers, such as the state of the art AT&T speech engine used in this work, have

---

1 Recently, tablet PC's have become popular. Examples of these devices include: Samsung's Galaxy Tab, Apple's iPad and Dell's Streak.
Applications: Integration and Implementation

the inbuilt facility to adjust the speed2 of the voice. The facility to adjust the rate is beneficial as we can adjust it to the required WPM, as governed by the requirements of the end application (literacy levels, language learning, medieval languages studies). Guidelines for assigning appropriate rates have been discussed in the section above on eye fixation, saccades durations and reading while listening (RWL) rates. The user can use this synthetic speech as cue information, which is quite similar to the way that a metronome is used by music composers for maintaining consistent tempo around a fixed beat. The three ways that this synthesizer can provide appropriate cue information are: (i) audio cues, where the narrator listens to the synthetic speech in their headset and reads along in unison with the synthetic speech (ii) visual cues, similar to the audio cues except the visual highlighted word is used as the cue information and (iii) mixed audio-visual cues, which is a combination of the first two types. All three have been experimented upon and the latter was found to be the best compromise. The visual cue on its own can be quite discrete in nature (highlighting words visually) which can be off putting to the voice actor/narrator, which can result in some loss of natural qualities in their speech. Audio signals have a more continuous nature as a cue, which in fact works quite well in conjunction with the visual cue as a supplementary aid.

3. Segmentation Process. This stage performs automatic speech segmentation at the boundaries of meaningful linguistic units, such as the individual word boundaries. This is for the purpose of synchronization and tablet/touchscreen tracking. The task of automating the segmentation procedure was technically the most difficult aspect of automating the production process. The underlying, motivation, theory and algorithmic processes that underpin this stage have been covered in detail in earlier Chapters of this thesis.

4. Synchronization Process. Text synchronization comprises of text based 2D animation, which is achieved by adapting or modifying the text in a suitable manner so that the words are highlighted accordingly. The term highlight is used rather loosely as there are many different schemes for drawing the viewers attention to a particular word in synchronization with the audio. The more notable ones that have been investigated include: (i) Times Square scrolling, (ii) Rapid serial visual presentation, (iii) Colour highlighting, (iv) Opacity highlighting, and (v) Magnification highlighting, all of which are illustrated in Fig. 6.4. Of these five techniques, the most robust and effective compromise was found to be a hybrid combination of the colour highlighting used in conjunction with subtle magnification highlighting of each word. This was determined empirically from laboratory experimentation to be a suitable framework for use within the system.

The magnification of each word is only in the vertical direction, i.e. vertically scaled by a percentage amount, determined empirically (typically 5-10 percent). The reason for

---

2In addition to adjusting the speed some other synthesizers offer the facility to adjust the pitch, e.g. Microsoft speech synthesizers.
6.1. Content Generation System

Figure 6.4: The different types of text presentations and highlighting that have been considered: (a & b) Times Square Scrolling (c & d) Rapid Serial Visual Presentation, (e & f) Text Colour Highlighting (g & h) Text Opacity Highlighting and (i & k) Text Magnification Highlighting. Credit: graphics by Liam O’Sullivan, TCD.
this is that horizontal increases in font sizes or scaling causes a horizontal shift of all the
other letter/words in the text stream, which can become quite off putting and visually
disturbing for the end-user.

The use of colour highlighting required taking different criteria into account, such as Lu­
minosity Contrast Ratio (LCR), Colour Difference (CD) and Brightness Difference (BD),
as defined by the World Wide Web Consortium (W3C) in [319,320].

LCR is given by,

\[
L_n = .2126 \times R_n + .7152 \times G_n + .0722 \times B_n
\]

\[
LCR = \frac{(L_1 + .05)}{(L_2 + .05)}
\]  

(6.1a)  

(6.1b)

where R, G, and B are linearized values, e.g. \( R = \left( \frac{R}{FS} \right)^{2,2} \), where FS is the full scale value
(255 for 8 bit colour channels).

Colour Difference (CD) is given by,

\[
C_D = [\max(R_1, R_2) - \min(R_1, R_2)] + [\max(G_1, G_2) - \min(G_1, G_2)]
\]

\[
+ [(\max(B_1, B_2) - \min(B_1, B_2)] > 500
\]

(6.2a)  

(6.2b)

Brightness Difference (BD), given by,

\[
B_n = \frac{(R_n \times 299) + (G_n \times 587) + (B_n \times 114))}{1000}
\]

\[
BD = B_1 - B_2 > 125
\]

(6.3a)  

(6.3b)

In relation to font colour and highlighting, it is important to adhere to the guidelines
mentioned above (LCR, CD and BD). Offering additional choices between font colour
schemes is important to cater for some related eye-sight impairments. Studies have been
conducted in the area of font styles (sans serif versus serif etc) and font sizes. Sans Serif
are thought of as being easier to read, [67, 30, 29, 95]. A study by Bernard et al. [29]
used the fonts: (i) Times New Roman, Courier New, Arial and Comic Sans MS for two
different font sizes, 12 and 14 point. They found that Comic Sans MS at size 14 point to be
"easier and quicker to read, as well as being more attractive and more desired to be used
in schoolbooks" compared with the others studied. The types of fonts used in this work
were also heavily influenced by educational aspects, where for the literacy and language
learning, double-storey letters, e.g. a’s and g’s were considered to be more inappropriate
than single-storey a’s and g’s, see Fig. 6.5. Also old style medieval letters were used with
regard to: The Book: Discovering Sounds Initiative Application. These letters will be
encountered later in this Chapter in the display images which showcase the applications.
5. **Evaluation.** This process allows for monitoring and evaluation of the material. Firstly, in the PSGC case there are some automated procedures that can determine measures from statistical analysis of the segmented time values to enable monitoring for quality control purposes. A good example of this is the WPM rate (reading pace). If the WPM is too high or low the sentence or phrase might need to be re-recorded. This can be automatically prompted on screen to the media designer and the narrator, suggesting a re-recording of a particular sentence within the passage. Also there is the facility for the media designer to experience what the media content looks and sounds like upon play back in synchronism with the visual content in rich media format similar to the end application. This is quite an important piece of inbuilt functionality, as it allows the freedom to view the content during the production process and to act accordingly, i.e. make requests, suggestions and recommendations surrounding re-takes if necessary.

For the case of the technologie's role within the UGC framework, much of the same functionality is present. However, the evaluation stage within a UGC system is proposed to be less sophisticated in nature and to be more of a simplified piece of functionality. First, the speech rate (pace) can be checked automatically and benchmarked against suitable values for the target audience, for instance the intended readers proficiency level. The narrator (voice artist) is then prompted if a retake is necessary. However, it should be stressed that it is only on rare occurrences that a re-recording is necessary as the Controlled Recording stage alleviates many of the issues around speech rate control. It is more to the point that the users have the added advantage of being able to make re-recordings easily at their own discretion.

6. **Digital Technologies for Education.** The technologies and resources produced by the ACG system have been explored in conjunction with three important applications which are discussed in greater detail in the next section. These are as follows: (i) The ReciTell

---

Figure 6.5: Lowercase ‘a’ and ‘g’ for double and single-storey variants. Credit Jim Hood.
Application, which is specifically aimed towards the PSCG framework for the purpose of professional and marketable literacy and language learning products and educational technologies. (ii) The iMARK Application, which is a UGC tool that has been created in collaboration with this research for the generation of rich media based learning objects\(^3\) using custom-designed desktop software, and (iii) The Book: Discovering Sounds Initiative (TBDSI) is also described, which in many respects is quite similar to the ReciTell Application as it is more aimed towards the high-end PSCG framework, which requires a certain degree of specialist expertise from media designers and sound engineers respectively. However, there are some avenues that are briefly alluded to that could be further explored in a UGC context.

6.2 Applications

6.2.1 Recitell Virtual Learning Environment

6.2.1.1 Background Information

ReciTell Ltd is an early stage campus company from Trinity College Dublin, which was established to exploit innovative educational technology that was developed as part of this research; see the logo graphic in Fig. 6.6. ReciTell is a Business to Business venture for partnering companies such as Educational Publishers and Media Production Houses. This type of open innovation partnership is commonplace in industry for developing highly innovative products with clearly defined Unique Selling Points (USP’s). Collaboration of this nature leverages a favourable balance between market pull and technology push \([60]\), towards the research, design, and development of innovative solutions that best meet the needs of end-customers. Some relevant examples of industry interaction include: (a) Collaboration between C.J. Fallon, a leading Irish publisher, and Houghton Mifflin Harcourt (formerly Riverdeep) in developing an Irish curriculum-specific version of Destination Maths Software, (b) Pearson’s acquisition of eCollege, Ordinate Corporation, Effective Education Technologies, PowerSchool and Chancery in recent years, (c) McGraw-Hill’s acquisition of Turnleaf Solutions in 2005, and the (d) Merger of Riverdeep educational technology company with Houghton Mifflin Harcourt Publishers.

The ReciTell’s technology is aimed at K-12 markets for literacy and language learning products. The creation of ReciTell’s novel technology has been developed in a practical and pragmatic manner through fostering close relationships with both teachers and educational publishers. Through these connections a good knowledge and understanding was obtained surrounding the importance of naturalness, robustness and teacher-student engagement when developing innovative learning environments.

\(^3\)The term learning object is defined as digital or non-digital resources that can be used for learning, education or training \([119]\).
6.2. Applications

Figure 6.6: The logo graphic created for the ReciTell VLE Application.

6.2.1.2 Problems Addressed

The ReciTell application addresses important issues surrounding poor literacy levels and the difficulties in the acquisition of second languages, in particular focusing on English as a Second Language (ESL). There is international recognition of the problem of poor literacy levels in many developed economies, which is compounded by the cost, in time and resources, for the provision of effective teaching methods. There is also an apparent absence of objective literacy assessment methods at present. The ReciTell’s Virtual Learning Environment (VLE) Application, is a novel technology that is aimed at solving important market needs around the areas of literacy and language learning. Some key areas that have been identified where market needs currently exist include:

1. **English for Speakers of Other Languages (ESOL)**. There are many challenges at present facing teaching, in Ireland and the United Kingdom, about the number of children who do not have English as a first language. Evidence of this was reported in the latest Annual Schools Census 2009 data [73] for England, where 15.2% of primary pupils and 11.1% secondary students do not have English as a first language. Teachers may experience challenges, since many lack the necessary resources to bring these students up to the same literacy levels as the average students. Increased class sizes and budget cuts to support services also exacerbate these challenges. The ReciTell VLE offers great potential for
helping ESOL pupils improve basic literacy skills in a self-paced learning environment. This technology offers learning modules and resources that can be used in the classroom setting, as well as in the home environment.

2. **Challenging Higher Abilities.** Teachers usually aim their lesson plans at average pupils. As mentioned before this can present challenges for struggling pupils in class both ESOL and otherwise. Furthermore it also has implications for challenging higher performing pupils, as they can become demotivated and lose interest in class work. ReciTell’s technology offers a low cost interactive computer software solution that can be used to challenge higher performing pupils, with higher abilities, with extended reading, comprehension and assessment resources. This can be used in a complementary capacity to enable the teacher to focus on the mainstream pupils. Technology’s use in this regard has been discussed by Moe and Chubb [207], stating that schools that rely more on technology can be run relatively cheaply compared to the alternative, which is labour intensive resources, that tends to be much more expensive.

3. **No Child Left Behind (NCLB).** This education initiative was signed into law in the United States in 2002, and requires states to implement standards-based assessments in reading and mathematics for pupils in grades of 3-8. There is also a requirement that all states participate in National Assessment of Educational Progress tests in 4th and 8th grade, for Reading and Mathematics [214]. This has increased accountability considerably in schools across all states, since it places schools under increased pressure to ensure pupils achieve baseline proficiency in the areas of Literacy and Mathematics. The impact has created greater need and market demand for quality interactive educational software that can be used to assist teaching and perform assessments in schools. Statistics have shown that pupils who use classroom software perform better on standard tests, the biggest gains occurring among the low-achieving students [174,240].

4. **Basic Literacy and Foreign Language Learning.** Recently, primary and secondary schools in the UK have installed Virtual Learning Environments (VLE). The strategy behind this initiative was to bring more learning exercises and activities into the home environment. Although this infrastructure is in place, its effectiveness is severely limited due to the lack of quality learning resources and modules on the market at present. The ReciTell system is programmed in Flash and can be easily integrated into a web-based VLE to provide interactive learning materials and resources to support teaching basic literacy and for foreign language learning, e.g. French, Spanish, German, etc.

6.2.1.3 **Proposed Solution**

The ReciTell Virtual Learning Environment (VLE) application places particular emphasis and attention on each of the main learning modalities. ReciTell’s technology incorporates a comple-
6.2. Applications

Figure 6.7: The ReciTell Virtual Tutor System

mentary fusion of visual, auditory, and kinaesthetic learning modalities, to provide an enhanced learning experience for the pupil. The automated content generation system, within the PSGC context, is used to facilitate the rapid production of ReciTell's applications.

An example of the ReciTell VLE application is given in Fig. 6.7. This application operates within the rich media based Flash environment. As observed in the figure, the VLE is rendered as a sophisticated interactive digital book. This virtual book has the same appearance as a physical book on screen with a hard outer cover (hardback) and flexible interactive pages. The functionality is included to be able to turn the virtual pages as illustrated in the diagram by clicking the cursor on the corner or the edge of the page and dragging across to turn the page. There is also the option to navigate using the control bar as shown, with semi-transparent grey buttons next and back along the bottom of the interface. The audio accompaniment is controlled using the play and stop buttons in the control bar. During the audio narration the words are visually highlighted in synchronism. This is shown in Fig. 6.7, where the word /little/ is being highlighted in the contrast black along with a slight adjustment in vertical font size as was described previously in ACG system. This delivery platform, with fusion of synchronized audio and visual information, creates an engaging and immersive learning environment for pupils. This system aids pupils in associating the written and spoken word in a manner which is similar to traditional teaching methods of words, following in a one-to-one teacher pupil session, referring back to Fig. 6.2. The style and format of the VLE rendered in a virtual book format has many aesthetically pleasing aspects, which sustain some of the same experiences encountered
Applications: Integration and Implementation

The gingerbread man.

An old woman had some flour and some ginger.

She made a gingerbread man.

Figure 6.8: The tablet based assessment system (a) the overlay assessment sheet, and (b) the stylus position relative to time which is indicated by a colour coded scheme, where time is with respect to the audio narration.

Coupled as part of the ReciTell VLE is a novel tablet based assessment system which has been investigated for measuring the pupil’s ability to follow the passage that has been read aloud (audio). This is achieved using a peripheral computer based table device and stylus as indicated in Fig. 6.8. The sheets of text as shown in Fig. 6.8 form overlays for the tablet. The assessment system works as follows: As the audio is played to the pupil they are required to follow the text by moving the tablets stylus in synchronism with the audio narration. This process can measure the pupil’s ability to keep up with the audio reading. This can be further supplemented with the addition of sentence branching, which can be used to test the pupil’s listening comprehension and word association. As they are listening to the text being read, at a slow rate, they move the stylus along the correct path corresponding to what has been read during the audio narration. Upon completion of the text, the system calculates an assessment metric which relates the position of the pupil’s stylus and the correct path. The results (scores) obtained from the assessment are then displayed to the teacher in an intuitive graphical display, which can be stored or logged in memory. In this way progression can be closely monitored for each of the pupil’s.

In addition, this tablet based system has somewhat of a dual-purpose in that as well as providing the assessment functionality, as described above, it can also provide a controlled reader functionality, where the tablet can be used to read the words back to the pupil as they...
move the stylus along the bottom of the words. This empowers the pupil by enabling them to
dictate the pace of reading and have any words repeated if necessary. This is a novel way of
introducing a form of kinaesthetic learning modality into the process of teaching pupils to read
through utilizing a form of stylus dexterity skills.

Some technical details of the underlying software which supports both of these functionalities
described above are: firstly, the segment boundaries of the words in the audio stream are known
to the system as the speech segmentation algorithm has given these segmentation boundary
points. These segment boundaries in the audio stream can be related to the two dimensional
spatial positions of the words on the tablets surface overlay. These spatial positions are known to
the system beforehand, as the overlays are generated in a controlled way that register the words
spatial positions. Thus the two can be inter-related to one another and interpreted appropriately
by the software algorithm.

6.2.1.4 Outcomes

Working prototypes of the proposed technologies have been developed, demonstrated and trialled
in Schools by Educational Publishers C. J. Fallon. C. J. Fallon is the leading primary school
publisher in Ireland and has also gained a strong presence in the secondary school market. They
produce the most widely used national tests, namely, MICRA-T (for Reading) and SIGMA-T (for
Mathematics), in the primary schools market. There have been a number of demonstrations and
appraisals with their school representatives, educational advisors and senior management. The
outcome from these demonstrations has been very positive with a strong interest in collaborating
for the integration of the technology into a new literacy programme. C. J. Fallon has provided
soft-copy versions of their content, text, illustrations and graphics for use within the development
work. The ReciTell VLE has great potential for stand-alone use as a modern educational software
package and to be integrated into a modern Learning Management System\(^4\) (LMS) for schools.
There are many examples of these including: Moodle, Blackboard, and WebCT.

6.2.2 iMARK: interactive Media Annotation Resource Kit

6.2.2.1 Background Information

iMark stands for interactive Media Annotation Resources Kit, which is an acronym encompassing
the tool's role within the context of media based annotation and creation of UGC; see Fig. 6.9.
This project has built upon the experience of a pre-existing learning resource that was in use
in the School of English at TCD. This system was the outcome of a Learning and Innovation
project developed in 2004. The core of this system contained digitized pages from the Trinity
Manuscript MS 212, the C-text of the allegorical poem by Piers Plowman (c. 1400). Interactive
content was also added to show what the manuscripts can tell us about the author of the text,

\(^4\)SCORM is the industry standard for LMSs [222].
the scribe, as well as uncovering information about the people who owned it over the years, who sometimes glossed or annotated the text along the sides of the pages. Up until recently, the system has formed an integral part of the English course, *The Book: Rethinking Learning*.

### 6.2.2.2 Problem Addressed

This early resource had a fundamental limitation surrounding re-usability. As the system was designed by an experienced flash designer, it was not possible for the academic to change, update, or to add new content. This limited the educators' creative freedom to revise learning resources due to the specialized environment upon which it was created in (Flash/Actionscript).

### 6.2.2.3 Proposed Solution

The solution explored to solve this problem involved developing an intuitive software program which empowers educators to create their own interactive learning resources and materials. These learning resources are rendered as an interactive digital book in a similar style to the ReciTell VLE system. The iMARK resource allows the educator to add interactive annotations to the document at different locations. Before proceeding it is important to define what is meant by the term *annotation*. An annotation can comprise of:

1. **Text field inserts** - with additional information, explanations, comments, transcriptions, and translations.

2. **Uniform Resource Locator (URL) links** - to additional resources, websites, and even to other iMARK applications.

3. **Audio inserts** - audio accompaniment to assist in explaining the content can be added via MP3 recordings.
Figure 6.10: The three stage process involved in creating an interactive learning resource using the iMARK Application.

The educator has the facility and freedom to place (populate) annotations at relevant points of interest in the interactive document. This in turn is rendered as an Interactive Digital Book, which can be integrated into a Internet browser based VLE (HTML) or can run as computer desktop application using the Adobe Air runtime environment\(^5\). The annotations appear as small subtle white squares (icons) in the interactive book. The end-user can then click on these annotations icons to present the annotation dialog box with the content, text, URLs, audio.

The process to create these \textit{learning objects}, using iMARK, is illustrated in Fig. 6.10. This is a straightforward procedure that is intuitive and user friendly. It can be broken down into three stages as follows:

1. \textbf{Content}. The educator starts the process with content that they wish to transform into interactive learning objects, using the iMARK system. This content comprises of JPEG images of the material they want to use to form the pages of their interactive book. As well as the content in relation to the annotations such as the text fields for the annotations, the MP3 audio files either acquired or recorded themselves, selection of URL links to additional websites, and to other iMARK resources.

2. \textbf{Authoring Tool}. The iMARK \textit{Authoring Tool} is a desktop application where the \textit{learning objects} are created. The tool runs within the Adobe Air\(^6\) environment, which can run on most platforms including: Windows and Macintosh. It is very intuitive to use, which makes the creation of iMARK learning objects both quick and easy. The user has the freedom

\(^5\)The Adobe Air runtime environment is free to use and easy to download and install.

\(^6\)Adobe AIR, also known as Adobe Integrated Runtime (AIR), is a cross-platform runtime environment developed by Adobe Systems.
to insert their interactive annotations throughout the virtual book (learning object).

3. **Interactive Learning Resource.** The iMARK interactive learning objects that are produced by the authoring can also run within the Adobe Air runtime environment as a computer desktop application. Alternatively, the learning objects can run within a web browser using a built-in Flash player. This is an advantage as it allows the iMARK learning objects to easily form part of any educational website.

### 6.2.2.4 Outcomes

The iMARK application forms part of National Digital Learning Resources (NDLR) repository and is available from the NDLR website\(^7\). The resource was launched at the NDLR Conference in May of 2010 at Trinity College Dublin. Already, the resource has been utilized across many fields and disciplines for the creation of learning objects. Although the application was motivated by studies in the School of English on Early Printed Books and Manuscripts, such as the Piers Plowman resource, the iMARK application has a certain degree of generic usability and can be used to create learning resources in many areas. It is currently one of the NDLR’s top ten learning resources and has 40 users as of September 2010. Two showcase examples of learning objects are presented in Fig. 6.11. To highlight the potential and versatility of the resource, it is worth mentioning other areas where iMARK could be utilized, such as in the fields of: (i) *Astronomy*: annotation of images of planets and star constellations, (ii) *Mathematics*: annotation of mathematical theorems with explanations, corollaries and notational information, and (iii) *Anatomy*: annotation of human anatomy images\(^8\) for training and instructional purposes.

While the applications for iMARK have been discussed, there is however further scope for it to overlap with other areas addressed in this research. Proposed extensions to the iMARK technology could provide a tool for the creation of UGC for literacy and language learning applications. The current version of iMARK allows users to create their own content and learning resources, i.e. UGC as described above. Although the underpinning technology that has been exploited in the ReciTell system is aimed at professional product based literacy and language learning which is generally directed towards educational publishers, the technology that is utilized within the ReciTell system could be integrated into the iMARK system so that educators, instructors and parents/guardians could create their own interactive learning objects for literacy and language learning applications.

This would have suitable applications in language learning, so that teachers, lecturers and parents/guardians could create their own reusable literacy and language learning objects. They would have the facility to read in the text themselves, which the system could capture using a microphone and automatically interpret and transform it into an interactive audio book that has the synchronized text similar to Recitell’s VLE resource. This ability would add unique

---

\(^7\)The NDLR website may be located at the URL: www.ndlr.ie.

\(^8\)Examples include Magnetic resonance imaging (MRI) scans.
6.2. Applications

Figure 6.11: Showcase examples of learning objects created using iMARK, (a) Land War Document learning object by Fota House Arboretum and Gardens, and (b) Magazine of Magazines 1751 learning object by The Eighteenth Century Research Group, University of Limerick.

benefits, as learning from a familiar narrative voice, such as a parent/guardian or teacher would create a more reassuring and supporting learning environment for the pupil. This would help keep the pupil’s attention and interest in the reading and listening educational activity.

Another area where this UGC tool would have valuable applications is adult literacy learning, where there is presently a serious dearth of appropriate content targeted towards illiterate adults. The vast majority of literacy content and resources is aimed towards pre-literate children, with themes including picnics and fairy tales, which is not appropriate or relevant for adult literacy learning. This is a major problem that poses a limitation on the effectiveness of educational materials and resources which are not customized to meet the needs and stimulate the interests of adult learners. This is not a new problem or unique to digital learning resources as much earlier Knowles [151] reported that much of the theories of adult learning were heavily based on research into children’s learning, which was in fact originally founded upon studies of animal learning.

6.2.3 The Book: Discovering Sounds Initiative (TBDSI)

6.2.3.1 Background Information

This project has formed part of a research initiative for the Trinity Long Room Hub. It has explored a novel means of annotation and interaction with content in the field of Arts, in particular studies relating to manuscripts and Early Printed Books (EPB), using advanced rich media (audio-visual) based technology.

9The Trinity Long Room Hub is Trinity College Dublin’s (TCD’s) Arts and Humanities Research Institute.
6.2.3.2 Problem Addressed

The TCD library collection is richly diverse and is home to more than 20,000 collections of manuscripts and archives, dating as far back as the 13th century B.C. There are many prestigious items within this collection, such as the Book of Kells, The Book of Durrow and the Book of Armagh. Two example pages from the Book of Kells are presented in Fig. 6.12. All of these manuscripts and EPBs are extremely precious and valuable and as a result of this they are highly limited in terms of their accessibility because of concerns in relation to preservation and protection from damage and deterioration.

6.2.3.3 Proposed Solution

The project has focused on making these precious manuscripts, EPB and archives more accessible to academics, students and to a wider audience. An application has been developed to display these valuable manuscripts and EPB's in digital format, accompanied with additional synchronous audio visual media, to create a more informative and engaging learning experience, see Fig. 6.13. In addition to making the manuscript more accessible, these audio visual elements enhance the learning outcomes from the manuscript. The project, as an initial pilot, has focused on taking scanned images of pages from an important TCD Library manuscript, MS 212 titled *Piers Ploughman* C-text written by William Langland, which is an allegorical poem from the fifteenth century (c.1400) [164].

The solution involved taking page scans and transforming them into an interactive medieval book that has a similar format and features to the other applications described, and is rendered as a sophisticated virtual book. This gives great aesthetic appreciation of the manuscript and

---

Figure 6.12: The images from left to right, The Old Library in the Long Room Building in Trinity College Dublin, The Book of Kells: Folio 32v shows Christ enthroned, The Book of Kells: Folio 309r from the Gospel of John written in Insular majuscule by the scribe known as Hand B.
resembles the real thing, although some purists would argue that aspects such as touch, smell, relative size, weight etc, are part of the real thing which are not applicable on this interactive digital format. The addition of a zoom functionality has been incorporated to allow close up views of the manuscript for research purposes. A professional audio narration of the some of the early passages from the Piers Ploughman M.S. has been recorded, along with an introduction describing the manuscript. These audio recordings of the passage, which is in middle English, have been provided in synchronism with text highlighting of the words. This is achieved using suitable computer based typographical fonts overlaid on top of the manuscript with the corresponding scribed version of the text being emphasized immediately above; this is shown in Fig. 6.13d. Since the text of the manuscript is in middle English, it aids with interpretation and transcription and can also teach people to read the old English alphabet in a scribed format, e.g. characters include thorn /θ/ and yogh /ȝ/.

6.2.3.4 Outcomes

The Book Discovering Sounds Application now forms a valuable learning object for Undergraduate English Studies at TCD. This interactive resource, with synchronized audio visual media, provides an enriching way for students to experience medieval English in its original manuscript form. The resource has important benefits for helping to improve interpretation and transcription. There are also aspects of increasing accessibility to precious manuscripts to a much wider audience, which was not possible before due to preservation issues and concerns. This was an exploratory pilot project that has focused on using one important manuscript (i.e. MS 212 Piers Ploughman) from the TCD collection, however, using the ACG methodologies described earlier in this Chapter there is great potential for expanding this towards transforming larger volumes of manuscripts into similar formats.

6.3 Conclusion

This Chapter began by examining some of the background theory which has supported the implementation of the Automatic Content Generation technology. By the same token, this theory has also contributed to the development of the educational applications that have been explored. These areas include, (i) eye movement characteristics, (ii) appropriate reading rates, (iii) some cognitive aspects for audio-visual simultaneity, and (iv) some discussion on different learning styles. The ACG system has been described in detail. The three main educational application areas, ReciTell, iMARK and TBDSI have been described under four areas: (i) Background information, (ii) Problems addressed, (iii) Proposed solution, and (iv) Outcomes. Showcase examples for each application have also been presented.

11This particular manuscript contains many annotations along the sides of the pages that has been made by the owners of it over the years.
Figure 6.13: The Book: Discovering Sounds Initiative (TBDSI) Application for the Piers Ploughman M.S. 212 (a) the front cover of the virtual medieval manuscript, (b) interactively navigating through the manuscript, (c) synchronous audio-visual media reading of the text in middle English, and (d) a close-up view of (c), the original scribed version of the text is highlighted above and the computer based typographical text is placed below with synchronized audio-visual highlighting accompaniment. Also notice in (d) the middle English alphabet character thorn /th/.
Automated Condition Monitoring of Rotating Machines

This Chapter describes a separate research problem that has been addressed in the area of Automatic Condition Monitoring (ACM) for rotating machines. Firstly, an introduction to the area of ACM is presented. The topic of acoustic noise is then examined in relation to Rolling Element Bearings (REB) with rotating machinery. The Chapter then describes two approaches to feature extraction that have been investigated, namely Envelope Analysis (EA) and a Data-Driven method. Later in the Chapter details on the Experimental Implementation are provided. Finally, the results are presented for both methods showing the performance for classifying the acoustical noise signal into wear (degradation) states.

Approaches to Automated Condition Monitoring (ACM) of rotating machinery typically involves the detection and diagnosis of defects in bearings. These techniques are widely reported upon in the literature [118, 197, 270]. While the detection and diagnosis of these defects is useful for ACM, there is an increasing demand to also predict the Remaining Useful Life (RUL) of machines. Predicting the RUL allows for (i) improved reliability of machinery, (ii) scheduling of maintenance prior to machine failure in order to prevent machine downtime, and (iii) the removal of the cost of unscheduled maintenance. The ability to predict the RUL has been explored to a much lesser extent in the literature [93, 272, 271]. ACM systems, reported in the literature, typically acquire and monitor the vibrational signal, both low frequency vibrations and acoustic emissions (high frequency vibrational signal) [118, 197, 270, 93]. The acquisition of such vibrational signals requires physical contact with the machine being monitored. However,
it is often the case that in real world conditions it is not always desirable or possible to acquire good quality vibration signals. In this work the acoustic noise signal (< 25kHz) is monitored in order to determine the RUL. This approach is advantageous as it allows for remote, non-contact monitoring of machines. It is worth mentioning that, while acoustic signals are obviously susceptible to ambient noise, vibrational signals are also susceptible to vibrational noise from nearby sources, e.g. machinery and equipment.

A technique known as Envelope Analysis (EA) is widely reported on in the literature, for detection and diagnostics of bearing defects using the machines vibration signal [118,197,270]. This approach to signal analysis is ubiquitous in the condition monitoring industry. However, it was Kavanagh et al. in [141] that first proposed applying this approach to the task of predicting the RUL of machines using acoustic noise signals (< 25kHz). Accordingly, the work in this Chapter describes a novel implementation of the traditional EA feature extraction technique applied to acoustic noise signals to determine the RUL of a rotating machine.

However, one major shortcoming of the EA approach for feature extraction, often discussed in the literature [118,302] is that the frequency band of interest, the cut-off frequencies, must be determined a-priori. In an effort to address this weakness of the EA method, a Data-Driven approach has been proposed by Scanlon and Bergin in [271]. This approach has also been explored in [141], and is presented herein for comparison with the EA approach. This Data-Driven approach utilizes Information Theory for a technique of Feature Subset Selection (FSS), to select only these salient spectral features across the entire spectrum and hence remove noisy features from the feature vector. Thus this approach does not require the prior knowledge of the particular frequency band which contains the information that is important for characterizing potential defects.

7.1 Acoustic Machine Noise

The rotating machines used in this study are comprised of several moving parts, including two Rolling Element Bearings (REB); see Fig. 7.1 for a diagram of bearing parts. As REBs degrade over time, faults such as localized defects, cracks or spalls begin to occur in the bearing structure and a fault signal is generated. This fault signal arises due to the interactions between the different parts of the REB and the defects as they make contact with one another as the bearing rotates. This interaction (impact) produces a force impulse which excites resonances in the bearing and the machine, producing vibrations which lead to acoustic sounds being generated. These impacts occur in a successive manner as the shaft rotates, giving rise to a series of impulse responses occurring periodically at time $\tau$. The amplitude of the impulses may be modulated because of two factors: (i) the strength of impacts, which corresponds to the amplitude of the impulses, which will be modulated at the rate the fault/defect passes through the load zone; and (ii) the variation in propagation of the fault signal due to the relative location of the defect, in relation to the fixed position of the transducer (sensor) [118,248].
7.1. Acoustic Machine Noise

The spectrum of the REB signal also contains a series of harmonic frequency components. The shape and resonance properties of a machine and housing cause harmonics to be boosted or dampened, i.e. harmonics whose frequencies are close to a resonance frequency of the machine or housing will be enhanced, and harmonics whose frequencies are not close to resonance frequencies become weakened. The period $\tau$ depends on which REB part the fault occurs in, i.e. whether it occurs in a ball/roller, inner race, outer race or cage. The parts all have their own associated time duration $\tau$. The rate at which these impulses occur, $\left(\frac{1}{\tau}\right)$, is known as the Characteristic Defect Frequency (CDF), labeled here $\text{cdf}_0$. Each of the four REB parts have their own associated CDF, which is a function of REB geometries and of the shaft rotational frequency ($F_s$); refer to Fig. 7.1 for the different parts. The formulae that govern where these occur in the spectrum include the ball-pass frequency of the outer race, ball-pass frequency of the inner race, fundamental train frequency (cage frequency) and the ball-spin frequency, [270].

Note that in normal operation, slight random fluctuations can occur between the temporal durations of impulse responses, $\tau$. These fluctuations are caused by random slip, due to the cage in the bearing, ensuring that the mean speed of the REB elements remains the same. These random fluctuations, although small, can have a pronounced effect on the spectrum causing the frequency components to smear laterally [118,248]. Such spectral smearing is highly undesirable from a feature extraction point of view as it can make diagnostic information difficult to uncover.

Figure 7.2 shows the standard log energy spectrum of data sampled at different points in time over the lifetime of the machine, that is for a new machine, a machine midway through its lifetime machine and a machine immediately prior to total failure (bearing seizure). This figure illustrates that there are obvious differences in the acoustic spectrum once a defect has fully

Figure 7.1: The main component parts for a Mechanical Bearing. Credit: Benutzer Niabot for Bearing Graphics.
In this study, spectral analysis approaches to feature extraction for machine monitoring are explored. Such analysis decomposes the signal into spectral components, so that the influence of individual mechanical components, including REB parts, can be ascertained. As considerable broadband noise from rotating machines can occur in the low frequency range, below 5kHz, the fundamental defect frequency and low order harmonic resonant frequencies may be swamped within the noise and therefore not useful for monitoring. However, the higher order harmonic resonances are more likely to have a higher Signal to Noise Ratio (SNR) which makes them easier to detect and monitor.

### 7.2 Envelope Analysis

Envelope Analysis (EA) is one of the most predominant feature extraction approaches cited in the literature in the field of diagnostic monitoring of rotating machinery [248,282,270]. The EA approach involves filtering the signal around the spectral region of interest. The CDFs themselves are typically swamped within the low frequency broadband noise that is inherently present in mechanical systems. Therefore, it is the higher order harmonics that are more pronounced (spectrally visible), due to strong resonances occurring within the bearing structure, thus yielding a much greater signal to noise ratio at these higher frequencies. A band-pass filter is used to focus on the particular region of interest to remove all other unwanted adjacent frequencies that are deemed noise. This band limited signal then passes through a stage of rectification or de-
modulation to produce an envelope signal. This envelope signal is then analysed in the frequency domain by taking the Short-time Fourier Transform (STFT). In the frequency domain the effect that this enveloping process has on this band limited signal is to shift the spectral components to baseband, i.e. down to 0 Hz (DC). That is, the process effectively finds the spectral repetition present (harmonics), resulting in an envelope spectrum that will have a peak at the CDF ($cdf_0$ fundamental) of the fault and at integer multiples of this $2cdf_0$, $3cdf_0$, … and so on. Different techniques of performing rectification have been documented in the literature [282,118,323]. The most common of which is by taking the absolute value of the discrete-time sequence, labeled here ($\epsilon_{abs}$). A variation on the classic envelope has been proposed by Ho and Randall [118], where rather than taking the absolute value, the square of the signal is taken to yield a squared envelope, labeled here ($\epsilon_{sqr}$). By simulating REB faults, their studies have shown that squared envelope and other higher order power envelopes gave better results when the Signal to Noise Ratio (SNR) is greater than unity. This effectively takes the components ($f_0, 2f_0, 3f_0, ...$), along with any AM sidebands present, and significantly emphasizes them above the noise floor in the baseband spectrum. Another approach to rectification considered here is by using an analytic signal, labeled here ($\epsilon_{an}$). An analytic signal is a complex signal obtained by taking the Hilbert transform of the signal, $x(r)$, to give a $\pm \frac{\pi}{2}$ phase shifted version, $x(i)$. This phase shifted version forms the imaginary component and the original $x(r)$ from the real component; see Eq. (7.1). This results in replacing each pair of positive and negative frequency phasors in the two sided spectrum, containing the real signal $x(r)$, with one phasor of double the amplitude that rotates in the positive direction. The magnitude of the analytic signal gives us the envelope signal, see Eq. (7.2).

$$z[n] = x_r[n] + j x_i[n],$$

where $x_i[n]$ is the Hilbert transform of $x_r[n]$, and

$$\epsilon_{an}[n] = (x_r^2[n] + x_i^2[n])^{\frac{1}{2}},$$

The next stage, after rectification, is to low-pass filter the envelope signal to remove any spurious high frequency mirror harmonics, generated as a result of the rectification process. This filtering stage is followed by taking the spectrum of the envelope signals, which is achieved by taking the STFT. In the case of fault diagnosis, the relative position of the peaks in the envelope spectrum can provide information as to which CDF ($f_0$), or REB part has the fault/defect. The envelope spectrum provides valuable diagnostic information in terms of monitoring how the spectrum evolves over the lifetime of the bearing.

The EA approach provides robustness against the effects of slight random fluctuations between the temporal durations of impulse, $\tau$, as discussed earlier in this section. These random fluctuations cause spectral smearing, which cause spectral components to smear laterally [118,248]. This is highly undesirable from a feature extraction view point as the diagnostic information is difficult to extract due to the spectral smearing. The *enveloping process* of the EA
approach is robust to such spectral smearing and uncovers diagnostic information that otherwise could prove difficult to distinguish from the conventional spectrum alone.

7.3 Data-driven Approach

Conventional spectral features are first extracted from the acoustic data. The acoustic data signal is split into windowed (hamming), non-overlapping time frames, and the STFT is applied, which results in an estimate of the short-term, time localized frequency content of the acoustic signal. This spectral feature extraction approach results in irrelevant, noisy parts of the spectrum being included in the feature vector. The EA approach attempts to isolate the frequency band of interest and attenuate noise outside of this band. However, it is proposed that the salient information for determining the RUL may exist in several different locations across the entire spectrum from 0 to 25kHz. Therefore, an Information Theoretic approach to feature subset selection (FSS) is employed to remove noisy data from the feature vector by only selecting features relevant to the task of predicting the RUL. Specifically, Mutual Information (MI) is used as a measure of usefulness of each spectral component. Therefore the MI criterion forms a basis for FSS of spectral components, MIFSS, in order to optimize the choice of features used as inputs to the classifier to predict the RUL of the machine. This approach does not require any a-priori information regarding the spectral location of potential defects and their associated resonances. It determines the relevant spectral features for monitoring by using information obtained from the data acquired over the lifetime of the machine only. This feature extraction technique is described in more detail in [271]. Fig. 7.3 illustrates which spectral components are selected using the MI-FSS criterion for given subset sizes of 0, 16, 64 and 128 features. The frequency bands used for the EA feature extraction approach, detailed in Section 7.2, are also illustrated for comparison on Fig. 7.3. The figure shows that the most important 16 features tend to exist primarily in the (10-15)kHz and (15-20)kHz frequency bands. Further increasing the number of features, it can be seen that relevant information for predicting RUL is spread out over all the different frequency bands proposed in the previous section. Thus, a subset of the spectral components is selected via MI-FSS and used as input to the pattern recognition algorithms to determine the RUL of the machine.

7.4 Experimental Implementation

Acoustic data was captured using a microphone, in close proximity to a rotating machine, over its entire lifetime. The rotating machine was maintained at a constant load and shaft speed (approx. 80 Hz), in high heat conditions to accelerate failure over a period of approximately 6 months. Ultimate machine failure was due to complete bearing seizure. Acoustic data was acquired at a sampling rate of 50,000 samples/second.

In this study, fixed frequency bands were used for selecting the pass-bands as follows: F1(1-
7.4. Experimental Implementation

2.5), F2(2.5-5), F3(5-10), F4(10-15), F5(15-20), F6(20-25) and F7(15-25) kHz. A similar approach was taken in [197]. Rectification was performed using the three techniques as described earlier in this section ($\epsilon_{abs}$), ($\epsilon_{sqr}$) and ($\epsilon_{en}$). The STFT\(^1\) was computed to obtain the Envelope Spectrum.

The Data-Driven approach to feature extraction involves first extracting 512 log-spectral features via a straight-forward STFT of the acoustic data. The resultant spectral feature vector spans the entire spectrum from 0-25kHz. The MI-FSS approach is then employed to select the top 128 features according to the MI criterion mentioned in Section 7.3. These 128 features were retained in the feature vector for classification.

In order to predict the RUL of the machine, a classification approach is proposed that determines what state of degradation, or wear states, the machine is currently in. In this work, 10 wear states are used to predict the RUL. The degradation of the machine progresses through several stages of physical wear. As the exact location in time where such degradation events occur, is difficult to ascertain, therefore the data is divided into 10 equal segments for labelling.

\(^1\)For both the EA and the DD approach the frame duration was 20msec with a 1024-pt FFT and then averaged over 100 consecutive frames.
Each segment or wear state represents a different time interval over the lifetime of the machine from 1 to 10, where 1 is new and 10 is approaching failure. The Nearest Neighbour (NN) classification algorithm is used to determine the RUL of the machine. In the NN algorithm the training samples are mapped into a multidimensional feature space, which is partitioned into regions based on the class labels. The class is predicted to be the class of the closest training sample, using the Euclidean distance metric\(^2\). The data used in the classification training step is sampled over the lifetime of the machine. For testing, separate data is sampled over the same lifetime. Once the features are extracted for every sample in the training set, the mean and standard deviation are computed for normalization. Each feature dimension in the training set is separately scaled and shifted to have zero mean and unit variance. These normalization parameters are then applied to the test set.

### 7.5 Results and Discussion

Both the EA and Data-Driven feature extraction approaches have been applied to the acoustic data signals. A detailed set of results is presented in Table 7.1, using the EA feature extraction approach and the pass-bands indicated in Section 7.2. The three different rectification techniques: Absolute Value ($\epsilon_{abs}$), Square Value ($\epsilon_{sq}$), and the Analytic Signal ($\epsilon_{an}$) have been applied. The results indicate that F4 and F5 perform considerably better than the other bands. Fig. 7.3 also illustrates that these bands contain the majority of relevant information for predicting RUL. From the results, it is observed that all approaches perform best in the band F5 (15-20)kHz. While the ($\epsilon_{abs}$) and ($\epsilon_{an}$) approaches are comparable, the ($\epsilon_{abs}$) performs marginally better with an accuracy of 93.7%. Table 7.2 compares the results from each of the EA approaches with or without a Low Pass Filter (LPF), for selecting a subset of the spectral components, i.e. the first 128 spectral components. It is shown that using the LPF does not improve the accuracy for the task of predicting the RUL. A summary of the important results, for the Data-Driven approach, to feature extraction is also presented in Table 7.2. Using all 512 features implies that no FSS is performed on the spectral components, whereas using 128 features implies that the MI-FSS criterion was employed to select the top 128 MI features for processing. It is shown that the MI-FSS Data-Driven approach to feature extraction greatly improves accuracy in predicting the RUL of the machine, with an accuracy of 97.7%.

### 7.6 Conclusion

The results presented in this Chapter indicate that there exists sufficient information, in the acoustic signal emitted from a machine to determine the RUL using both the novel implementation of the traditional EA feature extraction approach. This approach provides an accuracy of 93.7% and a Data-Driven approach to feature extraction was also employed providing a sig-

---

\(^2\)This metric was presented in Chapter 4.
Table 7.1: Classification performances accuracies for the different Envelope Analysis implementations.

<table>
<thead>
<tr>
<th>EA method</th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
<th>F4</th>
<th>F5</th>
<th>F6</th>
<th>F7</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\epsilon_{abs}$</td>
<td>32.7</td>
<td>31.1</td>
<td>38.1</td>
<td>84.2</td>
<td>93.7</td>
<td>36.3</td>
<td>93.3</td>
</tr>
<tr>
<td>$\epsilon_{sqr}$</td>
<td>29.5</td>
<td>26.7</td>
<td>35.0</td>
<td>80.8</td>
<td>81.4</td>
<td>31.1</td>
<td>76.8</td>
</tr>
<tr>
<td>$\epsilon_{an}$</td>
<td>31.4</td>
<td>31.1</td>
<td>33.7</td>
<td>79.2</td>
<td>93.2</td>
<td>28.7</td>
<td>92.5</td>
</tr>
</tbody>
</table>

Table 7.2: Classification accuracies comparing Envelope Analysis to Data-driven Approach.

<table>
<thead>
<tr>
<th>No. Features</th>
<th>128</th>
<th>512</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\epsilon_{abs}$ with/without LPF</td>
<td>92.7</td>
<td>93.7</td>
</tr>
<tr>
<td>$\epsilon_{sqr}$ with/without LPF</td>
<td>79.2</td>
<td>81.4</td>
</tr>
<tr>
<td>$\epsilon_{an}$ with/without LPF</td>
<td>92.7</td>
<td>93.2</td>
</tr>
<tr>
<td>Spectral with/without MI-FSS</td>
<td>97.7</td>
<td>94.8</td>
</tr>
</tbody>
</table>

significantly higher accuracy of 97.7%. It is proposed that the success of this approach is due to: (i) the noisy irrelevant features being removed from the feature vector, (ii) the inclusion of relevant information in the feature vector that is spread throughout the spectrum from 0 to 25 kHz, and (iii) the absence of a-priori assumptions in the feature extraction process. While some automated approaches to selecting the cut-off frequencies for the frequency band of interest exist for EA, this approach is still impaired due to the fact that information is spread over different locations across the entire spectrum as illustrated in Fig. 7.3. Therefore, simply widening the band to incorporate all such information will only result in noisy data also being included in the resulting feature vector. This is verified by the results of the wider pass-band of F7 (15-25 kHz) in Table 7.1, which performs poorer than the narrower pass-band of F5 (15-20 kHz). In addition, it is proposed that the Data-Driven approach is a more suitable candidate for use in ACM as it removes the empirical analysis required to determine the frequency range of interest when applying the traditional EA approach to different machines/environments.
Conclusions

The thesis has focused mainly on innovative applications in the area of digital educational technologies. More specifically, it has focused on addressing important scientific research problems that are present when developing an automated production process for the creation of interactive educational resources. This Automatic Content Generation (ACG) system has commercial importance as it enables the streamlined production of these sophisticated rich media based educational applications. This is highly advantageous, as it greatly facilitates the efficient and cost effective production of these resources, for both professional educational products and for User Generated Content (UGC) based learning objects.

Three educational applications have been investigated and developed as part of this research, namely, ReciTell, iMARK\(^1\) and TBDSI\(^2\). As well as addressing research problems surrounding the ACG system, this work has addressed pedagogic problems and aspects in relation to the development of these three applications.

Finding the solutions to these scientific research problems has involved the exploration of advanced discrete-time signal processing algorithms and methodologies. The following section describes some of the areas that have been presented in this thesis, and outlines some areas for future work.

\(^{1}\)iMARK, interactive Media Annotation Resource Kit.
\(^{2}\)TBDSI, The Book: Discovering Sounds Initiative.
8.1 Automatic Speech Segmentation

While it is relatively straightforward to define the speech segmentation problem, the complex nature of the speech signal itself makes the problem particularly difficult to solve. In relation to this, Chapter 2 has provided some background theory on acoustic signals, particularly focusing on speech signals. Different aspects in relation to the theory of speech production have been presented. The physiological areas are reported, such as the organs that are responsible for producing the various speech sounds. The different aspects of speech sounds are looked at in terms of *place* and *manner* of articulation, as well as *voicing*. An overview of the prosodic effects which contribute to the linguistic structure in a language have been presented. These prosodic areas include, *stress*, *intonation* and *co-articulation*. Some discussion has been presented for modelling speech production, in terms of linear planar models, as well as more complex models for aeroacoustic flow in the vocal tract. The last topic presented in Chapter 2, was speech synthesis. Synthesized speech signals have been discussed as they play an important role within the speech segmentation algorithm.

Chapter 3 has given a review of the motivation and methods on the topic of speech segmentation. The literature review has been presented under the areas of (i) thematic applications, (ii) segmentation methodologies and feature extraction techniques, and (iii) the evaluation measures that are used to assess segmentation performance.

Chapter 4 has explored the different methodologies for extracting pertinent feature representations for characterizing speech signals. The Chapter began by discussing the philosophy of speech feature extraction in order to give an understanding and appreciation for *why* features are needed to characterize the speech signal. The Chapter then moved on to investigating many different feature extraction methods, with both univariate features and multivariate features being considered. Some examples of these include: univariate features of: Energy, ZCR, Spectral Centroid, Spectral Roll-off and Spectral Flux, and multivariate features of: PLP and MFCC. In addition to investigating these state of the art features, this Chapter has proposed other features for the task of speech segmentation, such as STFT spectrum, Teager Energy Spectrum, OBA and $\frac{1}{3}$OBA. Dynamic feature extensions such as first order derivative (delta or $\Delta$), second order derivative (delta-delta or $\Delta\Delta$) and Relative Spectra: termed RASTA Processing, have been explored. Chapter 4 has also presented the methodology behind Principal Component Analysis (PCA) for the task of feature transformation and dimensionality reduction.

Chapter 5 has reported on the novel segmentation methodology that has been developed for application within the automatic content generation (ACG) system for the creation of rich media based educational applications and resources. A methodology has been proposed which uses the Dynamic Programming (DP) technique known as Dynamic Time Warping. The algorithm has been described in detail, and has incorporated the speech features and signal analysis techniques from Chapter 4. Chapter 5 has also provided a discussion on the topic of segmentation tolerances, taking into consideration the segment points (in time) role within the end-application in
the area of synchronized content for educational technologies. A detailed analysis of the results has been presented which shows the effectiveness of the proposed segmentation system using different approaches. Superior performance has been reported for the PLP Set within the Dynamic Multivariate Feature Approach. The PLP Set has achieved the best performance overall with segmentation accuracies\(^3\) of \(\hat{S}_{acc}(97.30, 99.57, 99.95)\%\) and segmentation error statistics of \((\bar{x} = 0.003 \& S = 0.031)\) sec. Furthermore, a similar performance has been reported with the incorporation of Principal Component Analysis, referred to as PCA Dynamic Multivariate Feature Approach. Using a transformed and dimensionally reduced set of PLP features (PLP Set Con), segmentation accuracies of \(\hat{S}_{acc}(97.25, 99.54, 99.95)\%\) were achieved, with a segmentation error of \((\bar{x} = 0.004 \& S = 0.032)\) sec. These two sets of results strongly show the effectiveness and the validity of the proposed automatic speech segmentation methodology. Moreover, they give great confidence towards the capability of the method to fulfill its important role within the ACG system.

Chapter 6 has covered the area of integration and implementation of the end-applications. This Chapter began by discussing some important background theory relating to the technology. Areas that have been reported upon include: (i) eye movement characteristics, (ii) appropriate reading rates, (iii) some cognitive aspects related to cross-modal synchronism in multimedia resources, and (iv) a brief overview of hypotheses of learning styles. The three applications that were presented, where the underpinning technology has been applied, include:

1. **ReciTell**: for professional innovative literacy and language learning products. The ReciTell VLE incorporates a complementary fusion of visual, auditory, and kinaesthetic learning modalities, to provide an enhanced learning experience for the pupil.

2. **iMARK**: for user generated content, to greatly facilitate the creation of interactive learning objects. This also has important applications in literacy and language learning.

3. **TBDSI**: for studies of manuscripts, to allow greater accessibility to these precious resources. More specifically, the focus has been on medieval manuscripts which aims to help increase the appreciation, understanding and transcription of these manuscripts, for which the texts are in older languages, such as medieval (middle) English. This has been achieved by developing an application to display these valuable manuscripts and EPBs in digital format, accompanied with additional synchronous audio visual media. This creates a more informative, engaging and enriching learning experience for the student.

### 8.2 Automated Condition Monitoring of Rotating Machines

A separate scientific research problem that has been investigated as part of this research is in the area of automatic condition monitoring of machinery. This is an industrial application that

\(^3\)The segmentation accuracy values given, are for the tolerances of 70,90,110 (in msec).
pertains to being able to automatically estimate the Remaining Useful Life (RUL) of rotating machinery using acoustic signal processing methods. Automatic prediction of the remaining useful life is valuable as it allows for improved reliability of machinery, scheduling of maintenance prior to machine failure, in order to prevent machine downtime, and the removal of the cost of unscheduled maintenance.

Chapter 7 has reported on a signal processing method that has been explored in the area of Automated Condition Monitoring (ACM) of rotating machines. The approach presented has used acoustic noise signals (< 25kHz) to estimate the RUL of a rotating machine. This involved investigating a novel implementation of an ubiquitous ACM analysis technique called Envelope Analysis (EA). An alternative Data-Driven approach has also been presented for comparison purposes. The results have found that there exists sufficient information present in the acoustic signal emitted from a machine to determine the RUL. The EA approach gave a classification accuracy of 93.7%, and the Data-Driven approach has increased on this, increasing the accuracy to 97.7%. Discussion has been provided for each of the methods in terms of their effectiveness and some associated limitations.

8.3 Future Work

There are many avenues open to further exploration for the research and technology that have been presented in this thesis. Two ideas to extend on this research are: (i) Oral Reading Interaction and Assessment and (ii) MediaSkim: Speed Reading Technology, these are presented below.

8.3.1 Oral Reading Interaction and Assessment

One area that could be investigated, with a view to increasing further the pedagogic benefits and learning outcomes from the ReciTell Virtual Learning Environment (VLE) involves incorporating the functionality to be able to record and assess the pupils reading ability. The current system provides for learning modalities of auditory, visual and kinaesthetic modes. The facility to support oral reading interaction for the pupil would be a favourable inclusion. This would require the technology to record the spoken audio for children’s reading from the text (storybook). This additional functionality would have great benefits in terms of the pupil’s learning outcomes and interaction within the VLE, as well as benefiting a modern approach to facilitate assessment of reading abilities.

This could be achieved as follows: the pupil is prompted on screen to read an extract from the text; this is recorded using a headset microphone and stored in memory linked with its associated text extract. There are different envisaged schemes to provide for different levels of proficiency. One such scheme is presented below.

Using a headset (audio and microphone) the pupil:
8.3. Future Work

1. Reads the text along in unison while listening to the spoken audio narration.

2. Hears the audio first, and then afterwards is prompted to read the text themselves.

3. Is prompted to read it on their own without any audio narration.

Once the pupil has finished all of the reading tasks, the teacher can then assess their reading ability through listening to their audio recording with its associated text extracts. This can be achieved through an interactive assessment tool, enabling the teacher to select input boxes where the pupil had difficulties reading, i.e. troublesome words etc. This data can then be stored in the memory of the VLE for each pupil, and can quickly highlight problematic areas of reading and can also be used to monitor progression. Such a system would be an extremely useful tool and resource for the teachers and for remedial support in schools.

Furthermore, a more sophisticated system is envisaged that could automatically assess and monitor reading ability. This would involve developing an advanced signal processing algorithm to automatically measure reading ability from the speech recording. It is thought that some of the front-end stages of proposed automatic speech segmentation methodology of this thesis could be used to enable automatic reading assessment. As one important similarity exists between the two problems of segmentation and reading assessment, in that they are both text-dependent, such that a reference signal can be used in each case. This reference signal can be a synthetic speech signal or a professional natural voice recording of the text. The reference signal comprises of a correct reading of the passage and can be time aligned with the children’s audio recording. While the segmentation algorithm performs segmentation at particular regions of interest, instead this method can derive distance/similarity measures from the two time aligned signals to assess reading. The concept would be that large recorded distances are a good indication of incorrect reading. On this basis appropriate distance/similarity metrics could be explored. Such an automated assessment approach would be an extremely valuable tool and resource, saving significant correction and assessment time for the teacher.

Perhaps a good starting point for commencing this research would involve taking sample speech utterances from the children’s speech corpus, CMU Kids Corpus [82] and conducting preliminary experiments.

8.3.2 MediaSkim: Speed Reading Technology

The notion of interactively skimming speech and audio recordings was proposed by Arons, [16,15]. This concept was based on the idea of being able to quickly browse audio documents in a quick, easy and efficient manner. The concept that is proposed herein, is called MediaSkim, which takes the idea of audio skimming and develops it further to incorporate audio and visual media components.

This concept is based around skim reading, see Fig. 8.1, and involves developing an interac-

\[^4\text{Skim Reading: refers to reading a document quickly or superficially [239].}\]
Conclusions

Reading Example:
This person is reading the text for understanding.
So even though not every word is fixated, the amount of time spent on each word is indicative of the processing of the word.

(a)

Skimming Example:
This person is skimming the text. This is most obvious from the pattern of fixations that are more dispersed and shorter fixation durations that is typical for this type of reading. The main gist maybe understood, but poorer memory for the text usually results.

(b)

Figure 8.1: The images show eye movements, where dots represent fixations and lines represent saccades. The image on the left shows the eye movement characteristics for reading the text for comprehension, whereas the image on the right shows the eye movement characteristics for skim reading. Credit: images by Rayner and Castelhano [252].

tive tool that guides the user towards the selected skim-content. This would greatly assist them in skim reading a document quickly and efficiently, allowing them to pause at any point should they wish to read or listen to a section more thoroughly. The technology would direct the user towards the selected skim content using visual word highlighted with synchronized speech.

There are different areas that could be investigated on the topic of selecting the appropriate words from the document to form the skim (subset) content. Two naïve approaches are: (i) randomly select words from the passage in a sequential order, from start to finish, and (ii) an information theoretic approach, extracting words from the text where the word level entropy is higher based on some entropy threshold or percentage to form a basis which defines the appropriate skim-content to select. These approaches for selecting the appropriate skim content, although quite straightforward to implement, would perhaps be limited in their effectiveness. Nonetheless, they could provide good statistical comparisons for other more sophisticated clever approaches. The current thinking leads to the suggestion of modelling the higher order semantics\(^5\) of the text document to deduce the key-words from the passage in order to efficiently assess meaning from the document. Accordingly, further studies are needed to validate this concept and its potential role and merits within possible end-applications. Three envisaged applications of where the MediaSkim concept could be valuable are: (i) search engine optimization, (ii) academic literature reviews, and (iii) prior art searches for patents attorneys and inventors. The ReciTell VLE system in its current format presents itself as an ideal vehicle to carry out these studies and experiments.

\(^5\)Semantics referring to linguistics and logic concerned with meaning, i.e. the meaning of a word, phrase, sentence, or text [239].
The properties shown in Eqs. (9.1) and (9.2) are the additive and the homogeneity properties respectively.

\[ T[x_1(n) + x_2(n)] = T[x_1(n)] + T[x_2(n)] = y_1(n) + y_2(n) \]  
\[ T[a x(n)] = a T[x(n)] = ay(n) \]  
These can be combined into the principle of superposition as shown in Eq. (9.3).

\[ T[a x_1(n) + b x_2(n)] = a T[x_1(n)] + b T[x_2(n)] \]
Figure 9.1: The figure shows the non-linear perception of sound over the audible frequency range, this data is sourced from the International Standards Organization [217]. These are referred to as Fletcher-Munson curves after original work by Fletcher and Munson [86].
Figure 9.2: The transfer function, $E(\omega)$, for the equal-loudness pre-emphasis stage.

<table>
<thead>
<tr>
<th>General terms</th>
<th>Modified terms [sic]:</th>
</tr>
</thead>
<tbody>
<tr>
<td>frequency</td>
<td>quefrency</td>
</tr>
<tr>
<td>spectrum</td>
<td>cepstrum</td>
</tr>
<tr>
<td>phase</td>
<td>saphe</td>
</tr>
<tr>
<td>amplitude</td>
<td>gamnitude</td>
</tr>
<tr>
<td>filtering</td>
<td>liftering</td>
</tr>
<tr>
<td>harmonic</td>
<td>rahmonic</td>
</tr>
<tr>
<td>period</td>
<td>repiod</td>
</tr>
<tr>
<td>analysis</td>
<td>alanysis</td>
</tr>
</tbody>
</table>

Table 9.1: Glossary of terminology, coined by Bogert, Healy and Tukey in 1963 [216,39], some of which are now in prominent use amongst the signal processing community, terms such as [sic]: cepstrum, quefrency and liftering, whereas the remainder have essentially faded into the background, they are included here interests of completeness.
### CONSONANTS (PULMONIC)

<table>
<thead>
<tr>
<th>PALATAL</th>
<th>DENTAL</th>
<th>ALVEOLAR</th>
<th>PALATO-ALVEOLAR</th>
<th>RETROFLEX</th>
<th>PALATAL</th>
<th>VELAR</th>
<th>UVULAR</th>
<th>PHARYNGEAL</th>
<th>GLOTTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nasal</td>
<td>m</td>
<td>m</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>N</td>
<td>Glottal</td>
</tr>
<tr>
<td>Plosive</td>
<td>p</td>
<td>b</td>
<td>t</td>
<td>d</td>
<td>c</td>
<td>k</td>
<td>q</td>
<td>g</td>
<td>Glottal</td>
</tr>
<tr>
<td>Fricative</td>
<td>φ</td>
<td>f</td>
<td>θ</td>
<td>s</td>
<td>z</td>
<td>j</td>
<td>x,y,x</td>
<td>k</td>
<td>Glottal</td>
</tr>
<tr>
<td>Approximant</td>
<td>l</td>
<td>j</td>
<td>⁹</td>
<td>⁹</td>
<td>j</td>
<td>⁹</td>
<td>⁹</td>
<td>⁹</td>
<td>⁹</td>
</tr>
<tr>
<td>Trill</td>
<td>B</td>
<td>r</td>
<td>⁹</td>
<td>⁹</td>
<td>⁹</td>
<td>⁹</td>
<td>⁹</td>
<td>⁹</td>
<td>⁹</td>
</tr>
<tr>
<td>Tap, Flap</td>
<td>V</td>
<td>r</td>
<td>³</td>
<td>³</td>
<td>³</td>
<td>³</td>
<td>³</td>
<td>³</td>
<td>³</td>
</tr>
<tr>
<td>Lateral Fricative</td>
<td>⁹</td>
<td>k</td>
<td>⁹</td>
<td>⁹</td>
<td>⁹</td>
<td>⁹</td>
<td>⁹</td>
<td>⁹</td>
<td>⁹</td>
</tr>
<tr>
<td>Lateral Approximant</td>
<td>⁹</td>
<td>j</td>
<td>⁹</td>
<td>⁹</td>
<td>⁹</td>
<td>⁹</td>
<td>⁹</td>
<td>⁹</td>
<td>⁹</td>
</tr>
<tr>
<td>Lateral Flap</td>
<td>⁹</td>
<td>j</td>
<td>⁹</td>
<td>⁹</td>
<td>⁹</td>
<td>⁹</td>
<td>⁹</td>
<td>⁹</td>
<td>⁹</td>
</tr>
</tbody>
</table>

Where symbols appear in pairs, the one to the right represents a modally voiced consonant, except for murmured ə.

Shaded areas denote articulations judged to be impossible.

### CONSONANTS (NON-PULMONIC)

- Bilabial fricated
- Labio-dental
- Dental or alveolar
- Palatal or alveolar
- Retropalatal or alveolar
- Labiovelar
- Alveolar fricative

### CONSONANTS (CO-ARTICULATED)

- Voiceless labialized velar approximant
- Voiced labialized velar approximant
- Voiced labialized palatal approximant
- Voiceless palatalizedpostalveolar (alveolo-palatal) fricative
- Voiced palatalized postalveolar (alveolo-palatal) fricative
- Simultaneous x and y (disputed)

### VOWELS

- Front
- Near front
- Central
- Near back
- Back

- Close
- Close mid
- Open mid
- Near open
- Open

Vowels at right & left of bullets are rounded, unrounded.

### SUPRASEGMENTALS

- Primary stress
- Extra stress
- Level tones
- Contour-tone examples:

### TONE

- Level tones
- Top
- High rising
- Long
- Half-long
- High
- Falling
- Short
- Extra-short
- Mid
- Low rising
- Standard break
- Linking
- High falling
- Intonation
- Minor (foot) break
- Tone-terminating
- Major (intonation) break
- Upstep
- Peaking
- Global rise
- Global fall

### DIACRITICS

Diacritics may be placed above a symbol with a descender, as ə. Other insymbols may appear as diacritics to represent phonetic detail: 'r' (fricative release), 'e' (breathy voice), 'a' (glottal onset), 'epenthetic schwa', 'a' (epenthetic devoicing).

### SYLLABILITY & RELEASES

- Syllabic
- Non-syllabic
- Nasal release
- Lateral release
- No audible release

### PHONATION

- Voiceless or breathy voice
- Nasal voice
- Strident
- Centralized
- Lowered

### PRIMARY ARTICULATION

- Dental
- Apical
- Laminal
- Stress
- Centralized

### SECONDARY ARTICULATION

- Labialized
- Palatalized
- Velarized
- Retracted
- Advanced
- Centralized
- Retracted tongue root

Figure 9.3: The most recent version of the International Phonetic Alphabet (IPA) 2005 [124]
Figure 9.4: Examples of some window functions in the frequency domain.
Bibliography


171


