# Psychoacoustics Modelling and the Recognition of Silence in Recorded Speech

Thesis by

Derek Wilson

Submitted In Partial Fulfilment of the Requirements
for the Degree of
*Doctor Of Philosophy*

School of Computing Science
Newcastle University, Newcastle Upon Tyne, UK.

September 2018

# Abstract

Over many years, a variety of different computer models purposed to encapsulate the essential differences between silence and speech have been investigated; but that notwithstanding, research into a different audio model may provide fresh insight. So, inspired by the unsurpassed human capability to differentiate between silence and speech under virtually any conditions, a dynamic psychoacoustics model, with a temporal resolution of an order of magnitude greater than that of the typical Mel Frequency Cepstral Coefficients model, and which implemented simultaneous masking around the most powerful harmonic in each of 24 Bark frequency bands, was evaluated within a two stage binary speech/silence non-linear classification system. The first classification stage (deterministic) was purposed to provide training data for the second stage (heuristic) — which was implemented using a Deep Neural Network (DNN).

It is authoritatively asserted in the Literature — in a context of speech processing and DNNs — that performance improvements experienced with a 'standard' speech corpus do not always generalise. Accordingly, six new test-cases were recorded; and as this corpus implicitly included frequency normalisation it was feasible to assess whether the solution generalised, and it was found that all of the test-cases could be successfully processed by any of the six trained DNNs. In other tests, the performance of the two stage silence/speech classifier was found to exceed that of the silence/speech classifiers discussed in the Literature Review; but it was interesting to note that the Split Sample Technique for neural net training did not always identify the optimal trained network — and to correct this, an additional step in the training process was devised and tested.

Overall, the results conclusively demonstrate that the combination of the dynamic psychoacoustics model with the two stage binary speech/silence non-linear classification system provides a viable alternative to existing methods of detecting silence in speech.

The author gratefully acknowledges the contributions to the success of this work by those who — often selflessly — helped him to broaden his horizons; those who were encouraging and actively supportive of this enterprise at his last place of work; those from other institutions who helped him to achieve a perspective on, and to develop a context for this work; and ...

Dr. Paul Ezhilchelvan, Dr. Leo Freitas and Dr. Phillip Lord who provided timely advice during the annual progression process.

Dr. Lindsay Marshall and Dr. William Blewitt — supervisor and associated supervisor respectively — who both exemplify all that is good about their profession.

Newcastle
University

This author of this work is not affiliated with, nor is the work sponsored, authorised or approved by, any of the equipment, software, or service providers, named within this thesis.

The rights in all of the registered or copyrighted material or trademarks reproduced within this thesis — whether such material is explicitly identified or not — remain with the respective owners.

# Contents

# List of Figures

# List of Tables

# Chapter 1: Introduction

## 1.1 Motivation

The realisation of an effective automated process for the Recognition of Silence in Recorded Speech is difficult because a great deal of recorded speech — and the silence therein — is noisy; and algorithms for the analysis of clean speech do not function well with noisy speech (Smaragdis [2013]).

In latter years, the problem of recognising silence in speech has largely been ignored by the research community, because a process which implicitly identifies silence exists in the form of Automatic Speech Recognition (ASR) Forced-Alignment (Huang et al. [2001], Viterbi [2006]) — which is the process of aligning speech with its phonetic transcript. But forced-alignment does not provide the ideal solution: firstly because a great deal rests upon the quality of the implementation and the accuracy and availability of the speech models, and secondly because it has been demonstrated (Brognaux and Drugman [2016]) that prior knowledge of silence pauses can be used to improve forced-alignment. This, together with the results of a search of the Literature, give credence to the belief that the problem of detecting silence in speech has not been solved. So in this work alternative methods for recognising silence in recorded speech are explored; specifically methods with no dependency upon complex speech processing techniques, and with no requirement for a specialised knowledge base.

From the Literature (Atal and Rabiner [1976], Deekshitha et al. [2015]) it is clear that the most difficult aspect of recognising silence in speech is discriminating between unvoiced speech — those speech sounds such as fricatives and sibilants, where the vocal chords are inactive — and the noise during silence. This is an interesting problem in itself, because a solution may pave the way for an alternative method for recognizing other speech features.

## 1.2 The Detection of Silence in Recorded Speech

Silence in recorded speech comprises noise of various types, including periodic and aperiodic noise, transient bursts, electrical interference, white noise (for example from the air impeller of a cooling fan), breathing sounds, involuntary vocal effects such as non-silent releases (for example, from the relaxation of the vocal tract at the end of a block of continuous speech), the rustling of paper, the sounds of machines — the list goes on. The effect of this 'noise' is that it is difficult to automatically separate speech from silence. This problem is illustrated in Figure 1.1, where a short spike and an inhalation fricative are evident during the silence. This noise may trigger a speech/silence discriminator that uses a simple threshold and be erroneously marked as speech.

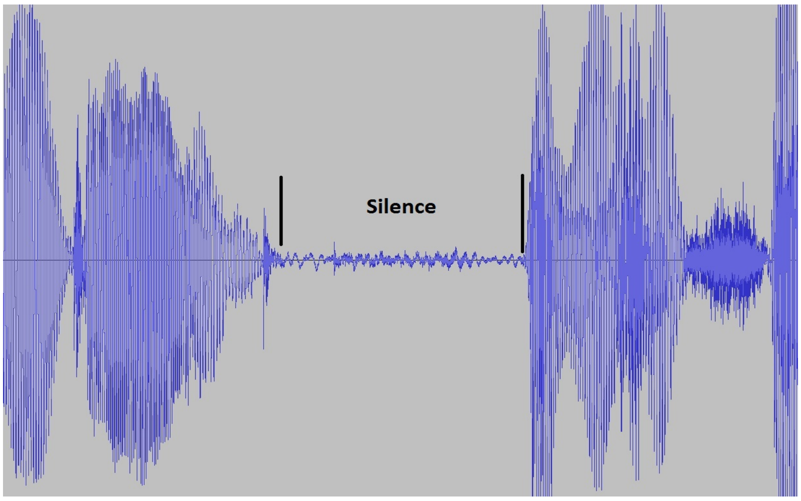


Figure 1.1: The Noise in Silence

The separation of silence from speech may be a necessary, or a useful first step in several automatic processes, including:

- Automatic Speech Recognition (ASR) and Transcription — for the automatic insertion of punctuation.
- To improve the 'silence' model for the Forced-Alignment of speech with its phonetic transcript.
- Indexing Speech Recordings
- Lossless data compression : Silence removal for compact storage.
- Speech analysis.

Silence and speech are not linearly separable on the basis of energy alone; because the zero sound level — which it may be thought corresponds to silence — seldom obtains, because of noise. So some means of characterising speech is necessary, such that audio which does not have the characteristics of speech can be classified as silence. Speech has the Markov Property (pg307: Oxford Dictionary of Computing [2008]), which is where the next state is often a known stochastic function of the current state; whereas the noise during silence does not conform to any rule. Automatic Speech Recognition (see Section 2.2), which often makes use of the Markov Property in the form of Hidden Markov Models (HMM), may be thought to be the obvious candidate for characterising speech; but it is probable that modelling speech as we perceive it — psychoacoustics processing — is the better choice, because we have a far greater capability for recognising both speech, and silence pauses in speech, than the best ASR system.

## 1.3 Psychoacoustics

Psychoacoustics (Chapter 15, Gold et al. [2011]) is the science of how we perceive sound. Empirical work by Fletcher and Munson [1933], Stevens et al. [1937], Stevens and Volkmann [1940] and Zwicker et al. [1957], among others, showed that our response to auditory stimulus is largely logarithmic. That is, we perceive logarithmic increases in sound magnitude as linear increases in volume

(Fletcher and Munson [1933]), and for frequencies higher than around 500 Hz to 1 kHz, we perceive logarithmic changes in the frequency of a tone as linear changes in pitch (Stevens et al. [1937], Stevens and Volkmann [1940], Zwicker et al. [1957]).

Empirical work by Egan and Hake [1950] and later by Zwicker et al. [1957] (building on work by Fletcher and Munson [1937] and Fletcher [1940]) demonstrated the effect of Masking(Jeffress [1970]). Simultaneous Masking is the psychoacoustics effect where frequencies in the audible spectrum are not perceived because of the response of the human auditory system to nearby stronger frequencies (Egan and Hake [1950]). Forward temporal masking is where a softer sound may not be perceived if it follows a louder sound of a similar frequency; and backward temporal masking is where a softer sound may not be perceived if it is followed by a louder sound of similar frequency (Pg 92, Johnson [2012]). It was shown by Zwicker et al. [1957] that beyond a certain bandwidth around the masking tone (*the Critical Band*), the effect of masking ceases; and that critical bandwidth increases with increase of frequency. Moore and Glasberg [1983] refer to a useful conceptual description of critical bandwidth by Scharf [1970] who described the critical band as, *"that bandwidth at which subjective responses rather abruptly change"*. In 1961 Zwicker proposed the *Bark* auditory frequency scale; a scale that spans the audio frequency spectrum with 24 critical bands. (According to Zwicker the name "Bark" was chosen in memory of Barkhausen, who may have defined the unit for the sound level.) Earlier an alternative auditory frequency scale, the *Mel* scale was empirically derived by Stevens and Volkmann [1940], who set the datum for the Mel scale to 1 kHz, and assigned this a value of 1000 Mel. Although the units for the Mel and Bark scales are very different, the characteristic shape of the scales are similar (see Figures 3.1 and 3.2 in Section 3 herein).

In perceiving speech, we naturally achieve the separation of silence and speech, and my aim is to evolve an automated system that mimics human perception — a psychoacoustics model of sound — in its capability to reject (or ignore) the noise during silence for the recognition of silence in speech.

### 1.4 Hypotheses

Because of the random nature of noise, the parts of a recording that are most consistent are the speech parts; and to recognise silence in a recording, it is also important to recognise certain attributes of speech — but only to the extent necessary for the essential differences between the silence pauses and the speech to be captured. This will not necessarily involve the complexity of Automatic Speech Recognition (ASR), because the ultimate aim is only to have knowledge of the temporal location of the speech, not to decode it. One 'model' for capturing the differences between speech and silence is the psychoacoustics model of speech perception (Fastl and Zwicker [2007]) – a model based upon the modalities of human hearing. The human capability for identifying pauses in speech is unsurpassed; and though the human capability for speech perception is not understood, we have practical knowledge of these modalities. For example, both Krasner [1980] and Schroeder et al. [1979] employed psychoacoustics techniques to reduce the audible noise introduced during speech encoding, and psychoacoustics processing is specified for the MP3 lossy speech compression format (Brandenburg [1999], MP3-Standard [1995]).

**First hypothesis : An audio model based upon an interpretation of the psychoacoustics model of hearing will include sufficient information to facilitate the recognition of silence in speech.**

The consequence of using a model to represent both the speech and the silence is that the discrimination between speech and silence becomes a pattern recognition problem (Atal and Rabiner [1976]). Accordingly to test the first hypothesis it will be necessary to embed the speech model in an evaluation system which includes a supervised pattern recognition process. So, an automated process for generating the training data will also be required for the evaluation system; and preferably this method will be deterministic rather than statistical.

**Second hypothesis: A deterministic speech/silence binary classifier can**

**provide enough information to facilitate the generation of accurate speech and silence training data, such that the errors in classification by the deterministic classifier can be eliminated by a subsequent supervised classification process which uses a psychoacoustics audio model.**

The system to test these hypotheses comprises three main components: a deterministic speech/silence classifier, a speech model based upon the psychoacoustic modalities of speech perception, and a pattern classifier. An overview of the complete system follows.

## 1.5 A Two Stage Binary Speech/Silence Classifier

Figure 1.2 is an overview of the test system. The key components are, a deterministic speech/silence classifier (the D$_{eterm}$Classifier), a speech model based upon the psychoacoustic modalities of speech perception (the LogFB$_{dynamic}$), and a pattern classifier (the DNN_Classifier).

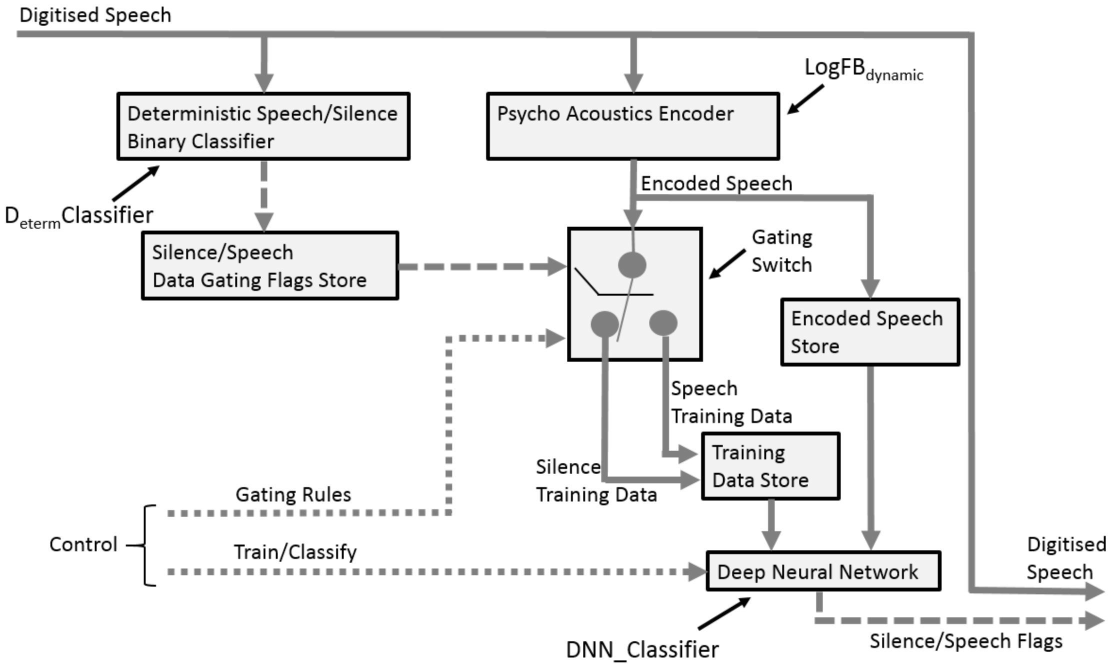


Figure 1.2: Two Stage Binary Speech/Silence Classifier

**Operation**

- Firstly, the D$_{eterm}$Classifier (Section 4.5) makes the Initial Speech/Silence Binary Classification and then stores the locations of all of the detected silence pauses in the Silence/Speech Data Gating Flags Store.
- Secondly:
  - the LogFB$_{dynamic}$ (Section 4.6) converts the complete audio into model slices of $1ms$ duration.
  - the Gating Switch uses both the Data Gating Flags that were generated by the D$_{eterm}$Classifier and the Gating Rules (Section 4.7), to classify some of the model slices as silence training data or speech training data. This data is then stored in the Training Data Store for use by the DNN_Classifier.
- Thirdly, the DNN_Classifier (Section 4.8) is then trained from the Training Data Store, in as many epochs as necessary.
- Fourthly, the now trained DNN_Classifier makes the Final Binary Classification of the data in the Encoded Speech Store; and a set of Speech/Silence Flags is generated.
- Finally, all of the temporal locations in the digitised speech that are flagged as silence, are set to zero.

## 1.6 Limitations

### 1.6.1 English Language Dependency

Although this work is predicated on the belief that all speech and silence is separable using a non-linear heuristic classification process that is trained from the results of a linear deterministic preliminary classification process, it has only been tested with the English Language. In fact a dependency exists within the processes of the $\mathrm{D}_{eterm}$Classifier, that the format of the text of the recording must comply with the punctuation rules prescribed for the English Language. The dependency is simply that the approximate number of silence pauses within each block of the text of the recording is derived from the punctuation therein. This number is then used as the notional 'set-point' for the control loop which optimises the silence/speech threshold for the $\mathrm{D}_{eterm}$Classifier. As discussed in Chapter 5, this method of establishing the set-point is not entirely satisfactory, and it is postulated that a more accurate set point may result, by empirically establishing a mean pause rate as a function of word rate. By removing the dependency on English Language punctuation in such a way, the change would also remove the need for any knowledge of the text of the recording; and the performance of the classifier could then be evaluated for any language.

### 1.6.2 Statistical Significance Versus Inference by Induction

There is no suggestion that the results herein are of statistical significance — and this for two reasons. Firstly, it would be necessary to process a vast amount of different voice recordings to achieve any measure of statistical significance; and secondly the detailed analysis required is a labour intensive process, such that the manual accumulation of sufficient 'evidence' is not feasible. That said, inferences, by induction, may be drawn from this work. Here the specific meaning of induction is that defined in the Oxford Dictionary of Computing [2008] to be where, "A general but not necessarily true conclusion is drawn from a set of particular instances".

## 1.7 Contribution

The filter-bank that is used within the $\mathrm{LogFB}_{dynamic}$ psychoacoustics speech model differs from the conventional form of the filter-bank (Figure 3.5) in three respects. Firstly, the most powerful harmonic in each of the Bark bands is selected, secondly a form of simultaneous masking is imposed around those harmonics, and thirdly the filters do not overlap.

The LogFB$_{dynamic}$ psychoacoustics speech model itself, differs from the Mel Frequency Cepstral Coefficients model of speech in yet another respect, and that is, the model slice duration is decoupled from the Discrete Fourier Transform window duration, such that any duration for the model slice is possible; thereby facilitating — in this case — a speech model with a temporal resolution that is at least an order of magnitude greater than that of the typical MFCC model.

The test results support the hypotheses, that:

- the LogFB$_{dynamic}$ psychoacoustics speech model is effective in capturing the essential differences between speech and silence.
- a deterministic classifier can provide enough information to facilitate the generation of accurate speech and silence training data.

An interesting result is that higher levels of noise during 'silence' directly degrade the performance of the D$_{eterm}$Classifier whereas for the DNN_Classifier there was no indication of any correlation between the noise levels and the absolute performance of the classifier.

The outcome of the split sample DNN training was a set of trained DNN Classifiers, which the testing confirmed to be valid for the subclass of speech and silence as defined in the DNN Training and Validation data sets. A peculiarity of the test corpus — where all of the test-cases are spoken by the same individual — is that the test-corpus implicitly emulates frequency normalisation. So it was feasible to assess the extent to which the solution generalised, and it was found that any of the six trained DNNs operated satisfactorily with any of the test-cases. This suggests the possibility, that with addition of explicit frequency normalisation — some variation of the Vocoder perhaps (i.e. a speech 'analysis-synthesis' system Gold et al. [2011]) — the solution might generalise to the class of speech and silence.

Though all of the trained DNNs obtained using the split sample technique were fit for purpose, the technique did not always identify the optimal trained DNN. That is, for the backpropagation training method used, the only significant independent variables in the system were the random initialisations of the weights for the DNNs, and it was found that a measure of quality of the trained DNN performance varied as a function of this initialisation. Specifically, the dispersion of the distribution for multiple training instances with the same training, validation and test data sets (but different weight initialisations) varied significantly from test-case to test-case — this variation possibly a function of the quality (accuracy) of the training data. So, to obtain the optimal trained DNN when using the split sample technique, an additional training step of establishing the distribution of variation in classification performance according to some appropriate quality measure, as a function of DNN weight initialisation, and then selecting the trained DNN from the centre of the distribution (where the density of trained solutions is at its greatest), is recommended.

A similar experiment to that described in the previous paragraph, but purposed to establish the variation in a measure of the quality of performance of the trained network as a function of the depth of the network, showed that the performance of a network with 6 hidden layers to be only marginally better than the performance of a network with 1 hidden layer.

The most difficult voiced/unvoiced/silence classification issue identified in the Literature is between unvoiced speech and silence; and this manifested itself during testing particularly in respect of discriminating between low energy unvoiced fricatives and inhalation fricatives during silence. The LogFB$_{dynamic}$ model may provide a platform, if enhanced as described in Section 6.3, for investigating this issue further.

The technique of using two classification stages (Qi et al. [2004]) was extensively reworked, and it may be that it can equally be applied to other features of speech — sibilants for example — that can be initially identified, using a deterministic (rules based) method.

## 1.8 Document Outline

**Chapter 2:** The "Technical Background", provides an outline of the broad subject areas that are referenced in this work. The descriptions provided only convey the essence of the subjects — to do full justice to each of the areas would require volumes of material. That said, the information provided is intended to provide a coherent context for the material in the Literature Review.

**Chapter 3:** The "Literature Review", is broadly in three parts. The first is concerned primarily with techniques for identifying silence in speech such as voiced-unvoiced-silence detection, voice activity detection, and speech segmentation, and also touches on pattern matching. The second part is concerned with the modalities of human hearing from two viewpoints: the psychoacoustic and the neurological; and the third part draws the narrative of the first two parts together thereby providing the rationale for the remainder of this work.

**Chapter 4:** The "Research Method", provides a description of and the rationale for the choice of test material, the algorithm for the $\mathrm{D}_{eterm}$Classifier, the detail of the $\mathrm{LogFB}_{dynamic}$ speech model, and the detail of the ANN speech/silence binary classifier (which uses a third party Artificial Neural Network).

**Chapter 5:** The "Results and Analysis", describes the performance assessment criteria and provides the results of the tests for both the $\mathrm{D}_{eterm}$Classifier, and the DNN_Classifier. The performance of the two classifiers are recorded as:

- Silence Deletions, where a silence pause that is perceived by the listener goes undetected by the automated speech/silence classification process.

- Silence Insertions, where silence is detected by the automated speech/silence classifier that is not perceived by the listener.

- Silence start and endpoints.

- Classification errors.

Additionally, Chapter 5 provides the results of tests intended to establish firstly, the degree of consistency of the Artificial Neural Network training process; and secondly, to determine the extent to which the solution generalises.

Chapter 5 also provides a consideration of the difficulties in pinpointing the temporal location of perceived silence pauses, and on the validity of estimating the number of silence pauses from the punctuation in the text, plus a discussion of the test corpus. The classification performance of the $\mathrm{D}_{eterm}$Classifier and the DNN_Classifier are separately considered, as are the short-comings of both classifiers. A consideration of the results of an experiment on speech recorded using the MP3 lossy compression format (detailed in Appendix C) is also provided; and Chapter 5 ends with a brief discussion of this work in the context of the work of others.

**Chapter 6:** "The Conclusions", provides a short précis of the ideas which led to the hypotheses, and an assessment of the extent to which the aims of this work have been achieved. Chapter 6 also includes speculation on how the work so far can be taken forward.

# Chapter 2: Technical Background

This chapter provides background material on the Ear and Psychoacoustics, Automatic Speech Recognition, Speech Models, the Discrete Fourier and Cosine Transforms, the Artificial Neural Network and other pattern recognition techniques; and is purposed to providing a context for the Literature Review.

Mathematical expressions are generally not included in Chapter 2, and this for two reasons. Firstly, they are concerned primarily with implementation rather than concept; and secondly, often they are immediately available on the Internet, or failing that, from the references cited. One exception to this is the section on the Discrete Fourier Transforms; where the specific forms for the equations are integral to the description of the DFT algorithm that is used throughout this work.

The **'Glossary'** provides additional background material and also includes a list of abbreviations and acronyms. It is located immediately before the Bibliography.

## 2.1 The Ear and Psychoacoustics

The ear comprises the auditory canal which funnels sound to the eardrum, three ossicles (tiny bones), the malleus, the incus and the stapes, which transmit the sound at the eardrum to the flexible membrane which covers the oval window of the cochlea, and the cochlea which is the biological transducer that converts the sound pressure wave at the oval window into nerve impulses (Crouch [1981]).

The cochlea is a coiled tube of bone that is closed at one end, having a notional tube length of around $32mm$ (Koch et al. [2017]), which varies from person to person; and which is divided for almost its entire length by the basilar membrane which is integral with the organ of corti. The cochlea is fluid filled such that the vibrations at the oval window — the result of the motion of the malleus, the incus and the stapes — causes waves within the fluid which move the basilar membrane; and these movements are detected at the organ of corti, where they are converted into nerve impulses that are transmitted by the auditory nerve bundle to the brain (Gold et al. [2011]).

The basilar membrane is relatively narrow and stiff near the oval window and becomes progressively less narrow and less stiff along its length, and when the sound pressure wave modulates the fluid within the cochlea, the higher frequencies cause motion in the basilar membrane and stimulation of the organ of corti receptors nearer the oval window, whereas the lower frequencies penetrate further into the cochlea and cause motion where the basilar membrane is less thin and less stiff (Gold et al. [2011]). So the motion of the basilar membrane — and hence the organ of corti receptors — at any given position is thought to be a function of frequency, and accordingly it is generally believed that the cochlea performs some form of spectral analysis — a belief that is supported to an extent by scans of auditory 'processing' in the brain which show that different frequencies appear to be tonotopic (processed in different parts of the brain). These beliefs have resulted in the adoption of the filter-bank — a set of bandpass filters which span the audible range (see Figure 3.5) — as the often preferred method for modelling short segments of speech (of 10$ms$ or so).

The sensitivity of the auditory system to loudness changes is generally thought to be considerably greater than the passive mechanisms described in the previous paragraphs can sustain, and it is commonly believed that there must exist within the auditory system a source of gain — whether due to resonance, amplification (positive feedback) or some other mechanism. It may be conjectured that the masking effect discussed in Section 1.3 is the result of this gain, insofar as frequencies near the boosted frequency will be perceived to be quieter by comparison. Whatever the cause, masking is a measurable psychoacoustic effect (see Chapter 3 and Appendix B).

The filter-bank in various forms has been, and remains, an important analytical instrument in the field of speech processing; yet our knowledge on the human processing of speech is very limited. The description of the cochlea provided above is very much a simplification of what is a completely inaccessible biological structure, and which itself is merely one component in the complex auditory processing system. Additionally, the human system of speech perception is further complicated in other ways. For example, the auditory channel does not always have primacy when decoding speech; and it can be demonstrated that speech sounds as perceived can be changed by contrary visual stimulus. McGurk and MacDonald [1976]/McGurk and MacDonald [1978] found that the perception of a consonant can be changed if the listener is simultaneously viewing the facial movements for a different consonant (the visual sense dominates). It can also be demonstrated that our capacity to understand speech that has been distorted so that it retains certain auditory clues but is otherwise indecipherable, can be increased if we are prompted in advance with the undistorted version (demonstrations of both of these effects can be found on-line). For the former of these, higher cognitive functions are modifying the perception of speech in the mind; and for the latter the mind is re-interpreting speech in the light of prior knowledge.

It is only possible to ascertain from the biology of the auditory channel a simplistic view of the actual mechanisms of speech perception, and from brain scans the locations of intense neural activity which correspond to the type of stimulus; but with psychoacoustics — the testing of our perception of auditory stimulus in a controlled environment — we can at least understand and measure to an extent what we consciously perceive to be true.

## 2.2 Automatic Speech Recognition

Automatic Speech Recognition (ASR) is the process of generating a model of speech from the digitised waveform and comparing the model with a pre-determined set of models with the aim of translating the speech into text. A key characteristic of speech is that it has the form — when viewed as a time series

— of a Hidden Markov Model (HMM); comprising a sequential set of discrete states with the Markov property, where the next state is a stochastic function of the current state only (Ghahramani [2001]).

**The Markov Property:** A simple analogy for the Markov property is described in Mlodinow [2009] which is the 'Random Walk', or the 'Drunkard's Walk'. The essential point of the analogy is that the route to the current position is not of significance; and yet — randomness not withstanding — the options for the next step are a limited function only of the current position. How this might apply to speech processing is shown in Figure 2.1. Assume that the waveform is sliced into 20ms sections as indicated by the vertical bars, and the magnitude spectrum for each of the slices is generated using the DFT. By normalising the spectrum, and then comparing the frequencies in the spectrum with a library of spectra for the different speech sounds, it is possible to classify the sound and identify the specific phoneme. It can be seen from Figure 2.1 that the rate of changes of the waveform when referenced to the 20ms slices is slow (for some slices quite possibly zero), and this means that two or more slices are likely to result in very similar spectra. This is an example of the Markov property in that a significant probability exists that the sound identified for a speech slice will be the same as for the previous slice. Also certain combinations of different speech sounds are more likely than others, so if the spectrum for a following sound is different to the current sound, the probability of what follows is still a function of the current slice.

**The Hidden Markov Model:** The state of an HMM is hidden and is manifested only through a sequence of observable events, so it can only be compared with other state templates in a probabilistic way, by using the observable sequence of events. So the probability of recognising speech is firstly a function of the probability that the model generates the sequence, and secondly a function of the degree by which the hidden process/state of the model can be associated with the observable process/state of the template (section 8.2 Huang et al. [2001]). That is, HMM ASR is a "doubly stochastic" process (Section 2.3. Virtanen et al. [2013]); and to identify the hidden state of the process, it is necessary to find the best match for the observable effect in the set of observable state templates, and adopt

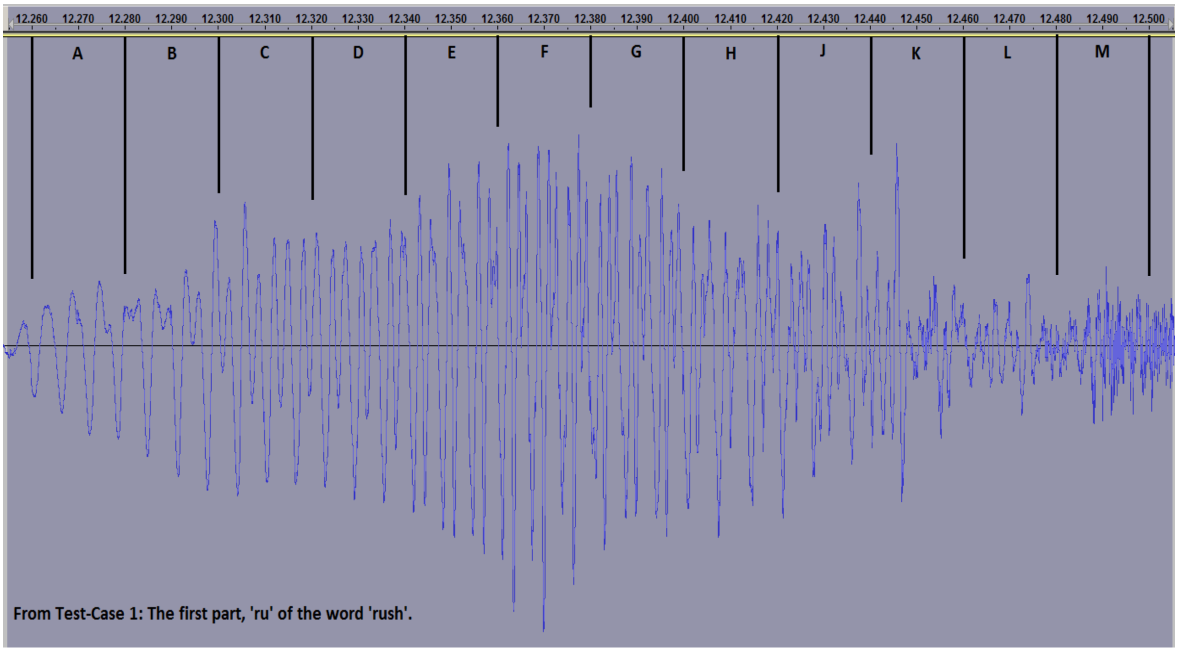


Figure 2.1: The Markov Property

The waveform should be sliced such that each slice is shorter than the speech phonemes ('the smallest significant unit of sound in a language' — Chambers [1999]) so that a DFT of the slice can accurately model individual speech sounds. Because the slice is shorter than the phonemes, there is a significant probability that a speech sound will be followed by a similar speech sound. Also because not all combinations of speech sounds are equally common, the probability of what follows a speech sound (if not the same) is a function of that sound. These are both examples of the Markov property, which is where the probability for the next state is a function of the current state only.

that in lieu of the hidden state. Identifying the states from a library of templates is the basis of HMM ASR.

**Posterior Probabilities** For ASR, finding the next state would involve matching the HMM with a potentially large set of templates each with its own probability; but with knowledge of the current state, the probabilities for the next state could be modified according to Bayes' Decision Rule (section 4.1 Huang et al. [2001]). Intuitively, these posterior probabilities can be seen to be more accurate because of the Markov condition that the next state is a function of the current state.

## 2.3 Models of Speech

### 2.3.1 Linear Predictive Coding

The Linear Predictive Coding (LPC) model (Section 21.2 of Gold et al. [2011]) is an all pole (resonance) model of speech, which assumes that speech is produced from a simple concatenation of tubes each of the same length, but having different cross sectional areas, and as a consequence different resonant frequencies. With the gains for each of these 'resonators' made variable, then 6 sections would be enough to span speech of 5 kHz bandwidth to represent the five formants and the variations in the driving waveform that constitute a vowel. The difference between the model and the original signal is the 'prediction error', and when the gain terms (coefficients) are chosen to minimise this, then the prediction error can be considered to be an approximation of the excitation function.

The purpose in creating the Linear Prediction Envelope is to de-emphasise the harmonic representation of the speech whilst capturing the formants. See Figure 3.19, pg 76, in Johnson [2012], which is a comparison of the FFT Spectrum, with its corresponding LPC Spectrum. This shows that the LPC identifies the formant peaks of the vowel, and does not suffer from aliasing as may occur with the DFT, where every peak in the spectrum is at an integer multiple of the reciprocal of the DFT window duration. Johnson goes on to outline the shortcomings of LPC analysis, and these include a lack of sensitivity to anti-formants which results in poor performance with nasals, laterals, and some fricatives. Conversely Moore (Bristow [1986] pg 133) states that LPC is, "Particularly good", for identifying the spectral peaks of vowels.

### 2.3.2 Mel Frequency Cepstrum Coefficients

The Mel Frequency Cepstrum Coefficients (MFCC) speech model is purposed to provide a better representation of the, "Perceptually relevant" features of the short-term speech spectra (Davis and Mermelstein [1980]); and was for many years of importance in the field of ASR, though in recent years interest in the model has declined in favour of often simpler models (Xiong et al. [2016]) — perhaps simplified to the point of using the raw speech data (Sainath et al. [2015]) — for ASR with Deep Neural Networks.

To create the MFCC, the audio spectrum is divided into equally spaced (in Mel) overlapping bands, and separately a set of filters which overlay these bands is constructed ( see Figure 3.5). The harmonics in each of the bands are scaled by the corresponding filter, and then the log-energy for each of the bands is calculated (White and Neely [1976]). Finally the discrete cosine transform (DCT) of the scaled log-energies results in the Mel Frequency Cepstral Coefficients. Conventionally (Jurafsky and Martin [2009]) the term MFCC is taken to mean the first 12 of these coefficients plus the energy in the period, plus a further 13 corresponding velocity terms, plus a further 13 acceleration terms resulting in 39 MFCC features.

## 2.4 The Analysis of Speech In the Frequency Domain

### 2.4.1 The Discrete Fourier Transform

Any non-sinusoidal periodic waveform can be specified as the sum of a series of sine terms and cosine terms plus a constant term. This, the Trigonometric Fourier Series, is defined by the equations: 2.1 to 2.5 below (pg 102–109 Ivison [1978], pg 910–921 Gullberg [1997])

$$f(t) = \frac{a_0}{2} + a_1 cos\omega t + a_2 cos2\omega t + ...a_n cosn\omega t \quad + \\ b_1 sin\omega t + b_2 sin2\omega t + ...b_n sinn\omega t \tag{2.1}$$

where $\omega = 2\pi freq$ is the angular frequency expressed in radians per second.

The coefficients $a_i$ and $b_i$ are given by

$$a_i = \frac{1}{\pi}\int_0^{2\pi} f(t).cosi\omega t.d(\omega t) \tag{2.2}$$

$$b_i = \frac{1}{\pi}\int_0^{2\pi} f(t).sini\omega t.d(\omega t) \tag{2.3}$$

Equation 2.1 can be written as:

$$f(t) = \frac{a_0}{2} + c_1 cos(\omega t - \phi_1) + c_2 cos(2\omega t - \phi_2) + ... + c_n cos(n\omega t - \phi_n) \tag{2.4}$$

where the magnitude ($c_i$) and phase ($\phi_i$) of the harmonics are calculated from Equation 2.1 as follows:

$$c_i = (a_i^2 + b_i^2)^{\frac{1}{2}} \quad and \quad \phi_i = tan^{-1}\left(\frac{b_i}{a_i}\right) \tag{2.5}$$

#### The Algorithm for the Calculation of Magnitude and Phase

It can be seen from the first terms of the sine and cosine expansions in Equation 2.1, that the fundamental frequency is the reciprocal of the period (the DFT window duration), and thereafter that the harmonics are integer multiples of the fundamental frequency.

According to Ivison [1978], and in accordance with Shannon's sampling theorem (Shannon [1949]), the highest harmonic, $n$ that can be generated to an acceptable degree of accuracy is a function of the number of samples in the period such that:

$$n \leq \frac{number\_of\_samples}{2} - 1$$

**Calculation of Coefficients :** (Ivison [1978], Gullberg [1997]) The coefficients $a_i$ and $b_i$ of equation 2.1 are calculated as follows:

Where $X$ = the number of samples in the period; and the symbol ":=" means "becomes equal to":

**The coefficients for the fundamental a_0 and b_0:**

```
angular_increment := $\frac{2\pi rad}{X}$
for i in (1 $to$ $X$)
loop
{
        $\theta$Array(i) := angular_increment * i
        $y$Array(i) := sample_magnitude at angular_increment
        $ycos\theta$Array(i) :=$y$Array(i).(cos($\theta$Array(i)))
        $ysin\theta$Array(i) :=$y$Array(i).(sin($\theta$Array(i)))
}
```

$$a_0 := \frac{2*\left(\sum_{i=1}^{X} ycos\theta Array(i)\right)}{X}$$

$$b_0 := \frac{2*\left(\sum_{i=1}^{X} ysin\theta Array(i)\right)}{X}$$

**The coefficients for the n$^{\text{th}}$ Harmonic:**

```
angular_increment := $\frac{2\pi rad}{X}$
for i in (1 $to$ $X$)
loop
{
        $n\theta$Array(i) := n.(angular increment * i)
        $y$Array(i) := sample magnitude at angular_increment
        $ycosn\theta$Array(i) :=$y$Array(i).(cos($n\theta$Array(i)))
        $ysinn\theta$Array(i) :=$y$Array(i).(sin($n\theta$Array(i)))
}
```

$$a_n := \frac{2*\left(\sum_{i=1}^{X} ycosn\theta Array(i)\right)}{X}$$

$$b_n := \frac{2*\left(\sum_{i=1}^{X} ysinn\theta Array(i)\right)}{X}$$

The magnitude and the phase for the fundamental and the harmonics are then calculated by substituting for $a_n$ and $b_n$ in the equations at 2.5.

### Waveform Reconstruction

The Discrete Fourier Transform is a reversible process in that the speech waveform can be constructed from the magnitude and phase spectra (see equation 2.1).

In the following pseudo-code, $H_1$ to $H_2$ is the range of harmonics in the reconstruction, $X$ = the number of samples during the Fourier window; and the symbol ":=" means "becomes equal to". The Local_Scaler_Array is populated with one of:

- unity for speech reconstruction.
- the conventional auditory filter — illustrated in Figure 4.3A.
- a scaling regime for simultaneous masking — illustrated in Figure 4.3B .

```
$Result\_Array(1\ to\ X) := 0.0)$
$for\ i\ in\ (H_1\ to\ H_2)\ \ loop$
    {
    $for\ x\ in\ (1\ to\ X)\ \ loop$
        {
        $Result\_Array(x)\ := \ Result\_Array(x) +$
            $Local\_Scaler\_Arr(i).(\ a_n(i).cos(n\theta Array(i)(x))$
                $+\ b_n(i).sin(n\theta Array(i)(x)\ )$
        }
    } $return\ Result\_Array(1\ to\ X)$
```

### The DFT and Spectral Estimation

In a work on analysing sibilant fricatives, Reidy [2015] uses the term, "Spectral estimator" for the process of using the DFT to convert what he terms random data in the waveform into a multivariate statistic (spectral estimate). Putting the question of the composition of sibilant fricatives aside, there is still an issue with the DFT where although the terms of the Fourier expansion correctly define the waveform, the resultant spectrum is a function of the DFT window duration. That is, the DFT translates all frequencies in the waveform into combinations of multiples of the DFT fundamental frequency, and this results in a smearing of the original frequencies. So to an extent, for all frequencies the DFT can be regarded as providing only a spectral equivalence of the waveform.

### The Hamming Window

Speech is not a true non-sinusoidal periodic waveform, and accordingly when a rectangular window of some duration is chosen as the period for the DFT, the first sample of the window will not be the same as the first sample of the next period. Such a discontinuity would result in additional harmonics in the spectra that are not a function of the required speech content. The Hamming Window (see Figure 2.2) is often used in place of the rectangular window to reduce this effect, and its use in ASR is formalised in the European Telecommunications Standards Institute standard on, ***"Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms"*** ETSI-ES-202-050 [2007].

When reconstruction of the speech is required, the Hamming window may not be the most convenient choice (see Figure 2.3). An alternative method is to use a $30ms$ rectangular window, stepping $10ms$, and to use only the $10ms$ at the centre of the window for waveform reconstruction. With this method, speech waveforms can be reconstructed by summing the terms of the Fourier Series; and there is found to be little difference between the reconstructed speech and the original speech. (The techniques for speech reconstruction are far from new. In 1979, Boll described a system purposed for the spectral subtraction of noise components of

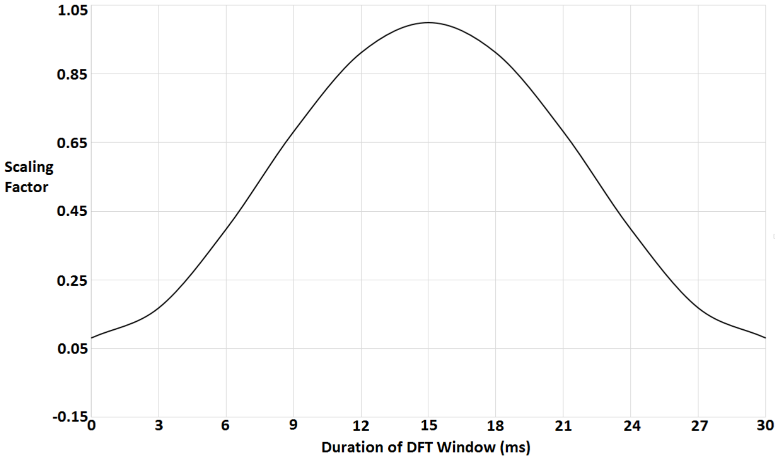


Figure 2.2: The Hamming Window

It can be seen that the waveform when scaled by the Hamming Window will adhere more closely to the non-sinusoidal periodic waveform necessary for the DFT, and the generation of harmonics due to the discontinuities at the start and end of the Fourier Period will be reduced. It is evident from Figure 5.20 parts (c) and (d) in Huang et al. [2001] that attenuation of the Hamming window is almost constant across the frequency range of interest. The graph is constructed using Equation 9.11 — Jurafsky and Martin [2009].

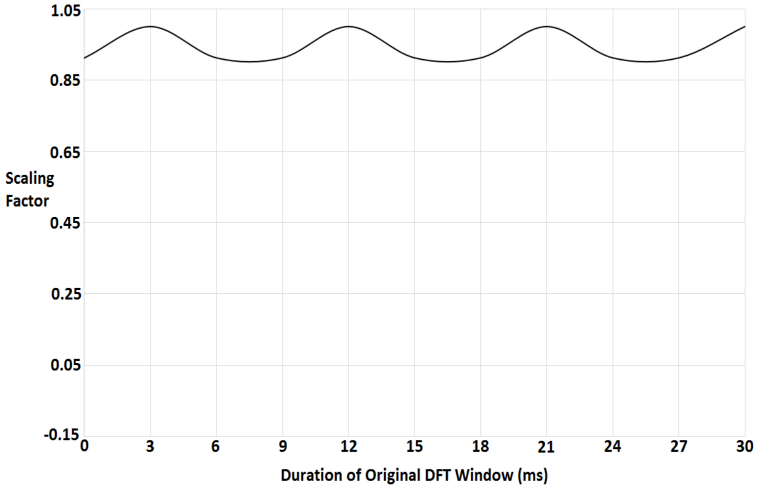


Figure 2.3: Waveform Reconstruction and the Hamming Window

If the waveform when scaled by the Hamming Window is reconstructed using only the central portion of the window, then this will result in an amplitude modulation of the speech – although this effect can be reversed by scaling the reconstructed waveform for the appropriate duration with the phase inversion of the amplitude modulation.

digitised speech.)

### 2.4.2 The Discrete Cosine Transform

**The Discrete Cosine Series** is a ‘half range expansion’ (Wylie and Barrett [1982]) in the frequency domain comprising of a constant plus the cosine series. Equation 2.1, the Trigonometric Fourier Series, reduces to:

$$f(t) = \frac{a_0}{2} + a_1 cos\omega t + a_2 cos 2\omega t + ...a_n cos n\omega t \tag{2.6}$$

**Rationale**: From Equation 2.1 for the Trigonometric Fourier Series it can be seen that each element of both the sine and cosine series is an integer multiple of the fundamental frequency and from the Magnitude and Phase equations (2.5) that the phase is embodied at each harmonic in the ratio of the coefficients. This means that all of the sine terms in the series have the same phase (coincident zero-crossing points); and the same as true for the cosine terms — though all of the cosine terms are shifted in phase by $\frac{\pi}{2}$ rad.

Figure 2.4 shows that the cosine function (A) has even symmetry about $t_i$ in that the magnitude at time $t_i - t_x = t_i + t_x$ for any time duration $t_x$ within the range $\frac{DFT\ Period}{2}$, whereas the sine waveform (B) has odd symmetry about $t_i$ in that the magnitude at time $t_i - t_x = -(t_i - t_x)$ for any $t_x$ within the range $\frac{DFT\ Period}{2}$. This means that the Fourier Transform of any even functions may only comprise a constant term (which may be zero) plus cosine terms. All of the sine terms must equal zero, because if any sine terms remain following the transform then were the waveform to be reconstituted from the harmonics, it would include odd symmetry. The converse of this is also true, specifically the Fourier Transform of any odd functions may only comprise a constant term (which may be zero) plus sine terms. All of the cosine terms must equal zero, because if any cosine terms remain following the transform then were the waveform to be reconstituted from the harmonics, it would include even symmetry.

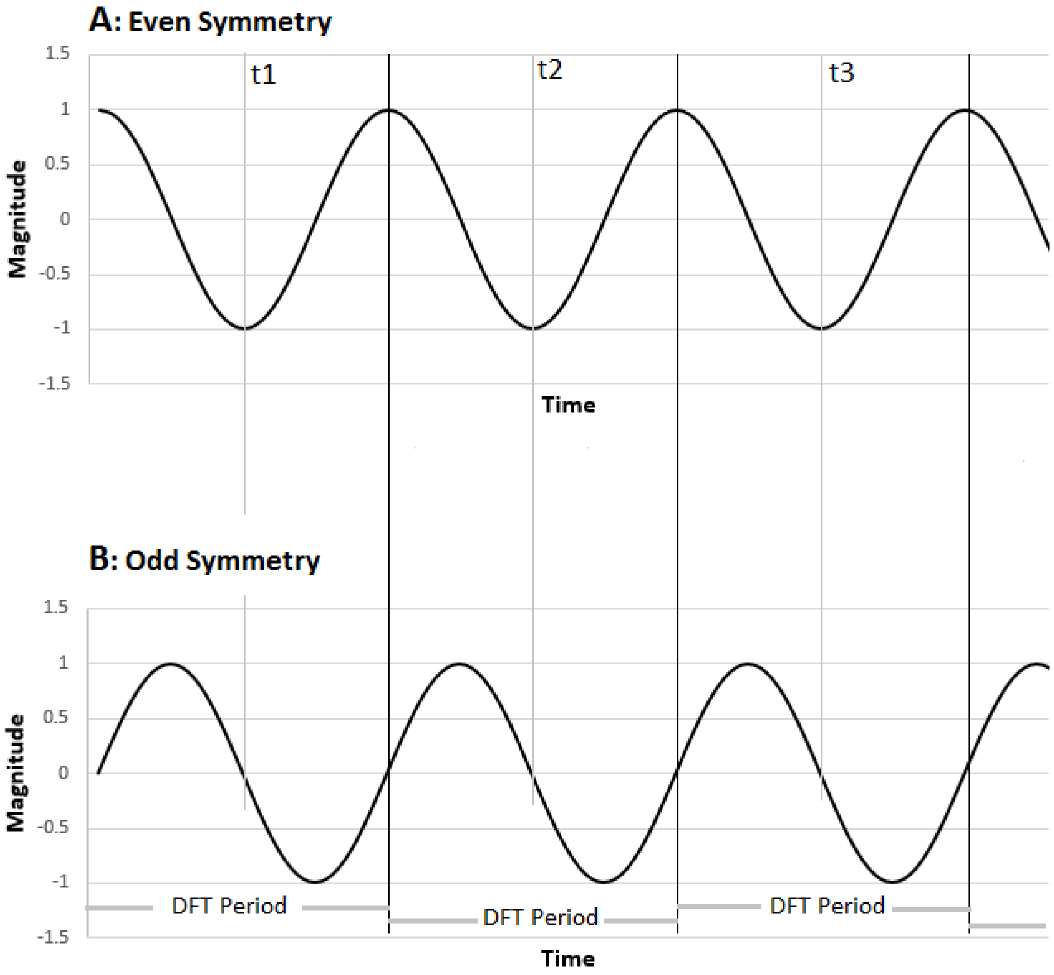


Figure 2.4: An Illustration of A: Odd Symmetry, and B: Even Symmetry

It can be seen that the cosine function (A) has even symmetry about $t_i$ in that the magnitude at time $t_i - t_x = t_i + t_x$ for any time duration $t_x$ within the range $\frac{DFT\ Period}{2}$, whereas the sine waveform (B) has odd symmetry about $t_i$ in that the magnitude at time $t_i - t_x = -(t_i - t_x)$ for any $t_x$ within the range $\frac{DFT\ Period}{2}$. This means that the Fourier Transform of any even functions may only comprise a constant term (which may be zero) plus cosine terms. All sine terms must equal 0, otherwise the function would contain odd symmetry, and would therefore not be an even function.

## 2.5 Pattern Recognition

### 2.5.1 The Artificial Neural Network

The Artificial Neural Network (Priddy and Keller [2005], da Silva et al. [2017]) is a so-called ‘intelligent’ system where the ANN ‘learns’ through experience, and can acquire a set of responses during training which can be more generally applied to different stimuli of a similar type.

The first modern expression of the Artificial Neural Network was by Rosenblatt [1958] who coined the term 'Perceptron' when expanding on the work of McCulloch and Pitts [1990] (a reprint of the 1943 original paper). Rosenblatt in addressing the questions of how information is remembered, and what the influence of stored information on behaviour and recognition is, wrote that (and this forms the basis of the model of the Artificial Neurone):

> **"...there is never any simple mapping of the stimulus into memory, according to some code which would permit its later reconstruction. Whatever information is retained must somehow be stored as a preference for a particular response; i.e., the information is contained in connections or associations rather than topographic representations."** – Rosenblatt [1958]

In Figure 2.5, 'The Model of an Artificial Neurone ', Rosenblatt's associations are represented by the weights on each of the inputs. Figure 2.5 also shows the Bias Input, and the Activation Function. The Bias Input is necessary for those artificial neurones which comprise the internal layers of an ANN because if all of the inputs to the neurone are zero, then none of the inputs can be meaningfully weighted and the output state would equal zero. With the bias input node fixed at 1.0, this ensures that at least one of the nodes can be meaningfully weighted during training. The Activation Function (i.e. the transfer function) dictates the configuration and performance capabilities of the ANN, and may be, for example, a step function, a linear ramp, or a non-linear Sigmoidal or Gaussian function.

The Single Layer Perceptron uses either a step or linear ramp activation function, and its capability as a pattern recognition device is limited to linearly separable data. Later the use of a non-linear activation function, with Backpropagation training (da Silva et al. [2017] citing Rumelhart et al. [1986]) facilitated — for example — network configurations such as the Multi-Layer Perceptron (MLP) and the Radial Basis Function Network (RBFN); which

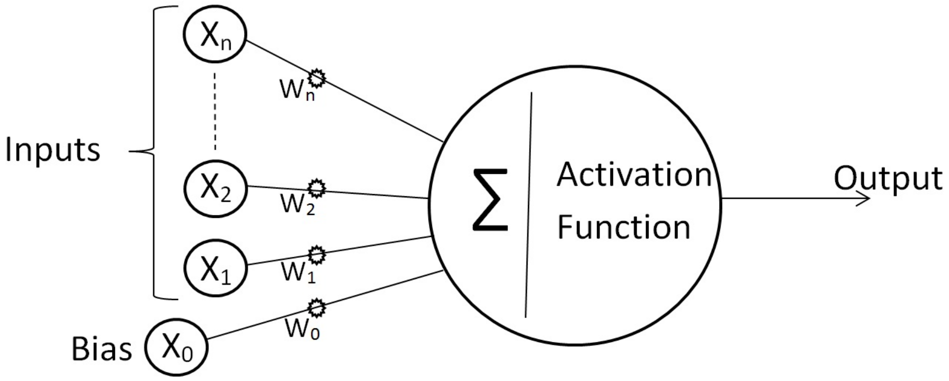


Figure 2.5: Model of an Artificial Neurone.

This illustration is based upon Figure 1.4 of Priddy and Keller [2005], and Figure 1.4 of da Silva et al. [2017]; and the Neurone implements the following feed-forward equation:

$$Output = (\sum_{i=0}^{n}(Xi.Wi)\ ).(activationfunction) \tag{2.7}$$

extended the capability of the ANN as a pattern recognition device to include non-linearly separable data.

The MLP is configured with an input layer, one or more hidden layers, and an output layer; and is trained using a Supervised Training technique — which is where the training data is a representative subset of the full data set, and consists of the input data together with the required classification results. The RBFN is configured with an input layer, one hidden layer, and an output layer; and is trained using a Supervised Training technique for the output layer only, whereas the hidden layer — which uses a Gaussian activation function — is firstly trained using an unsupervised self organizing mapping technique.

An Artificial Neural Network (see Figure 2.6) comprises one or more Neurones arranged in one of various architectures including: Single Layer Feed-Forward, Multiple Layer Feed-Forward, Recurrent, and Mesh (da Silva et al. [2017]). The Single Layer Feed-forward architecture facilitates the classification of linearly

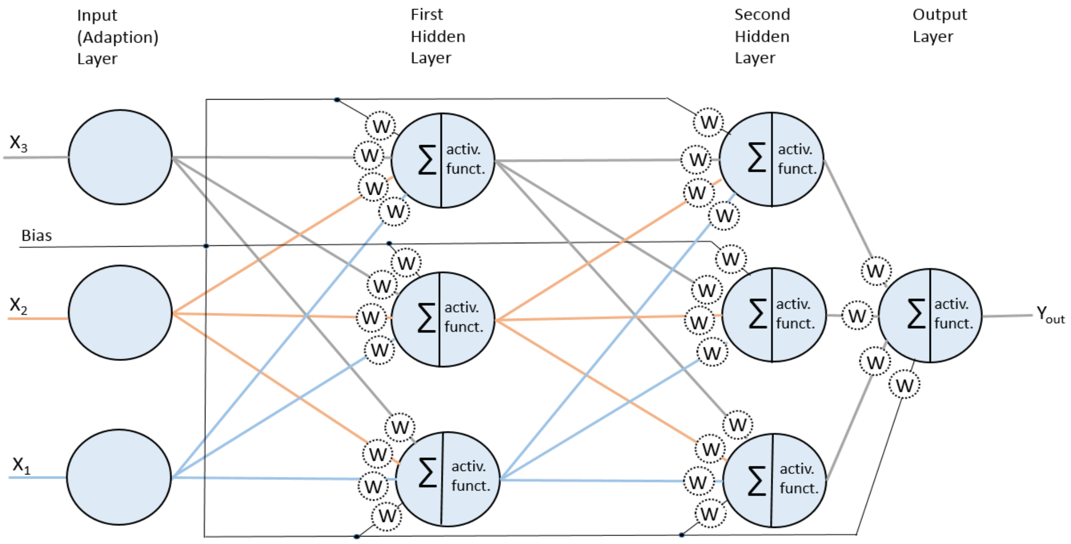


Figure 2.6: Artificial Neural Network with Two Hidden Layers

This is a schematic of a Multiple Layer Feed-Forward Deep Neural Network. Here the stimulus can only propagate from the input layer forward through the hidden layers to the output layer of neurons. i.e. There is no feedback mechanism—as would be the case with a recurrent neural network.
The input layer may have minimal functionality serving only to route the input data to each of the layers in the first hidden layer, or may include additional functionality — for example to rescale the input data to the operational range of the Neural network. The remainder of the neurones in the network each implement Equation 2.7.
For any given application, the configuration of the input layer, and output layer for an ANN are dictated by the system requirements; but it is less clear what might be the optimal number of hidden layers, and what might be the optimal number of neurons in those layers. Priddy and Keller [2005] state that ANN learning is a function of mapping so for best performance the minimum number of hidden layers which can support the functionality should be chosen; and suggest that the optimal number of neurons in the hidden layers can be determined from the validation-set error (i.e by trial and error). They state that the size of the training set must increase if the number of hidden neurons is increased—simply because more weights require more training.

separable data, The Multiple Layer Feed-forward architecture facilitates the classification of non-linearly separable data, the Recurrent architecture imposes some history (memory) by incorporating the previous result in the current data, and the Mesh architecture simulates neural topography — tonotopy for example — by arranging the Artificial Neurons so that particular regions of the neural network are made responsive to particular ranges of the data. The methods of training include variations of: Supervised, Unsupervised, Reinforcement, Off-line and On-line(da Silva et al. [2017]). For Supervised Training the training data set includes the desired outcome, for Unsupervised Training the training data set does not include the desired outcome and the Neural Network organises itself to recognise subsets amongst the training data, for Reinforcement Learning the actual response is compared continuously with the desired response and if satisfactory the weights are gradually changed to ‘reward’ this behaviour, for Off-line learning the full training set (1 epoch) is processed before the weights are adjusted, and for On-line learning, the weights are adjusted for each training sample.

### Forward Propagation

Forward propagation is a feed-forward process and is used with a trained Artificial Neural Network to classify fresh data into one of the trained groups; and is also an essential component in supervised training, where it is used to generate the outputs for all of the neurones in the network.

The process uses Equation 2.7 (from Figure 2.5) to generate the outputs for each of the neurones which constitute the hidden layers and the output layer; starting with all of the neurones in the first hidden layer and propagating the neurone outputs from the first hidden layer to the inputs of the next hidden layer, and so on, through to the output.

### Backpropagation

Backpropagation is a systematic process used for the supervised training of Artificial Neural Networks.

If the essential condition for backpropagation that the activation functions be differentiable is met, and knowing the output state of all of the neurones within the ANN in response to a training data item; then the network error — the difference between the actual and the required outputs of the network — can be apportioned to the individual neurones in the network, and the weights adjusted to reduce these errors, as follows.

For each training example, the forward propagation process is applied and the outputs for all of the neurones which constitute the network are stored. From Figure 2.6 it will be seen that if the outputs for all of the neurones are known then, ipso facto, the inputs for all of the neurones are known.

**Adjusting the Output Layer Weights:** The error across the output neurone referenced to input channel $i$ is a function of the weight at input $i$ and is denoted $\Delta E = \frac{dE}{dW_i}$, where $W$ is the weight matrix for the neurone; and the change that is to be applied to the weight at input $i$ is given by

$$\Delta W_i = -\eta . \frac{dE}{dW_i} \tag{2.8}$$

where $\eta$ is the learning rate (usually between 0 and 1.0), and the minus sign ensures that the error correction is applied so as to reduce the error gradient (i.e. Gradient Descent).

To solve Equation 2.8, the term $\frac{dE}{dW_i}$ must be simplified, and this is achieved using the chain rule for the differentiation of functions such that

$$\Delta E = \frac{dE}{dW_i} = \frac{\partial E}{\partial Y_i} . \frac{\partial Y_i}{\partial I_i} . \frac{\partial I_i}{\partial W_i} \tag{2.9}$$

and these terms equate to

$$\frac{\partial E}{\partial Y_i} = the\ quantity\ at\ input\ i\ of\ the\ output\ neurone \tag{2.10}$$

$$\frac{\partial Y_i}{\partial I_i} = the\ first\ order\ derivative\ of\ the\ activation\ function \tag{2.11}$$

$$\frac{\partial I_i}{\partial W_i} = -(required\ output - actual\ output) \tag{2.12}$$

**Adjusting the Second Hidden Layer Weights:** When adjusting the weights for the second hidden layer, Equation 2.8 is again used, except that for all of the neurones in the second hidden layer, the 'required output' used in Equation 2.12 is not known, and must be calculated. That is, the error attributed to the output neurone is backpropagated to the second hidden layer, and this is used to calculate the required outputs for that layer.

**Adjusting the First Hidden Layer Weights:** This is a similar process to that used to adjust the second layer hidden weights, except that in this case, the error attributed to the output neurone plus that attributed to the second hidden layer neurones is backpropagated to the first hidden layer, and this is used to calculate the required outputs for that layer.

The process described only results in weight corrections for a single instance of the many training data items which constitute the training data set, and must be repeated for the entire epoch — that is for all of the items in the training set. Accordingly, the final weight corrections can only be known after completion of the entire epoch; whether or not Off-line or On-line Training is used. Upon completion of each epoch, a measure of performance of the network — often the differences between the required output and the achieved output expressed as the Mean Squared Error (MSE) — is compared with the training target (the maximum acceptable MSE), and if the target has not been achieved then another training epoch is implemented. In practice, often many training epochs will be required.

**Resilient Propagation (RPROP):** Backpropagation as previously described may require tweaking of the learning rate to obtain a satisfactory outcome; and furthermore may be slow to converge to a solution requiring possibly many hundreds of training epochs. Resilient Propagation (RPROP) by Reidmiller and Braun [1993], is an example of an improved off-line backpropagation algorithm which eliminates the learning rate parameter, and which also significantly ameliorates the problem of slow convergence by replacing the weight corrections obtained during training, with weight corrections that are increased progressively epoch by epoch until the sign of any of the errors changes — which indicates to the algorithm that the particular weight correction is too large and must be reversed.

### 2.5.2 Other Pattern Recognition Techniques

The intention in this work is to mimic, insofar as that is possible, human interface methodologies; and in that the Artificial Neural Network originated from the field of neuroscience, it fulfills that criterion. However, at least two other distinct methodologies for pattern recognition (both statistical) exist; the first is the Support Vector Machine (grounded in Co-ordinate Geometry), and the second is the Random Forest (grounded in Binary Decision Trees). A brief introduction to both of these methodologies follows.

**Support Vector Machine:** (Cortes and Vapnik [1995]) (Section 8.4 Gold et al. [2011]) (James et al. [2013]) The Support Vector Machine (SVM) is a learning machine, where two sets of data that are not linearly separable can be made linearly separable using a transformation into an immensely high dimensional space. Conceptually, the support vector machine operates by mapping the independent variable(s) of a time varying function (for the case of speech processing) using one or more non-linear functions, thereby creating a multi-dimensional feature space that is modulated by the original data; and then finding the optimal hyperplane for separating the two data sets in the extended feature space.

The term, "Support Vector", is commonly used in connection with three different classifiers: the Maximal Margin Classifier and the Support Vector Classifier — both considered in Figure 2.7, plus the Support Vector Machine. The maximal margin classifier is suitable for the classification of linear separable classes, the support vector classifier (a soft margin classifier) is suitable for mostly linearly separable data but will tolerate some observations on the wrong side of the hyperplane, and the support vector machine is suitable for the classification of data that is not linearly separable.

The separating hyperplane has one dimension less than the number of co-ordinates of the data space, thus the hyperplane for a two dimensional classifier is a line, and for a three dimensional classifier a plane surface.

For the Maximal Margin Classifier — shown in Figure 2.7 — it may be seen that an infinite number of hyperplanes exist which would separate the classes (i.e. lie within the shaded areas), but that geometrically, a unique maximal margin hyperplane exists; and that is the hyperplane which is furthest from the training data instances and so yields the maximum margin. It can also be seen that only three separate data instances — the support vectors — are required to define the margins and the maximal margin hyperplane. The other training data instances have no effect.

The Support Vector Classifier is similar to the maximal margin classifier except that the former will tolerate some data instances on the wrong side of the margins

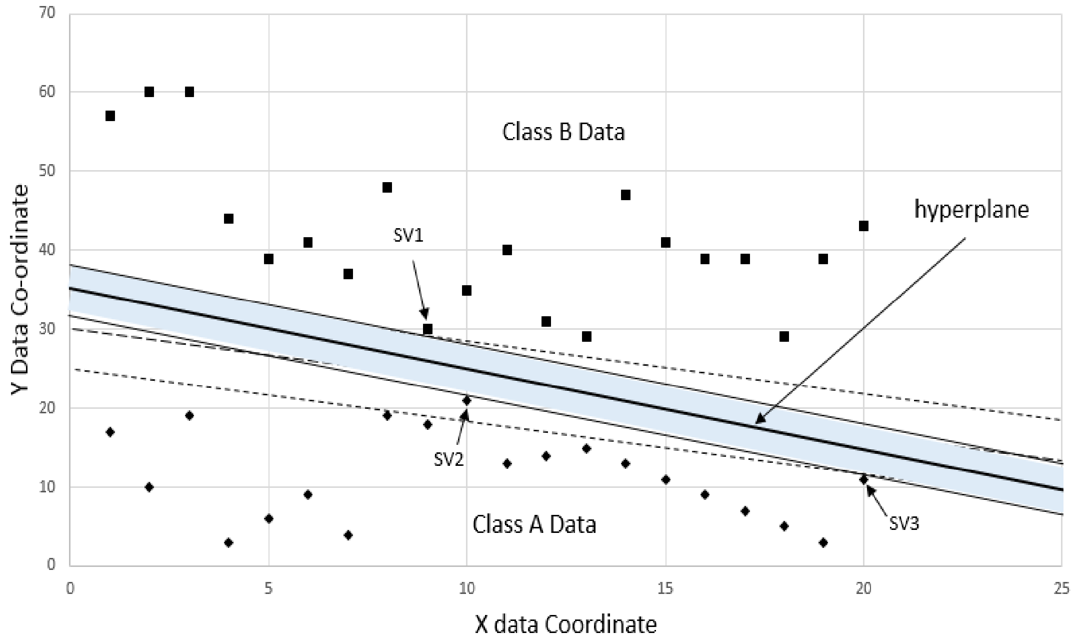


Figure 2.7: The Maximal Margin Classifier and the Support Vector Classifier

**Maximal Margin Classifier:** The illustration shows two linearly separable classes of training data instances — Class A indicated by diamonds, Class B by squares. Geometrically, three support vectors (arrowed) are required to delineate the maximal margins (shaded), and the maximal margin hyperplane (arrowed).
**Support Vector Classifier**: It can be seen that if support vector 2 (for example) is allowed to be within the margin, (i.e. effectively removed from the calculation of the maximal margin hyperplane), then a different, and wider separation of the classes (shown in dashed lines) is obtained; providing perhaps a better classifier at the cost of what may be an acceptable level of mis-classification.

and even the wrong side of the hyperplane. For the Support Vector Classifier, the support vectors are the training data instances that lie on the margin boundary, or the wrong side of the margin boundary for their class. The other training data instances have no effect.

The Support Vector Machine, unlike the Maximal Margin Classifier and the Support Vector Classifier, is non-linear classifier and operates conceptually by creating a multi-dimensioned feature space with the concomitant multi-dimensional hyperplane; and this is achieved by modulating the training data instance set using a kernel, which is the non-linear mathematical function that spawns the multi-dimensions, and which must be selected to accommodate the specific nature of the classification problem. Although virtually impossible to visualise, this can be realised using the inner product of the training data instance pairs rather than the actual instances themselves; and it can be shown mathematically, that the inner product of the training data instances that are not support vectors is zero. That the solution lies with the evaluation of the inner products of the training data instance pairs rather than in the multi-dimensional feature space simplifies what otherwise would be a complex and possibly intractable computational problem. i.e. Consider the level of complexity of three dimensional geometry compared with the complexity of two dimensional geometry, and then scale the complexity up to n-dimensional geometry — where n is a vast number and may be infinite.

An interesting point — in the context of statistical learning — is made by James et al. that since the SVM was introduced that,

> **"Deep connections between SVMs and other more classical statistical methods have emerged,"** — James et al. [2013]

and they also state — without citing their sources —that when the classes significantly overlap then the more classical statistical methods are often the preferred solution, but when the classes don't significantly overlap then SVMs tend to be better. If true, this is relevant because the extent to which the classes of noise during silence and speech overlap (i.e. inhabit each other's state space) cannot be quantified for the general case.

**Random Forests:** (Ho [1995])(James et al. [2013]): Random Forests are purposed to reduce the effects of the inherent limits on complexity for individual decision trees; and consist of many binary decision trees where each tree is built using randomly selected subsets of the feature vector. Each tree provides its own generalised classification, and when combined with other trees these classifications complement each other resulting in a monotonic improvement in overall classification performance.

A binary decision tree is 'grown' from the training data using the technique of recursive binary splitting; which is where the feature space initially forms a single area, and then this is split into two areas which correspond to two branches of the tree, and then the now two feature sub-spaces are each again split forming two more branches per split and 4 feature sub-spaces, and so on. The quality of each split is defined to be the extent to which any training observation which activates that branch of the tree belongs to the most commonly occurring class of training observations for that feature sub-space; and is quantified as the Classification Error, the Gini Index or the Entropy. These are all measures of statistical dispersion, and the smaller the number, the better the quality of the split.

The binary decision tree exhibits high variance, in that if the training set is divided into two parts, and a tree is grown for both parts, the solutions are likely to be dissimilar. High variance is a problem for any classifier, and this can be reduced for decision tree classifiers, by employing the technique of **B**ootstrap **Agg**regat**ing** (more commonly known as Bagging) — which may involve many hundreds (or even thousands) of individual trees.

The concept that underpins bagging is that if many training sets were available for the 'population', then many different decision-trees could be grown and the results of the classifications averaged; and this would reduce the variance of the classifier proportionally to the number of trees (i.e. the aggregating). In practice, many training sets are not available, and so must be synthesised from the available training data by repeatedly selecting samples from the master training data set and combining these in different ways to form many dissimilar training sets (i.e the bootstrapping).

Although bagging is purposed to solving the problem of high variance with binary decision-tree classifiers, it does so at the cost of introducing the potential for significant correlation. The correlation can occur because the technique of recursive binary splitting is a greedy technique which means the at each level of growing the tree, the best split is made for that node — whether or not a different split might result in a better tree at some later step. So if one of the features in the training data set is particularly strong, with other similar features moderately strong, then many of the bagged trees will use the strong feature in the top-split — and so will be similar and will generate correlated results; thereby re-introducing high-variance because the averaging of correlated results is effectively the averaging of similar results which provides no great reduction in variance.

Unlike Bagging, the decision trees which constitute a Random Forest are de-correlated; and this is achieved by using only a newly chosen random subset of the features vector (typically the square-root of the total number of features which constitute the feature vector) at each of the splits as each tree is grown. The results of this are that the strong feature and moderately strong features may not even be part of the feature subset — hence a reduced potential for correlation between the decision trees and for the concomitant high variance.

# Chapter 3: Literature Review

## 3.1 Introduction

There are two primary research areas involved with the topic of this thesis, and they are the detection of silence in speech and the perceptual modelling of speech. These are both purposed to deriving a model of the sound which can best capture the differences between silence and speech.

## 3.2 Detection of Silence in Speech

Atal and Rabiner [1976] rationalised that any single feature of speech will not provide enough information to support a voiced-unvoiced-silence classification, but that combining several features of the speech may provide a route to a more robust classification model. They chose 5 speech features: Short Term Energy, Zero-Crossing Rate, Auto-Correlation Coefficient, Linear Predictive Coding (LPC) First Coefficient, and Energy in the Prediction Error, but then state that a different set of parameters might provide better discrimination between the silence/unvoiced/voiced classes. With their model of speech, and employing a separate training set and test set with an unvoiced signal-to-noise ratio of 14 dB, and a voiced signal-to-noise ratio of 34 dB, the authors were able to detect 85.54 % of 94 Silences, 85.37 % of 82 unvoiced segments and 98.94 % of 375 voiced segments. Atal and Rabiner rated the effectiveness of their chosen parameters in minimising the classification errors for unvoiced/silence discrimination and found Energy to be the most effective, then Zero Crossing Rate, then Autocorrelation Coefficient, Energy in the Prediction Error and lastly the LPC First Predictor

Coefficient. This shows that the LPC derived parameters are least effective for unvoiced/silence classification. The total of 239 errors in unvoiced/silence discrimination was considerably larger than both the voiced/unvoiced error total of 91 and the voiced/silence error total of 117. These results reveal an inherent truth, and that is that the detection of voiced speech is considerably easier than the discrimination between unvoiced speech and silence. To give the results some context, with silence and unvoiced detection rates of ∼85.4%, near 1 in 7 silences or unvoiced segments will not be recognized, whereas with a voiced segments detection rate of 98.94% only 1 in 94 voiced segments will not be recognized. So although Atal and Rabiner provide a sound rationale for a pattern recognition approach, they also raise an important question: what is the best set of features for representing speech?

Ghiselli-Crippa and El-Jaroudi [1991] chose Atal and Rabiner's parameter set for their work to develop an ANN alternative fast training algorithm for voiced/unvoiced/silence classification. They used two training strategies: for the first they excluded the transitional frames between speech and silence — those frames difficult to tag as either speech or silence — from their manually derived training data set, and for the second the complete training data set was used whether transitional or not. For the former they achieved detection rates between 93.46 and 95.76%, and for the latter detection rates of between 95.31 to 96.63%. Although the authors provide no information on noise, this work is of interest firstly because it provides an early indication of the capabilities of the two layer artificial neural network, and secondly because the results of the classification improve when the training data includes the transitional frames.

Chen [1976] on the subject of achieving the optimal subset of features wrote that a full search of possible features — though "usually impossible", will provide the best feature subset. Sarma and Venugopal [1978] applied this to Atal and Rabiner's feature set, and by testing all possible combinations found that the optimal subset of features included only the Energy, the Zero-Crossing Rate and the Auto-Correlation Coefficient. Interestingly the LPC first coefficient is excluded from the best features subset, and this is not surprising as the LPC model is a primitive model of speech production. Linear Predictive Coding is described by Johnson [2012] as follows:

> ***" ...Physically sensible, if oversimplified, model of speech involving a sound source (vocal fold vibration) and a filter of several resonances"***
>
> — **Johnson [2012]**

It is interesting to compare Atal and Rabiner's parameter set with that chosen by Mondal and Barman [2015]. Mondal and Barman use six features: pre-emphasised energy ratio between consecutive frames, average zero crossing rate, short term energy, spectrum tilt, low to full band energy and spectral centroid. Mondal and Barman chose not to include in their parameter set the LPCs speech representation (a justifiable decision given the limitations of the LPC model), nor the Mel Frequency Cepstral Coefficients speech representation (which was not known about at the time of Atal and Rabiner's work). Although the work of Mondal and Barman post dates Atal and Rabiner's work by 39 years there is little evidence that the problem of voiced-unvoiced-silence detection is solved, and because Mondal and Barman manually selected a silence threshold knowing the background conditions their work does not address the question of fully automated voiced-unvoiced-silence detection.

In a short paper by Molla et al. [2015], the authors classified unvoiced speech as comprising both unvoiced speech and silence. So in the detection of what is effectively voiced speech they achieved a classification accuracy of 98% at 30 dB signal to noise ratio, which is not a significant improvement on Atal and Rabiners's results, and they do not address the problem of unvoiced/silence detection. A similar approach is taken by Upadhyay and Pachori [2015] where the authors used the same assumption that silence is unvoiced speech and achieved a similar order of performance. Both of these papers indicate a more general trend where the voiced-unvoiced-silence classification problem is reduced to one of voiced-unvoiced classification. Another example of this is the work by Kumar et al. [2015], where the authors segment the speech into voiced and unvoiced sections before their classification process. From Rabiner and Atal's work it is clear that the most difficult of the classification problems is that of separating silence and unvoiced speech, and papers such as those by Molla et al., Upadhyay and Pachori, and Kumar et al. don't contribute to this specific problem.

Deng and O'Shaughnessy [2007], describe a system based upon two binary classifiers, the first of which classifies each frame as either 'Voiced' or 'Unvoiced and Silence', and the second which reclassifies the 'Unvoiced and Silence' frames as either 'Unvoiced' or 'Silence'. With this system the authors achieve a voiced-unvoiced-silence classification accuracy of better than 91.15%; but the authors conclude that more, *"...Robust features..."* are required if the potential for their system is to be fully realized. Interestingly Deng and O'Shaughnessy provide a list of corrections they impose post classification to deal with very short voiced segments parenthesised by unvoiced segments, vice versa, and breath sounds, which they reclassify as silence if they are sustained for more than 90$ms$ before a voiced segment is encountered. Deng and O'Shaughnessy used the 'NTIMIT' Telephone speech corpus (Fisher et al. [1993]) in their evaluation.

Burileanu et al. [2000], described a system which makes an initial speech/not-speech decision based upon the ratio of the maximum energy throughout the speech versus the energy of each 15$ms$ slice of the speech, and which refined that decision—using trends in the zero crossing rate—in the vicinity of the potential silence to speech transitions previously detected. The authors found that 98% of the endpoints automatically detected were within 15$ms$ of manually measured values. Although, as the authors note, this method of silence detection is not robust to variations in noise, it is of interest because it develops the idea of using the ratio of energies. Sahoo and Patra [2014] extended the work of Burileanu et al. [2000] by using short term energy, zero crossing rate and the 'statistical' behaviour of the background noise. In their experiments on a system of speaker identification using the MFCC speech model Sahoo and Patra concluded that silence removal increased speaker identification rate by between 15 and 20%. The authors described a method of detecting silence which may improve the noise tolerance of the silence detection process; but because their aim was speaker identification their purpose was to remove both the silence and the unvoiced speech. Thus their basic aim was one of the detection of voiced speech; and Atal and Rabiner [1976] achieved an accuracy in this of better than 98% some 38 years earlier.

In pursuit of identifying sentence boundaries to facilitate the insertion of breaks into the stream of text output from an ASR process Anu and Karjigi [2014], employ a different set of speech features. The features are, pause duration, rhyme duration (where the speech slows towards the end of a sentence), and 4 pitch measures taken at inter-word boundaries — specifically slope, mean, maximum and minimum. With a Support Vector Machine (Cortes and Vapnik [1995]) trained as a 'non-probabilistic binary linear classifier', they achieved an accuracy of 81.176 %. This level of accuracy is low in comparison with the results of segmentation methods based upon the more deterministic features of speech; but Anu and Karjigi's approach may be of use in the sorting of already detected silences into categories such as stops, pauses and sentences.

The work by Mondal and Barman [2015] (previously discussed), brings to mind two other points. The first is on the technique of pre-emphasis, and the second — following tangentially from the selection of the 'low to full band energy' parameter — is on the optimal bandwidth for speech processing.

Pre-emphasis is a filtering technique and its use for ASR is formalised in ETSI-ES-202-050 [2007]. In Section 5.4.3.2 and Figure 5.2.1 of Huang et al. [2001], the authors describe a pre-emphasis filter as a First Order Finite Impulse Response Filter configured as a high pass filter. So when might it be advantageous to use pre-emphasis? Vergin and O'Shaughnessy [1995] describe pre-emphasis as a technique for flattening the speech spectrum which will (inevitably) boost the noise at higher frequencies as well as the speech. Ergo, pre-emphasis is a form of distortion and although the technique has often been adopted for ASR, for the recognition of silence in speech there is no particular advantage in using pre-emphasis with either the $\mathrm{D}_{eterm}$Classifier or the $\mathrm{LogFB}_{dynamic}$.

The technique of using the ratio of the energies across a broad spectrum can be traced back to Hughes and Halle [1956], where the authors process speech with a bandwidth up to 10 kHz; whereas Qi et al. [2004] sample the speech at 8 kHz, which supports a bandwidth of only 4 kHz (Shannon [1949]). So what might be the optimal bandwidth for speech processing? The answer to this question at least for the LogFB$_{dynamic}$ is just the maximum bandwidth available; and the reason for this is that though transmission systems may require a reduction in speech bandwidth to maximise communication channel throughput (presumably as with Qi et al.), the discarded high frequency content contains useful information (as shown by Hughes and Halle). So there is an advantage in using high bandwidth speech when it is available, and when there is no penalty in so doing.

The parameters used by Qi et al., are those defined in the Recommendations in G.729 [2012] on ***"Transmission Systems and Media, Digital Systems and Networks"*** and 729E Annex B describes the recommendations for a Voice Activity Detection System (VAD) which with the specified 8 kHz sampling rate supports a maximum audio frequency of 4 kHz. G.729 [2012] specifies four 'difference parameters' which quantify the differences between adjacent short term frames (10$ms$), and four 'differential parameters' which quantify the difference between each parameter and its long term average. The difference parameters include, the full band energy, the low band energy, the line-spectral frequencies and the zero crossing rate. The differential parameters include the full band energy, the low band energy, and the zero crossing rate, plus the spectral distortion. With these parameters, Qi et al. achieved a voiced-unvoiced Vs silence classification accuracy of 94.6%, and a voiced Vs unvoiced classification accuracy of 98.2%. These results — obtained with manually labelled training data — are for the Chinese Language and for bandwidth limited speech, and so are not entirely comparable with those of Section 5 herein. Qi et al., having chosen a Support Vector Machine as their pattern matching engine went on to identify potential disadvantages with the 'traditional' neural network, and though they also provided an empirical comparison of the SVM with the neural network, it cannot be ascertained from their work whether the SVM provided a performance advantage. Others though, have compared different pattern matching techniques for various purposes. Manchanda et al. [2007] evaluated various pattern matching techniques including Decision Trees, SVMs, Genetic Algorithms and Neural networks in the context of data mining, and they found Random Forests to be marginally more accurate than Multilayer Perceptrons (ANNs), and both of these to be more accurate than SVMs. Caruana and Niculescu-Mizil [2006] in a comparison of learning algorithms, found both Random Forests and SVMs to be marginally more accurate than ANNs; and Elizalde and Friedland [2013] in a comparison of three audio speech detectors specifically for speech segmentation that were based upon Gaussian Mixture Models (GMMs), the Support Vector Machine (SVM) and the Artificial Neural Network (ANN), report that the ANN variant was faster and more accurate than both the GMM and SVM variants.

Brognaux and Drugman [2016] used Voice Activity Detection to determine the location of silence pauses with the aim of improving the silence 'model' used in a Forced Alignment, and with this method they increased the accuracy of a system purposed at determining phoneme boundaries. The VAD method used by Brognaux and Drugman was based upon work by Sohn et al. [1999]. It is interesting to observe that Brognaux and Drugman selected a Voice Activity Detector that was developed around 16 years earlier; though why they chose this particular VAD, is not stated. This work is of interest because it questions the 'received wisdom' that the problem of the segmentation of speech at word boundaries has been solved with the technique of forced alignment. Rather the accuracy of an HMM forced alignment can be improved if the algorithm is primed with an accurate model of the silence; and to achieve that, the location and duration of silence pauses must be known in advance.

The work of Sohn et al. [1999] also features in a comparison of three VAD systems. The performance of VAD, is often evaluated using specific noise categories, such as in-factory or in-car noise, white noise and babble noise. This is because VAD is often used in real-time applications such as silence suppression for speech communication from uncontrolled acoustic environments, where the three noise categories may routinely be encountered. Using this evaluation technique, and in a comparison of their own work on Voice Activity Detection, with two other VAD systems, by You et al. [2012], and Sohn et al. [1999], Teng and Jia [2012] found that for in-factory and white noise the performance of their system expressed as a ratio of $\frac{P_(detection)}{P_(falsealarm)}$ is greater than that achieved by both You et al.

[2012] and Sohn et al. [1999], whereas for babble noise they found the converse to be true. Brognaux and Drugman did not cite Teng and Jia's evaluation of the work of Sohn et al., but from the Receiver Operating Characteristics (ROC) — a plot of Detection Probability Vs False Alarm Probability, published by Teng and Jia, it would seem that for Babble Noise — which is probably the most applicable noise category for Brognaux and Drugman's work, the VAD design by Sohn et al. is the better choice.

In an earlier work by Beritelli et al. [2002], the authors found that the performance of the VADs degraded with increase of noise in the speech signal, and were sensitive to the language spoken (for Italian, French, English and German).

This sensitivity to the magnitude and type of noise, and language, is an illustration of the intrinsic difficulties of VAD; but VAD, Brognaux and Drugman's  atypical use notwithstanding, is purposed towards real time communications and is subject to different constraints. For example, in the evaluation of VAD by noise injection of babble, white, and in-car/in-factory noise there is no specific accommodation of transient noise, which is a real issue when processing pre-recorded speech.

In the work by Deekshitha et al. [2015] on speech segmentation, the authors used short term energy, voicing information, most dominant frequency, and a spectral flatness measure, as the input to a supervised Artificial Neural Network (ANN) Classifier, and they achieved an average accuracy of 88%. In their aim of segmenting the speech into broad phonetic classes, whist using an MFCC based HMM model of the Malayalam language, they found an improvement in classification for 5 of the broad phonetic classes. Deekshitha et al. used 75% of the available data for training of the neural network, with the remaining 25% used for testing. They state that it is not possible to differentiate between silence and unvoiced speech, based upon the energy in the waveform — an assertion that is worthy of further investigation.

As well as silence pauses, speech can also include short periods of silence associated with stops and unvoiced speech; i.e. silences that are not consciously perceived by the listener. Such silences in speech can be painstakingly identified in the speech waveform — but this places a limit on the quantity of speech that can be investigated. Ananthapadmanabha et al. [2014] describe the stop as being a silence or a, “Low level acoustic signal” followed by a, “burst or transient”. The question is, can this silence or near silence be automatically detected, bearing in mind that the silence associated with stops can be short — of the order of a few milliseconds only.

For voiced/unvoiced/silence classification Qi and Hunt [1993] used a set of 13 cepstral coefficients (derived from 12 LPCs and the energy prediction error), the zero crossing rate and a function of RMS energy as the input to an ANN Voiced-Unvoiced-Silence classifier. With a signal to noise ratio of 30dB, Qi and Hunt achieved a classification accuracy of between 90 and 95%. Interestingly they compared the classification performance for just the 13 cepstral coefficients with the cepstral coefficients plus the zero crossing rate and energy function and found the latter two parameters made a significant contribution — particularly at lower signal to noise ratios, to classification accuracy. This result corroborates earlier results by Sarma and Venugopal [1978]), and also supports the view of Johnson [2012] on the limitations of the LPC speech model.

Working with Neural Networks for speech endpoint detection, and with a parameter set based upon the parameters defined by Rabiner and Sambur [1975], Hussain et al. [2000] investigated two ANN topologies. They found that the performance of a Multi-Layer Perceptron (MLP) network to be, “Slightly more accurate” than an ADALINE (Adaptive Linear) network, in that the ADALINE network showed a trend towards early endpoint detection. The authors state that this was because the ADALINE network — a linear classifier — classified unvoiced speech as noise. This is an encouraging result because it shows that the classes of speech and silence are not entirely linearly separable, and that the MLP — a non-linear classifier — does provide a better separation of the classes. Hussain et al. [2000] also compared the performance of both networks with the classifier defined by Rabiner and Sambur [1975], and found that the latter’s classifier was better at identifying speech endpoints; although for the endpoints detected, that the MLP provided the most accurate estimation of location. They concluded their work with the cautionary note that, “In short, further research should be performed to improve and adopt the MLP technique for endpoint detection” (sic).

Toledano [2000], improved the identification of speech/silence boundaries for the automatic segmentation of speech by replacing the fuzzy logic in his existing HMM ASR fuzzy logic system with a 3 layer ANN and found an improvement in performance. The speech model used by Toledano comprised parameters, based upon mean energy, zero-crossing rate and mean frequency in two windows either side of the point under consideration, and the differences between each of these features in the two windows, and correlations between the two windows including the mean spectrogram and the energy contour, plus the results of dip-detection on the mean spectrogram. The model also included deltas between the point under consideration and the adjacent time marks from the HMM ASR process. Although the ANN based system performed better than the fuzzy logic system, the author concluded that more work was needed; and noted that the test corpora was limited to a single speaker. This work mostly confirms that for this case at least, a multi-layer perceptron Artificial Neural Network is a better and more manageable non-linear classifier than the fuzzy-logic system it replaced.

The work of Oprea and Şchiopu [2012], Palaz et al. [2015], and Wei and Yanpu [2005], though not directly concerned with silence recognition in speech, is of interest:-

For an isolated word recognition system, Oprea and Şchiopu [2012] adopted a speech model based upon the Linear Prediction Cepstrum, and report a word recognition rate of between 61% and 76% for Female Speakers, and between 53% and 74% for Male Speakers. The authors do not address the question of why the system performance is better for female speakers. The word recognition rates serve only to provide a further indication of the limitations of the LPC model.

In a departure from the techniques so far described, Palaz et al. [2015] used raw speech as the input to a convolutional neural network (as described by LeCun and Bengio [1995]) for learning linearly separable features for ASR, and report that their system yielded, “similar or better performance” than Multi Layer Perceptron (MLP) Systems using cepstral based features. This is of interest because it illustrates that it is no longer necessary to work with reductionist models — such as the MFCC model — when processing speech. That said, raw speech might not be the optimal input because if the effects of masking are not taken into account, then the raw data must include redundant information. (The inevitable conclusion of this line of reasoning is that when the purpose is the identification of silence in speech, the input to the ANN should be in accord with the psychoacoustics model of speech.)

Wei and Yanpu [2005] used instantaneous spectral components with the greatest signal to noise ratio for speech enhancement; although they took no cognizance of masking or Bark bands. They found that by dynamically adjusting the speech/silence threshold, and reconstructing the speech from those harmonics above the threshold that the resulting speech was objectively and subjectively of, “Surprising quality”. This work leverages to an extent upon work by Boll [1979] on noise suppression in speech by harmonic subtraction. Boll dynamically obtained the spectrum of the noise during periods of ‘silence’, and subtracted this from the following speech — presuming that the noise during silence is additive noise which is also present in the speech. With normal speech recordings the noise is unpredictable and so Boll’s method may have only limited success; but the test environment used by Boll included the noise of a helicopter — a periodic non-sinusoidal noise that is particularly suitable for both Fourier analysis and the spectral component selection method. Of particular interest here are the techniques of spectral subtraction and waveform reconstruction.

## 3.3 Psychoacoustics Models of Speech

From a mechanistic viewpoint, the production of speech is very well understood and the operation of the human hearing transducer — the cochlea — is also to some extent understood. What is not understood is how we process the acoustic wave-front that we perceive as speech.

'Fechner's Law' (Gustav Theodor Fechner 1801-1887) states that the perceived effect is proportional to the logarithm of the physical stimulus; but Scheerer [1987] in an introduction to Fechner's work writes that Fechner, a psychologist, actually limited the scope of the law to the mind-brain relation. Whatever Fechner meant, Fletcher and Munson [1933], later confirmed that the human response to changes in sound magnitude is approximately logarithmic; and Stevens et al. [1937] and later Stevens and Volkmann [1940] in deriving the 'Mel' scale, found that the human response to changes in frequency becomes increasingly logarithmic as the frequency increases (Figure 3.1).

**The Mel Auditory Scale:** Stevens and Volkmann [1940] empirically derived the Mel scale as follows. A 'keyboard' was set up, which enabled five frequencies comprising fixed upper and fixed lower frequencies, two variable intermediate frequencies, and a variable centre frequency; and the test subjects were then tasked with adjusting separately the three variable frequencies using pitch bisections. Firstly they were to adjust the centre frequency until they perceived it to be midway between the fixed upper and fixed lower frequencies; and then they were to repeat the process for both the intermediate frequency between the fixed upper and the centre frequency and for the intermediate frequency between the fixed lower and the centre frequency. With this method Stevens and Volkmann were able to establish an approximation of the 'law' which relates perceived changes in pitch, to actual changes in frequency. The authors arbitrarily set the datum that 1,000 Mel should equal 1,000 Hz, and thereafter the application of the law — later to be expressed as a simple equation — resulted in the auditory Mel scale.

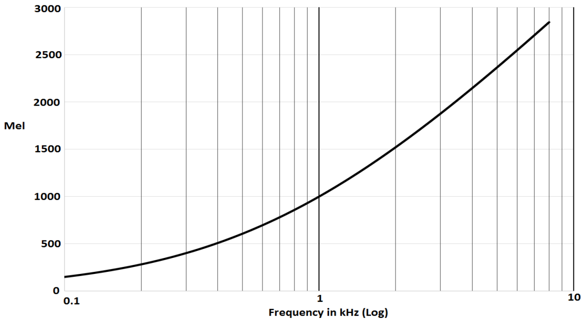


Figure 3.1: Mel versus Frequency

An approximation of the relationship between Mel and Frequency graphed using Equ. (8.4) — Stern and Morgan [2013]; attributed by Stern and Morgan to O'Shaughnessy [1999]. This graph is similar to Figure 2 — Stevens and Volkmann [1940]; which supersedes Stevens et al. [1937].

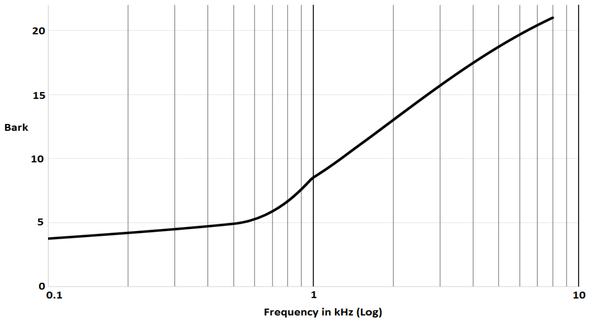


Figure 3.2: Bark versus Frequency

An approximation of the relationship between Bark and Frequency graphed from Equ. 6 — Traunmüller [1990].

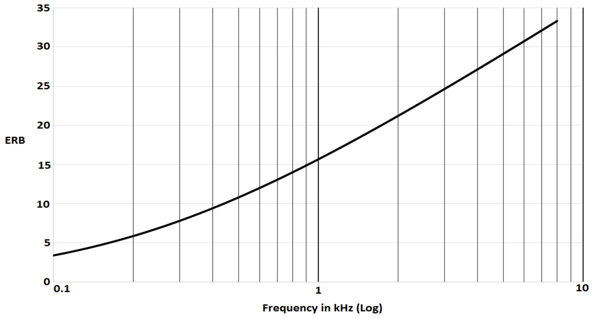


Figure 3.3: ERB Rate versus Frequency

An approximation of the relationship between ERB and Frequency graphed from Equ. (8.5) — Stern and Morgan [2013]. The graph is similar to Figure 2 — Moore and Glasberg [1983].

**Critical Bands and the Bark Auditory Scale:** Zwicker et al. [1957], with the aim establishing that the loudness within a critical band is constant and independent of the spacing of the tones, but increases when the spread of frequencies exceeds a critical value, derived the following experiment. Four tones of equal intensity were mixed, and the resulting 'complex' signal was played alternately with a pure tone to the test subject. The duration of each of the test signals was about 1 second, and they were separated in time by about half a second. The test Subject was required to adjust the loudness level of the complex signal until it was perceived as being equal to the loudness level of the pure tone; and finally the Subject was to refine the accuracy of their adjustment by increasing and decreasing the loudness of the complex signal about their best adjustment so far, to obtain their final adjustment. Thereafter the test was repeated, but this time it was the loudness of the pure tone that was varied. The authors repeated their procedure with different frequency spacings of the tones which constituted the complex signal, about centre frequencies of 500 Hz, 1kHz and 2kHz. The results obtained by Zwicker et al. confirmed their hypothesis. In the same work Zwicker et al. related the work of Fletcher [1940] who sought to determine 'position coordinates' on the basilar membrane — that part of the ear which effects the spectrum analysis — to their own work and that of three other researchers. On the subject of masking and the critical band, Fletcher wrote,

> **"For this type of noise the critical band width in cycles is numerically equal to the ratio of the intensity of the tone masked to the average intensity per cycle of the noise producing the masking. Regardless of where the band is located we will see later that these critical widths always correspond to a single element of length on the basilar membrane, namely $\frac{1}{2}$ mm".**

Zwicker et al. argued that the critical band defined by Fletcher is approximately proportional to the four sets of results achieved by different researchers using different techniques which involved Thresholds, Masking, Phase and — for their own experiment described above — Loudness Summation. The authors used the phrase 'approximately proportional' because the critical bandwidth derived in the four experiments is about 2.5 times the width of the critical bandwidth derived, "From the assumptions made by Fletcher". Assuming a notional length of the basilar membrane of around 32 mm or so (Koch et al. [2017]), then this suggests the number of critical bands is equal to $\frac{32.0}{0.5 \times 2.5} = 25.6$ critical bands. That is, the typical length of the basilar membrane divided by the product of Fletcher's critical bandwidth and Zwicker's constant of proportionality. It must be observed that Rask-Andersen et al. [2012] citing Hardy [1938] states that the organ of corti length (and hence the basilar membrane) can vary by as much as 10 mm (also Koch et al. [2017]); and according to Miller [2007], "there may be a small difference in the lengths of male and female human cochleas even though statistical analyses of the data are not decisive". What is important here is not the length of the typical basilar membrane or the absolute value of the constant of proportionality, but that the proportionality exists. Zwicker et al. also observed from their results and the results of the three other experiments that the critical band seemed to be similar to other auditory measures such as the Mel scale insofar as the number of Mels in the critical band is constant over most of the audible frequency range, and discussed the possibility that both equal Mel frequency intervals and critical bands may correspond to equal distance along the basilar membrane (the concept of critical bands having equal distances along the basilar membrane as expressed earlier by Fletcher). Zwicker [1961] then proposed the Bark scale as a useful subdivision of the audible frequency range into critical bands as an approximation of the manner in which the ear seems to carry out the subdivision process.

Although the Mel and Bark scales were derived by different empirical processes: they have a similar characteristic ( see Figures 3.1 and 3.2 ). The experiments on deriving the Bark scale are of particular relevance because of the implicit description of masking and the *'Critical Band'*.

**Equivalent Rectangular Bandwidth** Moore and Glasberg [1983] discuss experiments which culminate in yet another auditory scale, the Equivalent Rectangular Bandwidth (ERB) Scale, which is an approximation of the auditory filter, where the filters are approximated by rectangular passband filters. Figure 3.3 shows that the ERB scale has a similar characteristic to the Mel and the Bark scales, except that the filter bandwidth continues to decrease as the frequency at the centre of the ERB decreases below 500 Hz.

In summary, Zwicker et al. [1957] amongst others showed critical bandwidth to be a function of frequency (at higher frequencies the critical bandwidth is greater), and tabulated a non-overlapping fixed arrangement of *'Bark bands'* which for frequencies below approximately 500Hz to 1kHz are spaced at linear intervals and for frequencies above this range at logarithmic frequency intervals (see Figure 3.2 and Appendix A — Bark Band Ada Specification). Zwicker [1961] wrote that the critical bands are not fixed on the frequency scale, and also of a close correlation of the bands with the mechanical structure of the cochlea. An analogy for this is given by Scharf [1970], who describes a set of band-pass filters with variable centre frequencies. Scharf, citing von Békésy [1970] goes on to speculate on whether Lateral Inhibition (see Figure 3.4) is involved in the mechanisms of critical bands, and suggests other processes may be involved.

**Lateral Inhibition** (Figure 3.4) is described by von Békésy [1970] in terms of reaction to stimulus on the surface of the skin, where a point of stimulus (A in Figure 3.4) causes a ring of inhibition to other stimulus around that point; but if two stimuli are close together they will reinforce each other (if the stimuli are inside the ring of inhibition). Von Békésy's key points are that this results in heightened sensation at the stimulus point, and that with multiple simultaneous stimuli, the pattern of sensation may be different to the pattern of the stimuli. Von Békésy also pointed out that the mechanics of the cochlea are complicated, and this implies that the lateral inhibition model — though useful for Von Békésy's purpose of making an enlarged mechanical model of the cochlea using the response of the surface of the skin to stimuli — is not an adequate psychoacoustics model. That said, it is interesting that the idea of lateral inhibition in some form or other can be used to describe operational modalities of speech, vision, and touch (von Békésy [1970]).

**The 'Mexican Hat' filter :** In Figure 3.4 it will be seen that if the peak at A is considered to be the centre frequency of a filter, with increasing frequency on the x axis to the right of A and decreasing frequency on the x-axis to the left of A, then this describes the ‘Mexican Hat’ filter. Park and Lee [2003] used this filter shape in preference to the triangular filters shown in Figure 3.5 to include Lateral Inhibition in their filter-bank.

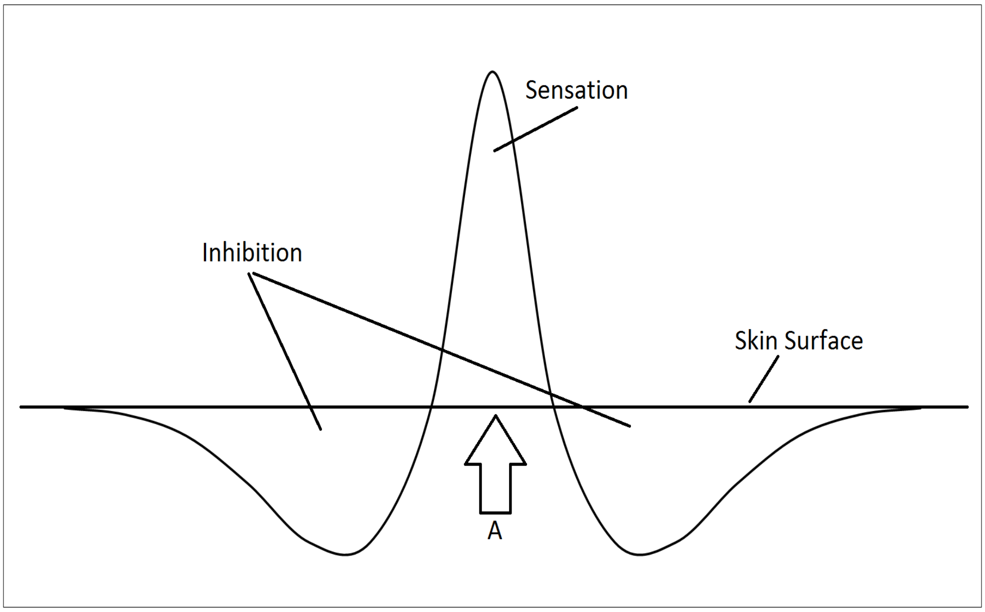


Figure 3.4: Lateral Inhibition and the ‘Mexican Hat’ Filter

This illustration is based upon Figure 19 (Page 325), of von Békésy [1970].
The Mexican Hat takes mathematical form as a type of Wavelet; specifically, the second derivative of a Gaussian (Addison [2002]). The curve above was constructed using the formula for the Mexican Hat Wavelet given on page 7 of Addison [2002].

**The Cochlea Amplifier :** Though evidence suggests that the processing of sounds in the brain is tonotopic — where ascending bands of frequencies are seemingly processed in adjacent locations in the brain (Moerel et al. [2012], Langers et al. [2014], Langers [2014]), there seems to be no physiological evidence of the Bark bands in the cochlea or neurological arrangement of the ear. That said, it is generally accepted that the ear boosts the most powerful harmonics in the spectrum though the mechanisms or processes for this are not understood (Davis [1983], Ashmore and Kolston [1994], Rask-Andersen et al. [2012], Fridberger et al. [2006], Dong and Olson [2013]).

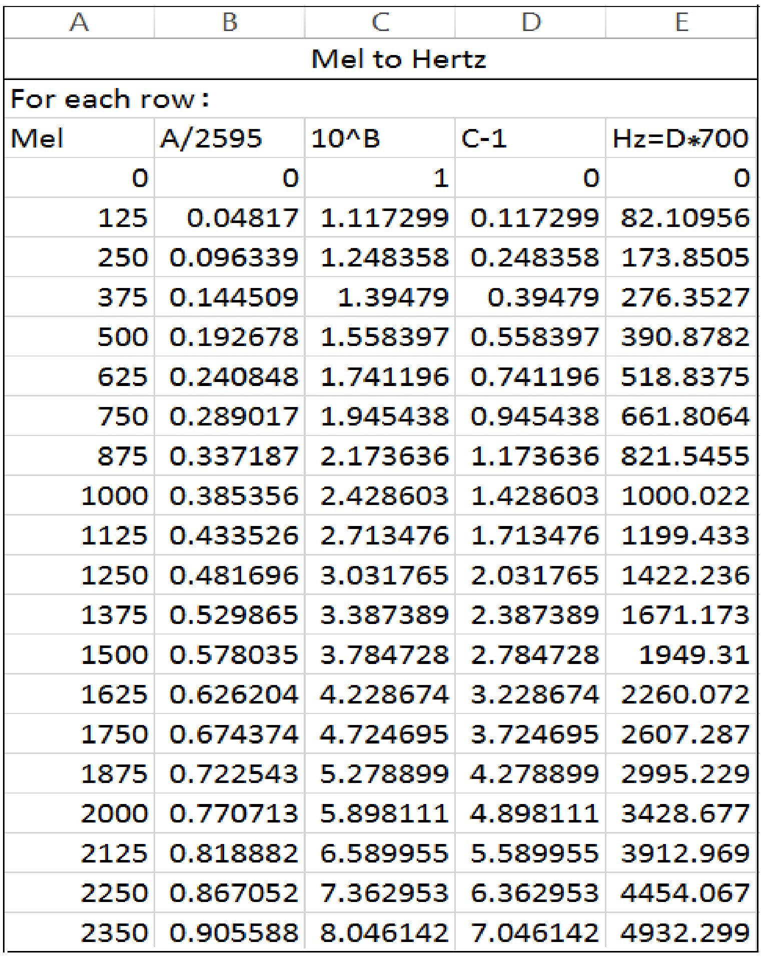
Davis [1983] describes a model of cochlea mechanics, but states that the mechanism for cochlea amplification is not understood. Ashmore and Kolston

[1994] discusses how the motion of the basilar membrane (Crouch [1981]) is not that which might be expected from a passive structure, and that the outer hair cells in the organ of Corti (Crouch [1981], Rask-Andersen et al. [2012]) include motor cells which can rapidly generate forces. Fridberger et al. [2006] discusses how the sensitivity to sound can be increased by "1,000–fold" (sic) by the outer hair cells, and Dong and Olson [2013] state that, "The cochlear amplifier has inspired scientists since its discovery in the 1970s, and is still not well understood."

Accepting then the reality of cochlea amplification and critical bands, the amplification of the most powerful harmonic within each critical band must coincide with the masking (both simultaneous and temporal) of other nearby harmonics. The conventional log filter-bank does not emulate this functionality, yet by using the most powerful harmonic in each of the critical bands, rather than the notional centre Bark frequency, an improved (psychoacoustics) model of speech will result - which will manifest itself in an acceptable speaker specific speech model.

The cochlea is the acoustic to neurological transducer and even though a wide variation exists in shape and dimensions (Rask-Andersen et al. [2012]), each cochlea is constant in its functionality (Fletcher [1940]). That the processing of sound in the brain may be tonotopic Moerel et al. [2012] Langers et al. [2014] Langers [2014] is suggestive that connections between the cochlea and the brain might also be grouped into frequency bands; and if the positive feedback mechanism does exist (Dong and Olson [2013]) then perhaps the masking is a side effect of the feedback mechanism. That said, it is not obvious from the tonotopic activity maps of the brain that the tonotopic relationship exists to the granularity of the Bark bands. Nor is there evidence in the literature on the exact mechanisms of cochlea amplification, which is not fully understood (Dong and Olson [2013]), or of any neurological correlation between the cochlea and the Bark Band frequencies. So, although there is no proof of a correlation between the tonotopic processing of sound in the brain and the Bark acoustic scale, there is the suggestion that assuming just such a correlation is not unreasonable; and it may be appropriate to ask whether a model which conforms to the Bark scale and comprises the minimum number of critical bands, with dynamic masking arranged around the most powerful harmonic in each band, and with the filter bandwidth limited to the upper and lower cut-off frequencies in each band, can provide an accurate representation of speech.

**The Mel Frequency Cepstral Coefficients Model of Speech :** Some of the psychoacoustics effects previously described were incorporated into mainstream ASR, following the work of Davis and Mermelstein [1980]. Building upon earlier work by White and Neely [1976], Davis and Mermelstein devised the 'Mel Frequency Cepstral Coefficients' (MFCC) speech representation which incorporated log magnitudes and the log filter-bank (see Table 3.1 & Figure 3.5) — both important characteristics of human hearing; thus providing an improved model of speech. The MFCCs model was static with fixed log filter-bank centre frequencies, presumably to facilitate generic speech templates, and accordingly could not include dynamic masking. According to Jurafsky and Martin [2009] the MFCCs separate the source from the filter. That is to say the shape of the vocal tract — the filter, is considered to be more important for distinguishing between different phones (phone: a single speech sound — Chambers [1999]) than information about the glottal source ($F_0$ for example); and this separation is achieved by using only the first 12 Cepstral values. The Cepstral values are more a model of speech production, than of speech perception, and it is necessary to add the energy coefficient, to exploit the correlation between the energy and the phone — for example voiced speech often has more energy than unvoiced speech. The MFCC model is completed by adding the velocities and the accelerations for each of the 13 features; thereby resulting in an MFCC model with 39 features.



| A | B | C | D | E |
|---|---|---|---|---|
| Mel to Hertz | | | | |
| For each row : | | | | |
| Mel | A/2595 | 10^B | C-1 | Hz=D*700 |
| 0 | 0 | 1 | 0 | 0 |
| 125 | 0.04817 | 1.117299 | 0.117299 | 82.10956 |
| 250 | 0.096339 | 1.248358 | 0.248358 | 173.8505 |
| 375 | 0.144509 | 1.39479 | 0.39479 | 276.3527 |
| 500 | 0.192678 | 1.558397 | 0.558397 | 390.8782 |
| 625 | 0.240848 | 1.741196 | 0.741196 | 518.8375 |
| 750 | 0.289017 | 1.945438 | 0.945438 | 661.8064 |
| 875 | 0.337187 | 2.173636 | 1.173636 | 821.5455 |
| 1000 | 0.385356 | 2.428603 | 1.428603 | 1000.022 |
| 1125 | 0.433526 | 2.713476 | 1.713476 | 1199.433 |
| 1250 | 0.481696 | 3.031765 | 2.031765 | 1422.236 |
| 1375 | 0.529865 | 3.387389 | 2.387389 | 1671.173 |
| 1500 | 0.578035 | 3.784728 | 2.784728 | 1949.31 |
| 1625 | 0.626204 | 4.228674 | 3.228674 | 2260.072 |
| 1750 | 0.674374 | 4.724695 | 3.724695 | 2607.287 |
| 1875 | 0.722543 | 5.278899 | 4.278899 | 2995.229 |
| 2000 | 0.770713 | 5.898111 | 4.898111 | 3428.677 |
| 2125 | 0.818882 | 6.589955 | 5.589955 | 3912.969 |
| 2250 | 0.867052 | 7.362953 | 6.362953 | 4454.067 |
| 2350 | 0.905588 | 8.046142 | 7.046142 | 4932.299 |

Column A tabulates equal Mel frequency intervals which are converted to Hertz in Column E using the equation for Figure 3.1. Each set of three adjacent frequencies in Column E provides the lower cut-off, the centre, and the upper cut-off frequencies for one of the 18 overlapping passband filters shown in Figure 3.5.

Table 3.1: Mel to Hertz Conversion.

Davis and Mermelstein [1980] limited the bandwidth of the speech in their experiments by using a low pass filter at 5 kHz and a sampling rate of 10,000 samples per second, and when contrasting the MFCC representation with other representations wrote,

> ***"Specifically, MFCC allow better suppression of insignificant spectral variation in the higher bands". — Davis and Mermelstein [1980]***

The decision to limit the audio bandwidth during their experiments to 5 kHz—without pre-emphasis is interesting. White and Neely [1976] in their earlier version of the filter-bank used 20 bandpass filters which in total spanned a bandwidth of 100 Hz to 10 kHz; and it is known (Hughes and Halle [1956] and others) that fricative consonants in normal speech have significant energy levels at

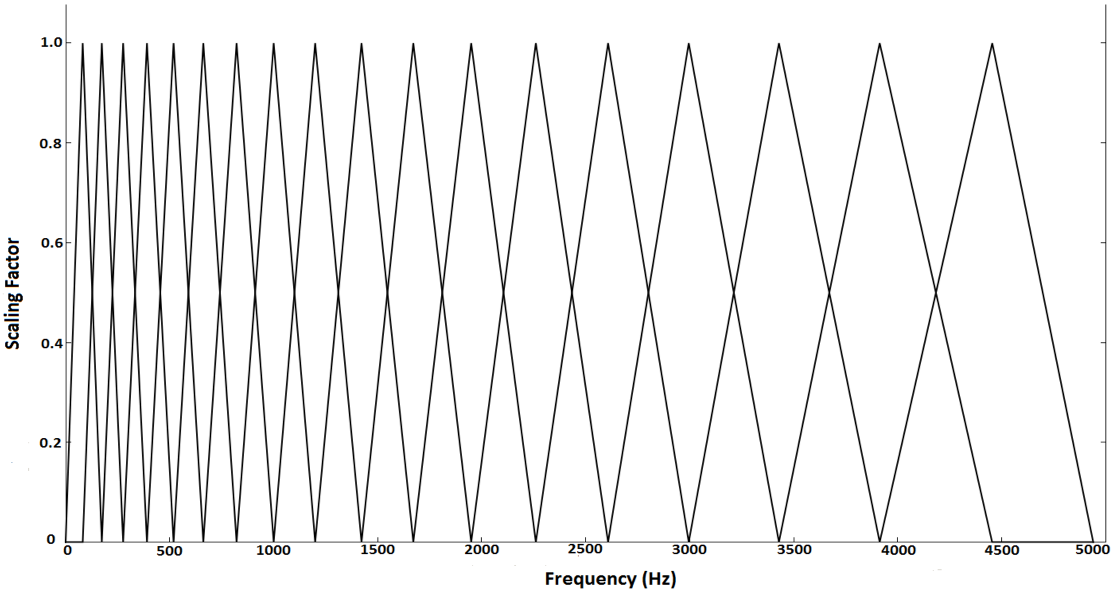


Figure 3.5: The Filter-Bank (Frequencies in Hertz)

The Filter-Bank with 18 separate overlapping filters spanning a frequency range up to a maximum frequency of 4,932 Hz. This is plotted from the frequencies calculated in Table 3.1 and shows less filters, but in other respects is similar to Fig 1: *"Filters for generating Mel-frequency cepstrum coefficients"* (Davis and Mermelstein [1980]).
In operation, the magnitude harmonics for the complete audio spectrum (as generated using the DFT) are scaled by each of the triangular filter shown — resulting in 18 overlapping bands of scaled harmonics.

linear frequencies up to and beyond 8 kHz — that is up to and beyond Bark band 22.

Davis and Mermelstein also experimented with stepping their analysis frame of 25.6$ms$ duration, by either 6.4$ms$ or 12.8$ms$; and found the speech recognition rate improved by an average of 1.7% with the shorter step.

Around the time Davis and Mermelstein published their work, Schroeder et al. [1979] and Krasner [1980] published work which described how the speech waveform includes more information than we perceive (because of auditory masking); Krasner [1980] seeking to reduce the bit-rate necessary for the transmission of speech of a given quality, and Schroeder et al. [1979] seeking to make the encoding noise (quantisation noise) imperceptible for higher bit-rates, and to reduce the encoding noise at lower bit-rates. A later encoding system which removes information from the encoded audio which, because of the masking effect, would not be perceived by the majority of listeners, is the MP3 Audio Lossy Encoder (MP3-Standard [1995], Brandenburg [1999]).

It is worth pausing to consider the accuracy of the Mel and Bark scales and the MFCC model of speech representation. Stevens [1957] in an evaluation of hysteresis in pitch (perceived frequency) — which is where the same pitch is perceived to be higher for ascending equal pitch intervals, and lower for descending equal pitch intervals — wrote, ***"All in all then, the evidence for hysteresis in pitch bisections is ambiguous"***. Greenwood [1997], in a re-evaluation of his earlier work on hysteresis for S.S. Stevens suggests that the evidence was not ambiguous; and more, stated that the Mel scale did not coincide with equal distances on the cochlea whereas the ascending and descending measures did.

Greenwood's view was supported by Thompson et al. [2012]. They not only confirmed hysteresis, they also found that other effects contributed to the level of hysteresis such as the size of the interval (e.g. 7 semitones versus 6 semitones), and whether the tone increased or decreased in intensity across the two pitches. Greenwood and later Thompson et al., showed that the Mel frequency scale is too simple to represent the response of the ear.

Rudnicki et al. [2015] compared three more-contemporary models together with the MFCC speech model in a Hidden Markov Model (HMM) Automatic Speech Recognition (ASR) engine and found that two of the three models outperformed the MFCC model (all at around 75 % recognition) by one or two percent. Park and Lee [2003] described a log filter-bank model which accommodated both simultaneous and temporal forward masking in the critical bands; and found the performance to be better than that of the conventional MFCC filter-bank; particularly under noisy conditions. To implement simultaneous masking, Park and Lee used a 'Mexican Hat' convolution filter (See Figure 3.4) to reinforce the dominant signal harmonics and to suppress the adjacent harmonics. Park and Lee [2003] did not include in their schema dynamic tracking of the most powerful harmonic in each of the critical bands. They found that simultaneous masking was more effective in reducing classification errors, than temporal masking. (Johnson [2012] writes on forward temporal masking that the effect is of little significance at the sound pressure levels which occur in normal speech, and that the effect lasts only of the order of $25ms$ or so; and comparatively, that backward temporal masking is of even less significance .) Dai and Soon [2013] also used the 'Mexican

Hat' filter — in this case a piecewise linear approximation — to integrate lateral inhibition into an MFCC feature extraction front-end. They also integrated temporal spectral averaging, forward masking and cepstral mean normalisation into their model. Dai and Soon [2013] conclude that their algorithm improves the speech recognition of an HMM system under noisy conditions.

Zhu and 0'Shaughnessy [2004] whilst seeking to improve the MFCC model for ASR introduced an emulation of simultaneous dynamic masking by pre-processing the frequency spectrum, before calculating the MFCCs. The authors selected a triangular filter about each harmonic and where the magnitude of adjacent harmonics was found to be lower than the instantaneous magnitude indicated by the triangular filter then those harmonics were suppressed. For their masking model, Zhu and O'Shaughnessy empirically derived the optimal upper and lower slopes for the filter, and also experimented using a logarithmic asymmetrical frequency representation of masking (using interpolation based upon 23 Mel Scale triangular filters). The authors found that the latter of these two models provided the greater improvement.

Montalvão and Araujo [2012] also introduced dynamic masking into an MFCC model for speaker verification, describing their implementation, "*as a sliding window (instead of fixed windows) from which energy peaks are taken, and all remaining spectral energy is discarded (masked) for each position of the sliding window*". In a comparison of this system with their baseline MFCC model which did not include the masking technique, the authors found that under strong noise conditions, a performance improvement was obtained.

**The Filter-bank and Deep Neural Networks :** It is interesting to consider the contribution to the log filter-bank debate from contemporary works on ANN ASR. Xiong et al. [2016] use a mel filter-bank comprising 40 filters. Unfortunately the authors do not provide a rationale for this choice, but does their choice imply that the MFCCs 39 feature speech model is either inadequate, or unsuitable for use with Deep Neural Networks? Contrast the work of Xiong et al. [2016] with that of Sainath et al. [2015] who also used a log filter-bank comprising 40 filters for comparative purposes, though they contend that for lower Word Error Rate

(WER) when using statistical modelling, that the log filter-bank is not, "guaranteed" to be the best option; and they suggest that using the raw speech might be the better choice for ASR. The first of these papers described activities to optimise a speech recognition system that used convolutional and recursive neural networks; and the second described a speech recognition system that learned from raw speech, and which used Convolutional Long Short-term Memory Deep Neural Networks (CLDNN). As both of these papers were published by companies, it is not known whether they were subjected to independent scrutiny. Even so, that separate organisations use a similar log filter-bank arrangement when working with Neural Networks is indicative of the current trend; which is one of incorporating more of the detail of speech into the model.

## 3.4 Summary

Work in the field of psychoacoustics (summarised in Fastl and Zwicker [2007]) led to the development of the Mel Frequency Cepstral Coefficients (MFCC) speech model (Davis and Mermelstein [1980]), which encapsulated some of the modalities of hearing. At the heart of the MFCC model is the conventional logarithmic filter-bank which is a set of passband filters (Figure 3.5) that are spaced at equal Mel intervals. Potential disadvantages with the log filter-bank are threefold: the filters overlap, are of fixed frequency, and the filter arrangement is not necessarily compliant with those observed in psychoacoustics experiments (Zwicker et al. [1957]).

The results of research into the theory and practice of the encoding of speech based upon the psychoacoustics model (Schroeder et al. [1979]) (Krasner [1980]) also emerged around the same date, when both Schroeder et al. and Krasner employed dynamic masking to remove noise from the speech. Their models were purposed to include only those parts of the speech we perceive; whereas Davis and Mermelstein's model was purposed towards a generic decoding of speech. Both of these approaches were of importance: in the fields of MP3 Audio Compression

(MP3-Standard [1995]) and Automatic Speech Recognition respectively. A comparison of the models of Schroeder et al. and Krasner with that of Davis and Mermelstein showed that the former's models — insofar as they were dynamic and included masking — incorporated more of the modalities of perception than did Davis and Mermelstein's.

Though Park and Lee [2003] improved the MFCC model by introducing enhanced masking, and both Zhu and 0'Shaughnessy [2004] and Montalvão and Araujo [2012] further improved the model by introducing dynamic tracking of the spectral peaks, and masking around those peaks; in recent times the MFCC model for ASR has lost favour. For example, researchers such as Xiong et al. [2016] and Sainath et al. [2015] are working with ASR which uses Deep Neural Networks, where the input to the DNN is either the raw speech data, or speech data that is filtered through a 40 element Mel filter-bank.

It is possible to conclude from the research that the difficulty in recognising silence in speech is mostly with differentiating between unvoiced speech and the noise during silence, that speech and silence are not linearly separable and that the recognition of silence in speech is a pattern recognition problem. Often the noise during silence is not predictable and is not Markovian in nature. So the pattern recognition technique must recognise the characteristics of speech, and then anything in the audio without these characteristics, inevitably must be silence.

It is self evident that a dynamic model of speech which tracks the spectral peaks is necessary to implement simultaneous masking. The reality of masking is proven, and it is fair to question whether dispensing with masking and working with the raw waveform is the best approach; as implicit in that approach is that there is more valid information carried in the speech than we can perceive.

From the narrative in this review, a strategy has evolved for applying psychoacoustics to the problem of the detection of silence in speech. This strategy is further defined in the next section, and experiments to test the hypotheses are described.

# Chapter 4: Research Method

Audacity® [2014] Version 2.0.6 audio recording and editing, open-source software was used to generate the waveforms and spectra for this work.

## 4.1 Overview

Subsidiary questions resulting from the literature review are:

1. What is the best parameter set for articulating the differences between silence and speech?

2. What is the most suitable window type and duration for the Discrete Fourier Analysis?

3. Is it feasible to accommodate the phase component of the DFT as well as the magnitude component in the speech model?

4. Is it possible to differentiate between silence and unvoiced speech based upon the energy in the waveform?

One can conclude from the research on sound perception that considerable evidence exists that shows the effects of masking and the 'existence' of critical bands (also see Appendix B: "An Experiment with Simultaneous Masking"); and my contention is that the fidelity of the log filter-bank as a device to model speech, can be significantly improved by:

(a) Substituting each of the log filter-bank filter centre frequencies with the dynamically derived most powerful harmonic for each of the Bark bands, and with simultaneous masking arranged around these harmonics;

(b) Returning the filtered Bark bands to the time domain using waveform reconstruction;

(c) Dividing this Bark Band representation of speech into time slices.

The advantages of this include:

1. The dynamic tracking of the most powerful harmonics together with the simulated masking about these harmonics will result in an accurate speech model.

2. As the model is entirely in the time domain, any relationship between the Fourier Transform window duration and the model time-slice duration is completely eliminated. Thus it is possible to select any time-slice duration for the speech model.

3. As each band in the model is individually reconstructed, then the model includes both magnitude and phase information.

To automate the detection of silence in speech my approach is to use a **Deterministic Silence/Speech Binary Classifier (D$_{eterm}$Classifier)** to achieve an initial categorisation of the audio, and then to repeat the classification using a **DNN_Classifier** that is trained from the output of the **D$_{eterm}$Classifier**. Of the various non-linear pattern classifiers available, the DNN was chosen because it most complies with the human interface paradigm which underpins this work. That is, whereas the Support Vector Machine is a development of co-ordinate geometry (Cortes and Vapnik [1995]) and the Random Forest a development of binary trees (Ho [1995]), the DNN is a development of neuroscience (McCulloch and Pitts [1990](reprinted from 1943), and Rosenblatt [1958]).

## 4.2 Test-Cases

The experiments will involve a detailed examination of two modes of operation:

- D$_{eterm}$Classifier automated silence detection versus the ground-truth.
- DNN_Classifier automated silence detection versus the ground-truth.

The chosen texts were all read aloud by the same male individual and recorded in monaural in a home environment using the internal microphone of a DP004 digital recorder (Tascam™ [2017]). The recordings were in the 'lossless' WAVE Format at 44100 samples per second (sps); so ensuring more than sufficient bandwidth for speech reproduction. In fact with this format the noise bandwidth considerably exceeds the signal bandwidth; and though this is not desirable, it is acceptable in this case because the audio is subsequently filtered using the techniques identified herein. Each of the test-cases was recorded in a single take. That is, any sections of speech with errors were immediately repeated, and were removed, post recording, with the Audacity® [2014] audio editor.

The signal to noise ratios for the recordings (Table 4.1) were obtained by measuring the RMS magnitude of the noise during silence, and the RMS magnitude of the signal during the speech for 45 contiguous seconds of the audio commencing after the first minute — using the silence and speech segments identified by the D$_{eterm}$Classifier — and applying the equation (Ivison [1978] pg 70) :

$$\text{Noise(dB)} = 20log_{10}\frac{RMS\ Speech}{RMS\ Silence}$$

The test-cases were all of about 12 minutes duration, and these were sectioned into Training, Validate, and Test Data Sets (Table 4.2) to support Split Sample Testing (Section 7.3.1 of Priddy and Keller [2005]). For each test-case the Training Data Set was used only for supervised training, the Validate Data Set was used to assess when training was complete (to prevent over-fitting); and the Test Data Set was used to assess the quality of the trained DNN_Classifier. Here results were recorded for the Test Data Sets only; but in practice all of the original speech data

— comprising the Training, Test and Validate Data Sets concatenated in the correct order — may be processed by the trained neural net.

| **Test-Case** | **Train S/N Ratio** | **Validate S/N Ratio** | **Test S/N Ratio** | **Average S/N Ratio** |
|---|---|---|---|---|
| TC1 | 33.395 dB | 32.18 dB | 30.825 dB | 32.133 dB |
| TC2 | 34.441 dB | 31.4 dB | 32.231 dB | 32.691 dB |
| TC3 | 28.047 dB | 29.042 dB | 33.418 dB | 30.169 dB |
| TC4 | 28.614 dB | 30.823 dB | 31.708 dB | 30.224 dB |
| TC5 | 31.248 dB | 35.372 dB | 32.584 dB | 33.068 dB |
| TC6 | 27.263 dB | 32.762 dB | 32.345 dB | 30.790 dB |

The signal to noise ratio was automatically calculated using the speech/silence classification data provided by the D$_{eterm}$Classifier. Variation exists between the train, validate, and test data S/N ratios because the data sets were processed separately by the D$_{eterm}$Classifier.

Table 4.1: Measured Signal to Noise Ratios for all Test-Cases

| **Test-Case** | **Train** | **Validate** | **Test** |
|---|---|---|---|
| TC1 | Pt1–255.6 | Pt3–325.47 | Pt2–294.133 |
| TC2 | Pt2–262.741 | Pt1–237.838 | Pt3–259.488 |
| TC3 | Pt2–244.01 | Pt3–227.373 | Pt1–244.001 |
| TC4 | Pt1–255.646 | Pt2–255.7 | Pt3–240.764 |
| TC5 | Pt3–259.016 | Pt2–244.336 | Pt1–244.009 |
| TC6 | Pt3–262.582 | Pt1–166.4 | Pt2–243.501 |

The texts and the audio were divided into three approximately equal parts and then the parts were assigned to be one of Training, Validate or Test data.
The number following the speech part identifier is the duration of that part in seconds. (The ordering of the parts as training data, validation data and test data was varied to negate the effect of any systematic variation in the speech as the recording progressed.)

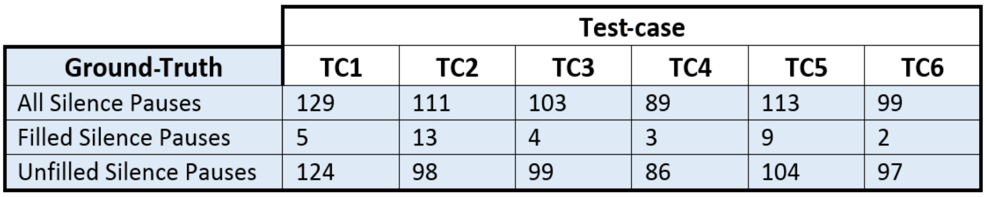
Table 4.2: Partitioning of the Test-Cases into the Training, Validation and Test Data Sets

## 4.3 The Ground-Truth

The ground-truth — information obtained by direct observation that is used to validate or confirm information that is obtained indirectly — that was used in the evaluation of the temporal accuracy of both the D$_{eterm}$Classifier and the DNN_Classifier, was established as follows.

Whilst listening to the test-cases, the locations of all perceived pauses were marked in a copy of the text. Subsequently the pauses were located in the waveforms, and from the waveforms the location of, and the duration for each of the silence pauses was measured to a resolution of 1mS, and manually recorded. The resultant datasets formed the ground-truth against which the computed results were later to be assessed; and a short form of the ground-truth — the pause counts for the six test-cases — is provided in Table 4.3, below.

The silence pauses, were found to be of two types: unfilled silence pauses with a recognisable period of silence (whether contaminated with significant amounts of noise, or not), and filled silence pauses with no recognisable period of silence. According to Green [1988], filled pauses are , "Nonlinguistic vocalisations", such as uhm, ah and er; but perhaps this definition is too narrow, because a pause can be suggested — and so perceived — simply by a change in speech cadence. For example in the phrase, "There was no recognisable period of silence", a pause can be suggested just by dwelling on the vowel in 'no'.



| | Test-case | | | | | |
|---|---|---|---|---|---|---|
| **Ground-Truth** | **TC1** | **TC2** | **TC3** | **TC4** | **TC5** | **TC6** |
| All Silence Pauses | 129 | 111 | 103 | 89 | 113 | 99 |
| Filled Silence Pauses | 5 | 13 | 4 | 3 | 9 | 2 |
| Unfilled Silence Pauses | 124 | 98 | 99 | 86 | 104 | 97 |

Table 4.3: Ground-Truth — Silence Pause Totals.

### Assessment Method

For all tests, the computed silence pauses were inserted into the filtered speech waveforms by overwriting the identified silence locations with true zero. Then the accuracy of the resulting silence pause starts and ends were manually measured to a granularity of $1ms$, compared with the ground-truth, and the differences were allocated to the appropriate error band; with each error band spanning $20ms$.

This is a similar order of accuracy to that discussed in the work of Wesenick and Kipp [1996], where the authors found for around three separate manual, broad phonetic segmentations of 64 sentences by 10 speakers in the German Language, that 96% were within $20ms$, 99% within $32ms$ and 100% within $64ms$.

### 4.4 The Programming Language

The D$_{eterm}$Classifier, the LogFB$_{dynamic}$ Speech Model, plus the Control Program for the outsourced ANN Software were all programmed in Ada using the GNAT Programming Studio© (©AdaCore). Ada 95 was selected because the author of this thesis has considerable experience of programming in Ada, and knows that Ada 95 has excellent number-crunching and serial file I/O capabilities — both of which are necessary to expedite this work. (An accessible introduction to Ada 95 is provided in Ben-Ari [1998]; and Ada 95 is defined by the Language Reference Manual: ISO/IEC-8652:1995, which is available from several sources on the Internet.)

Perhaps a criticism of the selection of Ada might be that it is not a particularly popular language, and for the more popular languages there may exist a greater variety of object libraries in the public domain. Another basis for criticism might be that the specification for Ada 95 is, as the name suggests, over 22 years old. The first of these points can be answered by recognising that a suitable ANN library written in C/C++ is available, as are the required Ada bindings; and the second by observing that Ada 95 is effectively a subset of Ada 2012. In fact, with a simple compiler switch, the Ada 2012 compiler could have been selected at the outset; but it would still remain the case that to implement the algorithms for this work, only a subset of Ada 95 is required.

Ada is often the language of choice for real-time high-integrity and safety critical embedded (cross-compiled) applications; but none of these attributes are particularly useful for this work. However the underpinning of these features by strict types, strong type checking and good programming practice ensures that virtually all coding errors are detected at compile time; and run-time errors can be virtually eliminated. Additionally exception handlers can be implemented within the source code — a useful capability when handling file I/O. (It seems to the author of this work that the separation of compilation errors and run-time errors is a useful decomposition of the problem of debugging; but that is not a universal view. In fact it is claimed on the website of one modern interpreted language, that the cycle of code/execute/debug is better.)

As implicitly suggested in the previous paragraphs, the choice of Ada as the primary programming language contributed to the software verification philosophy. Other than that the correct performance of the software system was verified using closed loop testing; where at various stages audio files were reconstructed and compared with the original recordings. For example, for the $\mathrm{D}_{eterm}$Classifier voiced and unvoiced speech files were built, and for the $\mathrm{LogFB}_{dynamic}$ model speech files were built for each of the Bark bands. The Filtered speech files (see Table 4.4) although purposed to removing noise due to air turbulence around the microphone and other low frequency noise, also served to verify correct operation of the DFT; and the insertion of the silence detected by the two classifiers into the Filtered speech files, in addition to providing a means for evaluating the performance of the classifiers, also provided a final verification of correct system operation.

## 4.5 The $\mathrm{D}_{eterm}$Classifier

For this work the three classification subsets of voiced speech, unvoiced speech and silence are upheld.

- Voiced speech is that part of intelligible speech that is produced by the vibration of the vocal chords (See Table 4.4).

- Unvoiced speech are those parts of intelligible speech that are not time co-incident with the sound produced by vibration of the vocal chords. For example, the unvoiced stops /p/, /t/ and /k/, the fricative /f/ and the sibilant fricatives /s/ and /sh/ (See Table 4.4).

- Silence comprises all parts of the audio recording that do not contribute to the intelligible speech. i.e. Silence is everything but vowels, semi-vowels and consonants.

### 4.5.1 Rationale

It may be thought that to distinguish silence from speech, it would only be necessary to classify silence as being that part of the audio with an energy level that is below some threshold. This is not the case because different recordings with different background noise conditions would require different thresholds, and bursts of noise, interference spikes and noisy releases would in any case inevitably exceed the background noise threshold. However, even though not all speech is linearly separable from silence, significant quantities of speech and silence can be separated with a simple energy threshold; and the purpose of the D$_{eterm}$Classifier is to automate the detection of the optimum silence/speech threshold — for any given recording.

It is the nature of speech that the energy in the lower frequency voiced parts of the speech may be several times greater than the energy of the higher frequency unvoiced parts of the speech (Atal and Rabiner [1976]). Accordingly the signal to noise ratio for the higher frequency components of the speech is significantly lower than the overall signal to noise ratio and the situation is worsened for weakly spoken consonants. So, the separation of the speech waveform into voiced and unvoiced speech is a useful decomposition of the problem of deterministic speech/silence classification, but according to Deekshitha et al. [2015] the discrimination between unvoiced speech and silence is not possible. Taking Atal and Rabiner's view (and implicitly testing the observation by Deekshitha et al.), the separation of the speech into voiced and unvoiced components can be achieved using a standard filtering technique: the Discrete Fourier Series (Section 2.4.1 herein, and Section 12.2: of Ivison [1978]). The short term energy waveform can be transformed into a harmonic series, harmonics deleted from the series as necessary, and the remaining harmonics summed to reconstruct the now filtered waveform (a method successfully used to subtract noise from speech by Boll [1979]). With this arrangement the search for the optimum unvoiced-speech/silence threshold is not influenced by the lower frequency voiced component of the speech and its attendant noise, and the search for the optimum voiced-speech/silence threshold is not influenced by the higher frequency unvoiced component of the speech and its attendant noise.

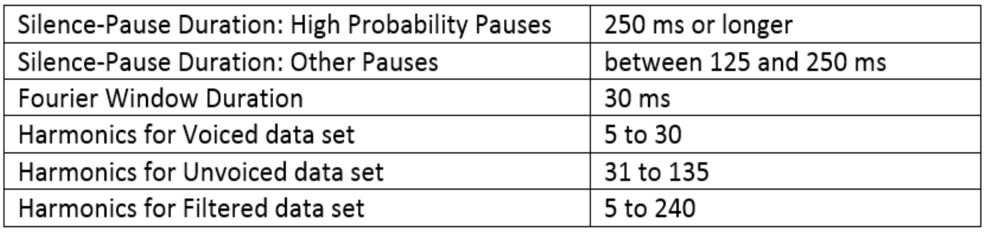
The advantage of using a deterministic system for locating silence in speech is that the likelihood that identified silence pauses will exist will be very high; but a disadvantage is that there is not likely to be sufficient information in the speech waveform to sustain classification of all silence pauses. It is the purpose of the D$_{eterm}$Classifier to provide an initial speech/silence binary classification of recorded speech so that accurate training data can be prepared for the DNN_Classifier. As training data is always some subset of all of the data, it is not necessary that all silences are identified; rather that those silences that are identified actually exist.

With a system which ultimately employs a fixed speech silence threshold (however derived), inevitably the noise during silence will occasionally result in periods during silence which are categorised as potential speech. If the short term power for this potential speech is compared with the short term power of a known segment of speech then the ratio of the two (Burileanu et al. [2000]) can be viewed as a probability that the potential speech is speech; and if this is repeated for all such potential speech bursts, an ordered table of probabilities can be built, and some criteria can be adopted for deciding which are speech and which are silence.

### 4.5.2 Detailed Description

1. The speech waveform was filtered using the trigonometric Discrete Fourier Transform and waveform reconstruction technique (see Section 2.4) with a Rectangular Window of 30$ms$. This window was shuffled in 10$ms$ steps through the audio data, and the filtered speech waveforms were reconstructed — using the 10$ms$ of the data that was at the centre of the window — by summing the harmonics (as listed in Table 4.4). The outcome of this process was three audio data sets: Voiced, Unvoiced and Filtered. The fundamental and first few harmonics were omitted from the reconstructed Voiced and Filtered data sets; the intention being to remove any non speech-sounds such as 50 or 60 Hz electrical supply noise (hum), and

the noise of low frequency turbulence around the microphone. The lowest fundamental frequency $f_0$ for the Voiced and Filtered speech of 167.045 Hz, is a little higher than the lowest voicing frequency of 124 Hz for men; as found by Petersen and Barney [1952].



| | |
|---|---|
| Silence-Pause Duration: High Probability Pauses | 250 ms or longer |
| Silence-Pause Duration: Other Pauses | between 125 and 250 ms |
| Fourier Window Duration | 30 ms |
| Harmonics for Voiced data set | 5 to 30 |
| Harmonics for Unvoiced data set | 31 to 135 |
| Harmonics for Filtered data set | 5 to 240 |

The selection for the ‘Silence Pause Duration: High Probability Pauses‘ was based upon work by Goldman-Eisler [1961] and Green [1988], and for the ‘Silence Pause Duration: Other Pauses’ was arbitrarily defined. The Fourier Window Duration was set to $30ms$ to provide an acceptably fine granularity for the harmonics; and the harmonic ranges for the voiced and unvoiced data sets were empirically chosen, by listening to reconstructed voiced and unvoiced speech for a variety of speakers.

Table 4.4: Constants For All Test-Cases

2. A preliminary value for the background noise threshold for the Voiced data set was then automatically determined by shuffling through the data in steps of $125ms$, and recording as the threshold the maximum value of the $n$ blocks of $250ms$ of contiguous samples which had the lowest overall signal magnitude—where $n$ was the number of expected pauses as automatically derived from the punctuation in the text.

3. This process was repeated for the Unvoiced data set and then the samples of each data set were compared with their corresponding preliminary noise thresholds as follows. For each data set, when the signal magnitude of the average of $1ms$ of contiguous samples was below the threshold then that block of samples was designated as potential silence; otherwise it was designated as potential speech. This process resulted in boolean Is-Voiced and Is-Unvoiced data sets.

4. Using logical ‘or’, the Is-Voiced and Is-Unvoiced data sets were combined and this resulted in a preliminary Is-Speech/Is-Silence data set. The duration of all silence-pauses identified in this set was measured, those greater than

250$ms$ were tagged as silence-pauses, and those greater than 125$ms$ and less than 250$ms$ were individually allocated a probability—proportional to pause duration—of being a silence-pause. (Goldman-Eisler [1961] defines the minimum duration for a pause to be around 250$ms$, and tags pauses of shorter duration than this as articulatory pauses, whereas Green [1988] defines the durations of pauses due to cognitive processing (i.e. not articulatory pauses) to be of the order of 250 to 300$ms$ or more.)

5. Separately a data set which was the magnitude of segments of 10$ms$ duration of the Filtered speech data-set expressed as a probability of silence was constructed; and when this data was combined with data derived in the previous steps, the result was a list of potential pauses.

6. From this list of potential pauses a list of the locations of all possible clauses was generated (‘clause’ is here used in the very limited sense of meaning a contiguous block of speech with no internal pauses), and this was used to gate the Filtered data set into a set of potential clauses. The maximum of the short term ($\frac{1}{3}ms$) magnitudes for each potential clause was calculated as was the maximum of the short term magnitudes for all voiced clauses (the clause prototype); and for each clause the ratio of these quantities (Burileanu et al. [2000]) was expressed as a probability where 1.0 indicated the certainty of speech.

7. A clause count as automatically obtained from the punctuation in the text was compared with the clause count derived during the previous steps. When the former was found to be greater than the latter, then the voiced and unvoiced preliminary noise thresholds were proportionally scaled up, and the automatic detection of potential clauses from the speech waveform was repeated — until either the clause count automatically detected from the speech waveform was found to be equal to or greater than the clause count obtained from the text (Figure 4.1) — or a limit was reached. This limit, the maximum value of the scaling multiplier, was empirically fixed at 9.0. Multiplying the voiced and unvoiced preliminary noise thresholds by more than 9.0 was found to result in erroneously high silence/speech thresholds.

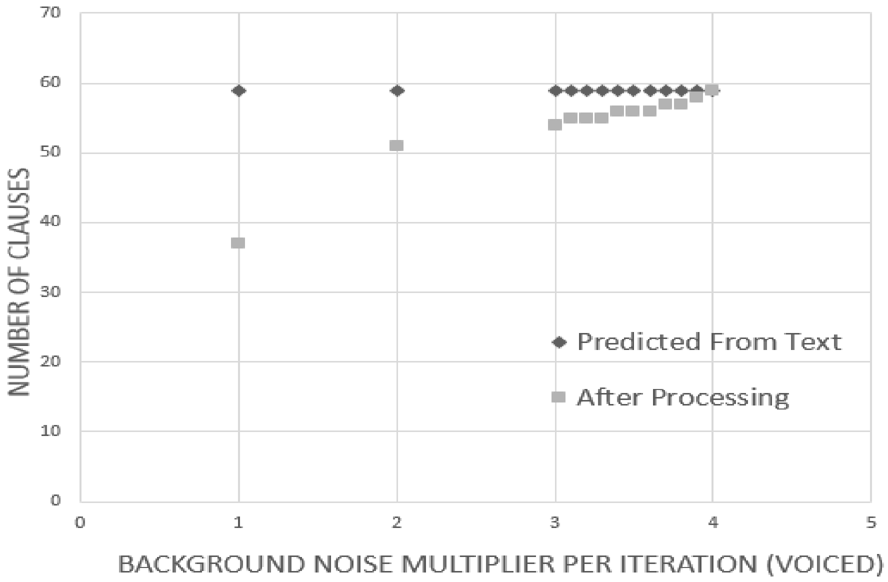


Figure 4.1: The Convergence of the Loop at Each Iteration.

8. The data set of probable clauses that was generated using the optimised noise thresholds was iteratively scanned—at each iteration with an increased probability threshold level; and for each threshold the clause count was recorded as the number of clauses with a probability greater than the threshold. When this data was plotted as a function of the incremental threshold, a characteristic curve resulted (Figure 4.2). The separation between the real clauses and the phony clauses could be identified from the data set where both the greatest change in threshold resulted in no change to the clause count, and the clause count was equal or near to the expected number of clauses. A threshold level in this range was adopted, and all potential clauses above this threshold were accepted as real clauses whereas those below the threshold were categorised as silence.

9. An Audio file was created where the Filtered speech was modified by setting all of the samples during the algorithmically determined silence pauses to zero, then the temporal location of all of the pauses in the audio waveform was measured to a resolution of $1ms$, and finally this data was compared with the location of the perceived pauses as defined in the ground-truth (Section 4.3).

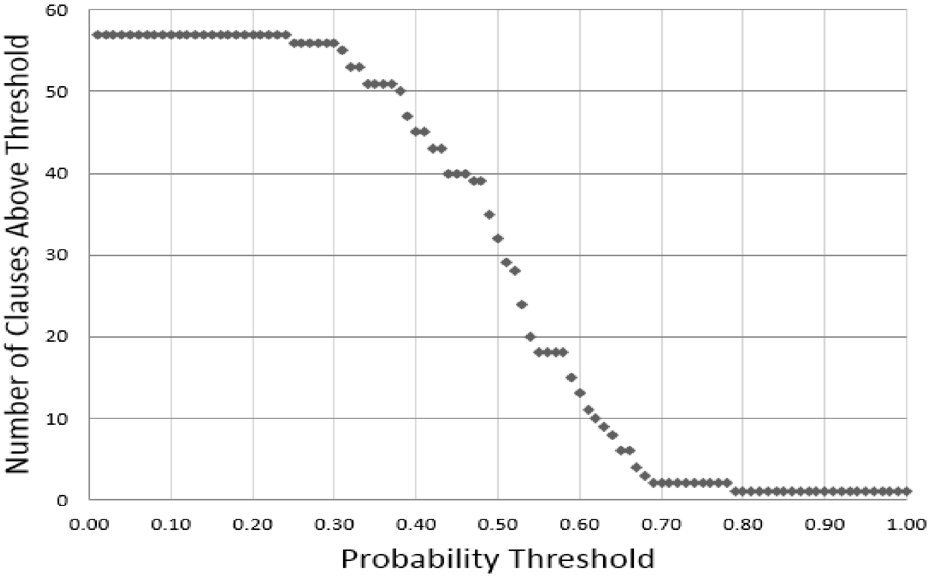


Figure 4.2: Number Of Clauses as a Function of $P_{threshold}$

## 4.6 The LogFB$_{dynamic}$ Speech Model

Appendix A provides relevant parts of the Bark Band Ada Specification.

### 4.6.1 Rationale

The question, *"What is the best parameter set for articulating the differences between silence and speech?"* was posed following the literature review, and one answer to this might be 'a psychoacoustics model of speech'. This is because the listener has no difficulty in discriminating between speech and often high levels of background noise; and my aim is to mimic this human capability to an extent.

A second question was, *"What is the most suitable window type and duration for the Discrete Fourier Analysis?"*. Because the quality of waveform reconstruction is important in this work, a rectangular window, without pre-emphasis is particularly suitable. Using pre-emphasis, and/or some variant of the Inverse Cosine window such as the Hamming Window, would introduce distortion — and this to no advantage. The second part of the question involved the resolution of the DFT and the consequent harmonic spacing, and a window with a duration of $30ms$ was selected. The reason for this is the harmonics are spaced at $33.33_{rec}$ Hz, and so even the lowest frequency Bark band will have a span of three harmonics; thus leaving scope for the selection of the most powerful harmonic in the band, thereby facilitating masking.

A third question was, *"Is it feasible to accommodate the phase component of the DFT as well as the magnitude component in the speech model?"*. The answer to this is 'yes', and this will be achieved with the LogFB$_{dynamic}$ Model by creating a set of audio files — one for each of the Bark bands, where each file is the summation of the harmonics for that band — scaled according to the simultaneous-masking regime.

A question so far not considered, is what might be the optimum duration for the model slices? The LogFB$_{dynamic}$ speech model is built by slicing the set of Bark band audio files into short equal slices and calculating the RMS magnitude for the elements which constitute each slice. The model data, the result of this process, is a set of 21 numbers. That is, one number per Bark band per slice (assuming that Bark bands 1, 23 and 24 are omitted because they span frequencies that are not within the spectrum of normal speech). As the model is in the time domain, it is not constrained by DFT windowing; and a shorter time-slice than the $\sim 10ms$ often selected for speech processing is feasible. So a duration for each slice of $\sim 1ms$ was chosen, because the greater temporal resolution would result in an improved representation in the model of short term events — such as transient noise. That said, it is still necessary to model the rate of change for the lower frequencies. The lowest frequency in Bark band 2 is $133.33_{rec}$ Hz (i.e. a period of $7.5ms$); and to accommodate this, 6 terms comprising the energy of the reconstructed waveform for the three $1ms$ slices before and the three $1ms$ slices after the slice of interest are added to the 21 Bark band energies, resulting in a model with 27 terms, a durational span of $\sim 7ms$, and a fine resolution of $\sim 1ms$.

With the duration of each of the consecutive segments of the speech model set to $1ms$, then this would limit the upper frequency limit of the model only insofar as changes to the RMS energies in any of the Bark bands would be accessed at

$1ms$ intervals.

### 4.6.2 Detailed Description

Equation 4.1, and Figure 4.3-A illustrate the conventional scaling technique for each of the triangular Bark band filters which constitute the Log Filter-bank (Figure 3.5), and Equation 4.2 and Figure 4.3-B show the scaling technique for each of the filters for the LogFB$_{dynamic}$. The conventional auditory filter arrangement (Figure 4.3-A) does not support dynamic tracking of the most powerful harmonic in each of the Bark bands, nor does it implement any form of simultaneous masking; whereas the filter arrangement of Figure 4.3-B — by imposing a progressively increasing attenuation of the harmonics to and from the most powerful harmonic in the Bark band — implements both dynamic tracking and a form of simultaneous masking. Note that this method of implementing masking is something of a simplification, because the masking effect has been shown by Egan and Hake [1950], as being asymmetrical about the masker, masking more of the lower frequencies for lower test signal levels and more of the higher frequencies at higher test signal levels.

$$Y_f = \sum_{n=PB_FL}^{H_p} (\frac{n - PB_FL}{H_p - PB_FL}) LogX_n + \sum_{n=H_p}^{PB_FU} (1 - \frac{n - H_p}{PB_FU - H_p}) LogX_n \tag{4.1}$$

$$Y_f = \sum_{n=PB_FL}^{H_p} (1 - \frac{n - PB_FL}{H_p - PB_FL}) LogX_n + \sum_{n=H_p}^{PB_FU} (\frac{n - H_p}{PB_FU - H_p}) LogX_n + LogX_{H_p} \tag{4.2}$$

In Equations 4.1 and 4.2, $Y_f$ is the energy in the band, $H_p$ is either the most powerful or the central harmonic of the passband filter $PB_F$; $PB_FL$ is the passband lower harmonic limit, $PB_FU$ the upper harmonic limit and $X_n$ is the magnitude of the n$^{th}$ harmonic. (For waveform reconstruction, the Log operator is replaced by 1.)

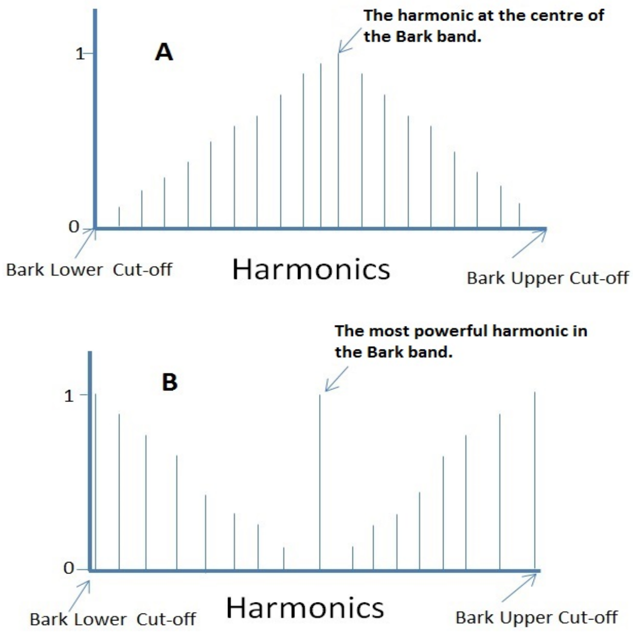


Figure 4.3: Simultaneous Masking Strategy

**A:** The conventional auditory filter arrangement similar to that used for MFCCs (Figure 3.5).

**B:** An arrangement which imposes simultaneous masking. This arrangement has some similarities with that used by Zhu and 0'Shaughnessy [2004], but differs in that the slopes of the filter used by Zhu and O'Shaughnessy are the independent variables, whereas here the slopes of the filter are variables dependent upon only the location of the centre (or most powerful) harmonic in each of the Bark bands, and the fixed Band upper and lower cut-off frequencies.

One possible disadvantage with the regime of Figure 4.3-B is that whereas the most powerful harmonic is depicted as being coincident with the central harmonic of the Bark band, in fact — because of dynamic tracking — it may exist anywhere within the band; and as the Bark lower and upper cut-off frequencies are fixed, then the masking — as implemented using Equation 4.2 — will as a result often be asymmetrical. Conversely, the advantage of the regime shown in Figure 4.3-B, is that the Bark band upper and lower cut-off frequencies are fixed from the outset, thereby eliminating the need for any 'tuning' of the cut-off frequencies. This approach factors on the point made in Chapter 3:

> **"...And it may be appropriate to ask whether a model which conforms to the Bark scale and comprises the minimum number of critical bands, with dynamic masking arranged around the most powerful harmonic in each band, and with**

**the filter bandwidth limited to the upper and lower cut-off frequencies in each band, can provide an accurate representation of speech"**

The LogFB$_{dynamic}$ Model is generated as follows:

1. The speech waveform is translated into the frequency domain using the trigonometric discrete Fourier Transform.
2. The harmonics which constitute each of the Bark bands (See Appendix A) are searched and the most powerful harmonic in each of the Bark bands is identified.
3. The magnitude of the harmonics within each of the Bark bands are scaled about the most powerful harmonic, and then summed (to implement simultaneous masking) in accord with Equation 4.2.
4. The result of scaling and summing the harmonics for each of the Bark bands is 24 audio files — one for each of the Bark bands; and the RMS magnitude over $1ms$ is calculated and normalised to between 0.0, and 1.0, for each of the bark bands. The result of this is an array of numbers 24 by the length of the recording in milliseconds; corresponding to a model slice width of $1ms$.
5. From the Filtered speech waveform, the RMS energies for the three $1ms$ time-slices immediately before the slice of interest are calculated, as are the RMS energies for the three $1ms$ time-slices immediately following the slice of interest. These 6 $\frac{dy}{dt}$ terms are similarly normalised, and then concatenated with the magnitudes for Bark bands 2 to 22, resulting in a speech model with 27 terms.

The LogFB$_{dynamic}$ Model comprises data in the range of 0 to $2^{15}$, which when normalised to the range of 0.0 to 0.1 — as required for the ANN — necessitates a resolution of 0.0000305176. When the finalised model data is stored (File Type B, Appendix D), each number is truncated by allocating 7 digits for the fractional part.

Because silence magnitudes are near 0, then to the data elements (already normalized to between 0 and 1.0), 0.5 is added to all values below 0.5, and 0.5 is subtracted from all values above 0.5. This places the model data with most relevance (silence), in the middle of the ANN dynamic range. This technique was validated, where it was found that the DNN_Classifier when trained with modified data identified more silence than when trained with unmodified data; and further that for two of three training cycles, the number of training epochs before auto-termination at the onset of over-training was reduced.

## 4.7 ANN Training Data Selection

The ANN is trained using supervised training, where each training input must have an associated fully specified output condition. So the training data must comprise the model data plus a flag to indicate whether the ANN output for that particular data instance is to be silence or speech. A list of the temporal locations in the speech recordings of all of the detected silence pauses with a duration of $250ms$ or more, is output by the D$_{eterm}$Classifier, and this is used for the generation of the silence/speech flags.

Figure 4.4 is a short section of the waveform for TC1, and shows the speech and silence training data gating rules, where — for illustrative purposes only — the location of the silence as detected by the D$_{eterm}$Classifier is represented as true zero. The speech training data blocks have a duration of 500 ms, and the silence training blocks a duration that is the lesser of the detected duration of the silence pause or $500ms$. The choice of two speech segments per silence segment is to ensure that both the speech end-points and onsets are represented in the training data, as these can have very different characteristics: speech onsets can be very abrupt whereas for end-points, the speech often just dwindles away to nothing.

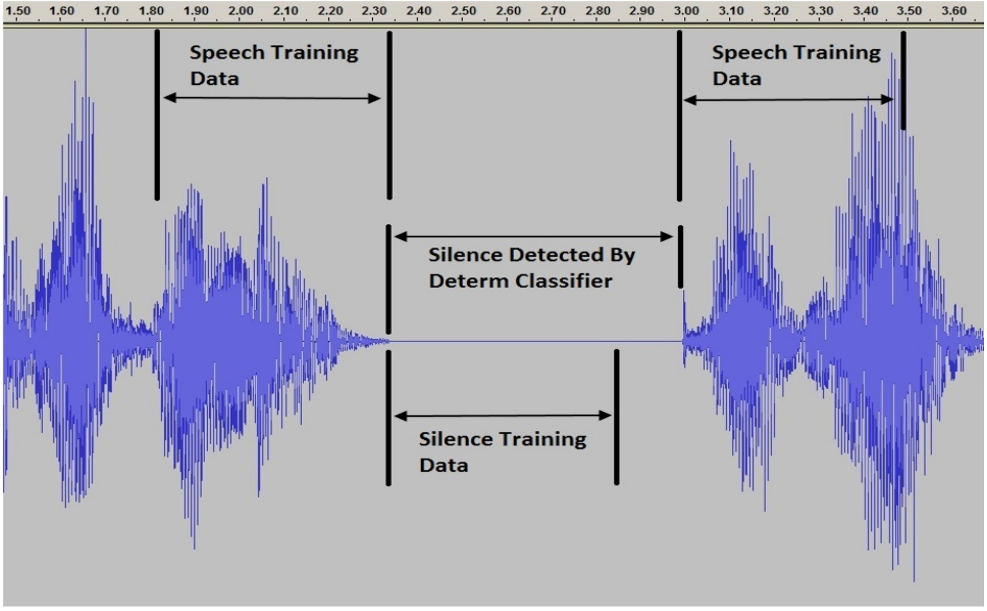


Figure 4.4: Speech and Silence Training Data Gating Rules.

### 4.8 The DNN_Classifier

This work was expedited by using the open source **Fast Artificial Neural Network (FANN) C library** (© Nissen [2003] as released under the LGPL License [Free Software Foundation, 1999],) , together with the **FANNAda Bindings** (© Andreasen [2015]). For training, the FANN© C library was configured to use its internal version of the Resilient Propagation (RPROP) back-propagation algorithm (Reidmiller and Braun [1993]). The FANN© Library was chosen because in addition to providing all of the ANN functionality required, it is open source — and so may be modified — should that prove necessary.

To train and evaluate the ANN Classifier, the Split Sample Technique (Priddy and Keller [2005] — Section 7.3.1) was adopted as follows. Each data set was split into three sections, the training data set (TrnDS), the validation data set (ValDS), and the test data set (TstDS). The ANN was then trained using the TrnDS, and at intervals of 5 epochs was validated using the ValDS. If the reported Mean Square Error (MSE) — a distance measure of the difference between the requested result and the achieved result — was found to be lower than previous, then the ANN was stored and the 5 epoch training cycle repeated; if the MSE from the validation test was found to be increasing for two successive validation cycles, then that indicated the onset of over-fitting and the training was ended. Following training the performance of the binary classifier was evaluated using the TstDS.

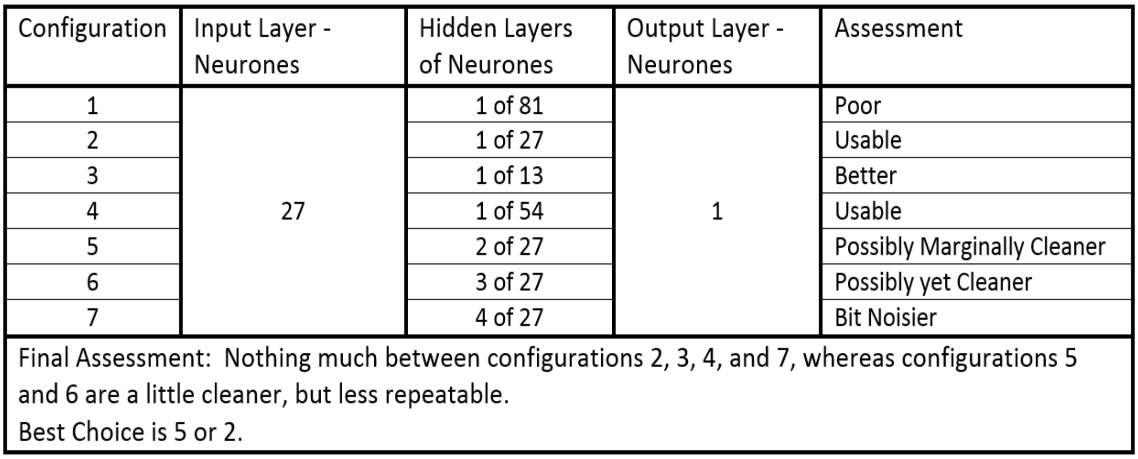
### 4.8.1 Configuration

The implication of the hypotheses (Section 1.4) is that the performance of the systems described herein is more a function of the quality of the audio model and the selection of the training data, than the configuration of the artificial neural network; and the outcome of the exploratory work for the, 'Experiments with MP3 Compressed Speech' ( Appendix C), was that the performance of a three layer MLP, with 27 Input Neurons, 54 Hidden Neurons and 1 Output Neuron, and with a Sigmoid Activation, was sufficient to support the testing of the hypotheses.

Following the completion of the 'Experiments with MP3 Compressed Speech' (chronologically the first part of this work), and due to a growing awareness from several sources, of the importance of the Deep Neural Network for ASR, the decision was taken that various potential network configurations should be further assessed to better inform the choice of final network configuration — before commencing the experiments for the main body of this work.

There was no similar imperative to consider changing the sigmoid activation function, the RPROP training algorithm, or the Train/Validate/Test technique; plus the experience with 'Experiments with MP3 Compressed Speech' showed these to be an excellent combination — and fit for purpose. When such a situation obtains, where the potential for performance improvement is slight, it becomes difficult to assess whether an alternative configuration would provide an improved — or just a different — classification performance. For example, from the results for the assessment of network topologies herein, it was possible to identify the topology which resulted in the least accurate classification, but almost impossible to identify the topology which provided the most accurate classification.

The results of the assessment of various network configurations — using the data set for TC1-N from the ‘Experiments with MP3 Compressed Speech’ and the training data selection rules as shown in Figure C.1 — are shown in Table 4.5.



| Configuration | Input Layer - Neurones | Hidden Layers of Neurones | Output Layer - Neurones | Assessment |
|---|---|---|---|---|
| 1 | 27 | 1 of 81 | 1 | Poor |
| 2 |  | 1 of 27 |  | Usable |
| 3 |  | 1 of 13 |  | Better |
| 4 |  | 1 of 54 |  | Usable |
| 5 |  | 2 of 27 |  | Possibly Marginally Cleaner |
| 6 |  | 3 of 27 |  | Possibly yet Cleaner |
| 7 |  | 4 of 27 |  | Bit Noisier |

Final Assessment: Nothing much between configurations 2, 3, 4, and 7, whereas configurations 5 and 6 are a little cleaner, but less repeatable.
Best Choice is 5 or 2.

Table 4.5: The NN Classifier: Results of the Assessment of Potential Configurations

With the exception of Configuration 1 (1 hidden layer of 81 neurones), the results for the various configurations were found to be generally satisfactory; and Configuration 5 (2 hidden layers each of 27 neurones) was adopted for the work herein, because the informal testing showed it to be marginally better than Configuration 2 (1 hidden layer of 27 neurones), at cleanly identifying sibilants. (Configuration 2 occasionally confused small sections of the sibilants — of the order of a few milliseconds duration — with silence.)

Although the performance of the various configurations was found — to a small extent, and for the same training data — to vary, the assessment was most effective at identifying the poor performing networks, and less successful in identifying the optimal network configuration. In Section 4.9, on the consistency of the DNN training process, the issue of identifying the optimal network configuration is again addressed.

In summary, the DNN is comprised 27 input neurons, 2 hidden layers of 27 neurons with the sigmoid activation function (range 0.0–1.0), and 1 output neuron; and is trained using the Resilient Propagation (RPROP) back-propagation algorithm (Reidmiller and Braun [1993]). The silence example span, and clause example span are both set to a maximum of 500 milliseconds.

### 4.8.2 Test Method

**Performance:**

1. The test-cases were processed using the $\mathrm{D}_{eterm}$Classifier as described in Section 4.5, and the locations of the silence pauses in the Training, Validation, and Test data sets were automatically recorded.

2. For each of the test-cases...

   (a) a speech model was created as described in Section 4.7, and for the TrnDS, the model data was partitioned into known speech and known silence, using the locations of silence pauses identified during the $\mathrm{D}_{eterm}$Classification.

   (b) The DNN was trained using the TrnDS created in the previous step, and at intervals of 5 epochs the 'fit' of the DNN was checked using the ValDS. At the onset of over-fitting the training process was ended. During the training process, a new DNN was stored only when the MSE from the validate process was found to be less than that of the previous DNN. The reason for this was to ensure that only the best trained DNN was stored, rather than the last trained DNN.

   (c) The TstDS in its entirety was processed through the DNN_Classifier, and the locations of all silence slices were recorded.

   (d) The original speech waveform was gated such that the locations of the automatically detected silence slices were set to zero. From the resulting audio file the location of all such pauses was measured to a resolution of $1ms$.

   (e) The silences indicated by the $\mathrm{D}_{eterm}$Classifier and the silences indicated by the DNN_Classifier, were compared with the ground-truth (see Section 4.3).

## 4.9 Consistency of the DNN Training Process:

### 4.9.1 Variation as a $f_{\text{(Initialisation)}}$

Whereas the result of the pattern recognition process using the trained DNN is deterministic, this is not the case for the DNN training process, which involves both a random initialisation of the weights for each of the artificial neurones and a training heuristic. So to test the extent to which different instances of training introduces variation in the resultant DNNs, each of the test-cases was trained 100 times with the same training, validation and test data sets. The dependent variables in this test were the durations of detected speech and silence (the sum of which is a constant), and the former of these was recorded. The original speech waveform was gated such that the locations of the automatically detected silence slices were set to zero, and samples of the audio files were qualitatively evaluated.

### 4.9.2 Variation as a $f_{\text{(Depth of Neural Network)}}$

The evaluation of the various configurations for the Neural Network, described in Section 4.8.1, was useful for identifying those network configurations that were least effective, but was less useful in identifying the configuration that would be most effective. However, when evaluating the consistency of the training process, it was necessary to consider various configurations for the Neural Network — to establish the extent to which network topology influenced the performance of the classifier. So, to further evaluate the performance of various configurations for the Neural Network, the method described in the previous section (Section 4.9.1) which comprised 100 training/execution cycles (with a fixed training, validation and test data set) was again used; but this time for only one of the test-cases, and for network configurations 2, 5, and 7 from Table 4.5, plus an additional network configured with 27 input neurones, six layers of hidden neurones (each comprising

27 neurones) and one output neurone. For this test the dependent variables were again the computed durations of the speech and silence; and the testing was extended by comparing the significant errors in classification for each of the configurations — for an audio sample with a speech duration around the mean for the test and with the detected silence slices set to zero — with the significant errors in classification as previously recorded for the chosen test-case.

## 4.10 Generalisability of the Solution:

The test-cases were all recorded under similar conditions — though on different days and at different times — and in the same voice, and therefore it is reasonable to expect that the neural nets for each of the test-cases should be similar to each other, such that any of the neural nets will provide a satisfactory classification for any of the test-cases. This was investigated as follows:

For each of the test-cases, and for each of the trained DNNs...

1. The TstDS in its entirety was processed through the DNN_Classifier, and the detected durations of both the silence and the speech were recorded.
2. The original speech waveform was gated such that the locations of the automatically detected silence slices were set to zero, and the audio files were qualitatively evaluated.

# Chapter 5: Results and Analysis

## 5.1 Selection of Test Material

An unwritten convention exists amongst part of the ASR research community that publicly available textual and audible test material should be used; and there are two reasons for this. Firstly a common test corpus provides the capacity for comparing the results of new processing techniques with research results already reported, and secondly with a common test corpus, it should be feasible to validate the work of other researchers by repeating their work.

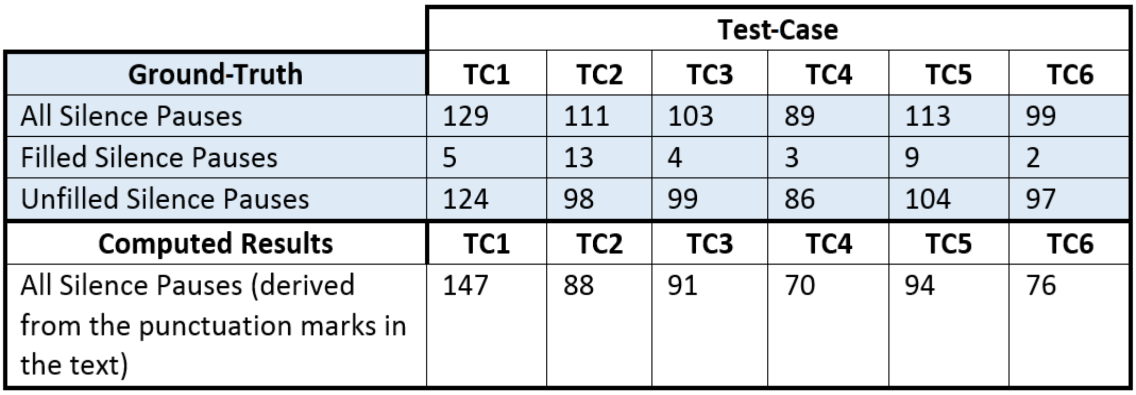
My choice is to ignore this convention and in this instance to work with new recordings. There are three reasons for this. Firstly the investigation of a new model should start with test material that has as few practical constraints on the bandwidth or recording quality of the speech as is feasible. Secondly only with full control of the recording environment is it possible to repeat recordings, or to extend the corpus in a controlled manner, should that be necessary. Thirdly, it is not always the case that performance of speech processes with a particular speech corpora is representative of the performance of those processes on other recorded speech. In the paper, 'Deep Neural Networks for Acoustic Modelling in Speech Recognition [The shared views of four research groups]' the authors write that:

***"Experience has shown that performance improvements on TIMIT do not necessarily translate into performance improvements on large vocabulary tasks with less controlled recording conditions and much more training data. Nevertheless..."***. — Hinton et al. [2012]

In the earlier stages of the development of the ANN_Classifier, a corpus which comprised eight MP3 Speech Samples was used. The report on this work has been relegated to an appendix, because without a full analysis of the speech samples which constitute the MP3 speech corpus, it is not possible to know the extent of the psychoacoustics encoding, and as a consequence to baseline the results. Even so, the results are of interest, and are discussed in Section 5.7.

## 5.2 Text Analysis

At step 2 of the algorithm for the D$_{eterm}$Classifier, for each test-case the notional number of silence pauses therein is computed by scanning through the text, and identifying and counting the punctuation marks. Table 5.1 provides a comparison for each test-case of this number with the silence pause counts defined in the ground-truth.

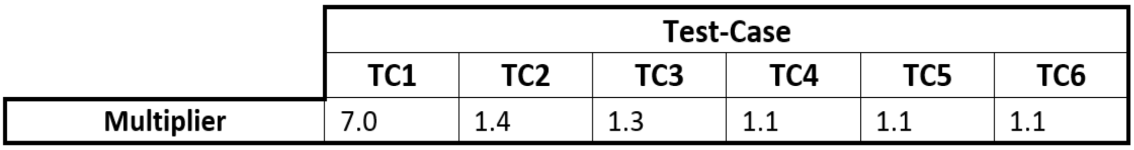


| | Test-Case | | | | | |
|---|---|---|---|---|---|---|
| **Ground-Truth** | **TC1** | **TC2** | **TC3** | **TC4** | **TC5** | **TC6** |
| All Silence Pauses | 129 | 111 | 103 | 89 | 113 | 99 |
| Filled Silence Pauses | 5 | 13 | 4 | 3 | 9 | 2 |
| Unfilled Silence Pauses | 124 | 98 | 99 | 86 | 104 | 97 |
| **Computed Results** | **TC1** | **TC2** | **TC3** | **TC4** | **TC5** | **TC6** |
| All Silence Pauses (derived from the punctuation marks in the text) | 147 | 88 | 91 | 70 | 94 | 76 |

Note the difference between TC1, and TC2 to TC6. One reason for this may be that the Test Data Set for the former comprises a reading of a narrative combined with conversational elements; whereas for the latter, just narratives.

Table 5.1: The D$_{eterm}$Classifier: Silence Pause Totals, as Derived from the Punctuation Marks in the Text

The results suggest that estimating the number of silence pauses from the punctuation in a script can yield different results depending upon the type of script. Specifically — and for a very small sample set — the method yields an underestimate of the silence pauses total from the punctuation in the script for the readings of the narratives (TC2 to TC6), and an overestimate for the reading of a narrative with conversational elements. The ramifications of this are that as the set-point (the computed pause count given in Table 5.1) for Step 7 of the algorithm for the D$_{eterm}$Classifier for TC1 is incorrectly high, then the control loop must drive the silence/speech threshold artificially high — by increasing the speech/silence threshold multiplier to the level given in Table 5.2 — to achieve the demanded pause count. The effect of this can be observed in the results (Table 5.3) where for TC1 the D$_{eterm}$Classifier identifies all of the Silence Pauses and several other silence insertions, whereas for the other test-cases the D$_{eterm}$Classifier identifies most, but not all of the silence pauses. This effect could be mitigated either by adopting a more complex algorithm for relating punctuation patterns to silence pauses, or by obtaining a statistical distribution of the relationship between word rate and pause rate for a broad range of samples in order to ascertain the most probable pause count (pauses per minute); and then using this in lieu of the pause count calculated from the punctuation in the text.



| | Test-Case | | | | | |
|---|---|---|---|---|---|---|
| | TC1 | TC2 | TC3 | TC4 | TC5 | TC6 |
| Multiplier | 7.0 | 1.4 | 1.3 | 1.1 | 1.1 | 1.1 |

For each of the test-cases, the multipliers of the initial speech/silence thresholds necessary to obtain loop closure are shown. The initial silence/speech thresholds are a function of the noise levels during silence (Step 2 of the D$_{eterm}$Classifier Algorithm), and since TC1 has the highest levels of noise during silence (Table 5.8), and the highest threshold multiplier (a function of the punctuation count), then the inevitable result is an incorrectly high silence/speech threshold.

Table 5.2: Extract From the Instrumentation Logs for Step 7 of the D$_{eterm}$Classifier Algorithm

## 5.3 The D$_{eterm}$Classifier

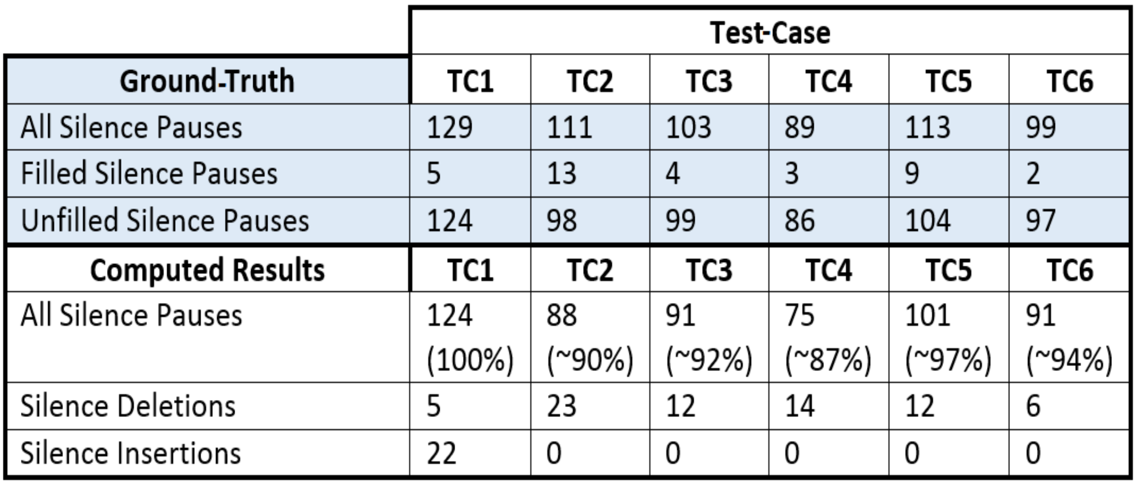
A basic measure of the performance of the D$_{eterm}$Classifier was obtained by comparing the Silence Pause counts with the ground truth; and the results of this comparison are considered in Section 5.3.1. Thereafter, in Section 5.3.2 the consolidated results of a detailed comparison of the accuracy of the D$_{eterm}$Classifier versus the ground-truth are considered; and in Section 5.3.3 classification failures are dealt with in detail. Note that the source data for the ground-truth, and for the classification results are not reproduced herein.

### 5.3.1 Basic Performance

Table 5.3 compares the Silence Pause counts as obtained using the D$_{eterm}$Classifier, with the those defined in the ground-truth; and also identifies the totals for the Silence Insertions and Deletions.

The D$_{eterm}$Classifier is configured to identify only the unfilled silence pauses, and it can be seen from table 5.3, that for TC1 all of the unfilled silence pauses were detected, and for the other 5 test-case the detection rate was between 87 and 97%. So, for this particular measure, the overall classification performance must be considered to be poor; but as the purpose of the D$_{eterm}$Classifier is to identify training data for the DNN_Classifier, then other measures of performance are also important. Those measures are, the timings for the silence/speech and speech/silence transitions, and the extent to which the silence pauses and the speech segments are clean. That is, it is not essential that the D$_{eterm}$Classifier identifies all pauses; but that for those pauses identified that the identification is accurate so that the DNN Training data is corrupted with as few bad training samples as is possible.

Only for TC1 were any silences other than those identified in the ground-truth detected. Such silence ‘insertions’ may be associated with Filled Silence Pauses, or be momentary periods of silence such as those occurring at stops. Examination of the detailed pause duration data (not reproduced herein) shows that the silence pauses identified in the ground-truth have a duration in the range of 0.051 to 4.518 seconds whereas the inserted pauses have a duration of 0.058 to 0.269 seconds; and because the two ranges overlap, it is not possible to classify the shorter pauses as



| | Test-Case | | | | | |
|---|---|---|---|---|---|---|
| **Ground-Truth** | **TC1** | **TC2** | **TC3** | **TC4** | **TC5** | **TC6** |
| All Silence Pauses | 129 | 111 | 103 | 89 | 113 | 99 |
| Filled Silence Pauses | 5 | 13 | 4 | 3 | 9 | 2 |
| Unfilled Silence Pauses | 124 | 98 | 99 | 86 | 104 | 97 |
| **Computed Results** | **TC1** | **TC2** | **TC3** | **TC4** | **TC5** | **TC6** |
| All Silence Pauses | 124<br>(100%) | 88<br>(~90%) | 91<br>(~92%) | 75<br>(~87%) | 101<br>(~97%) | 91<br>(~94%) |
| Silence Deletions | 5 | 23 | 12 | 14 | 12 | 6 |
| Silence Insertions | 22 | 0 | 0 | 0 | 0 | 0 |

The D$_{eterm}$Classifier is configured to automatically identify only those pauses which include a period of recognisable silence in the waveform — i.e. the 'Unfilled Silence Pauses' in the ground-truth. 'Silence Insertions' are short periods of silence that may be coincident with the Filled Silence Pauses, or may result naturally as part of the articulation of the unvoiced stops, /p/, /t/, and /k/.

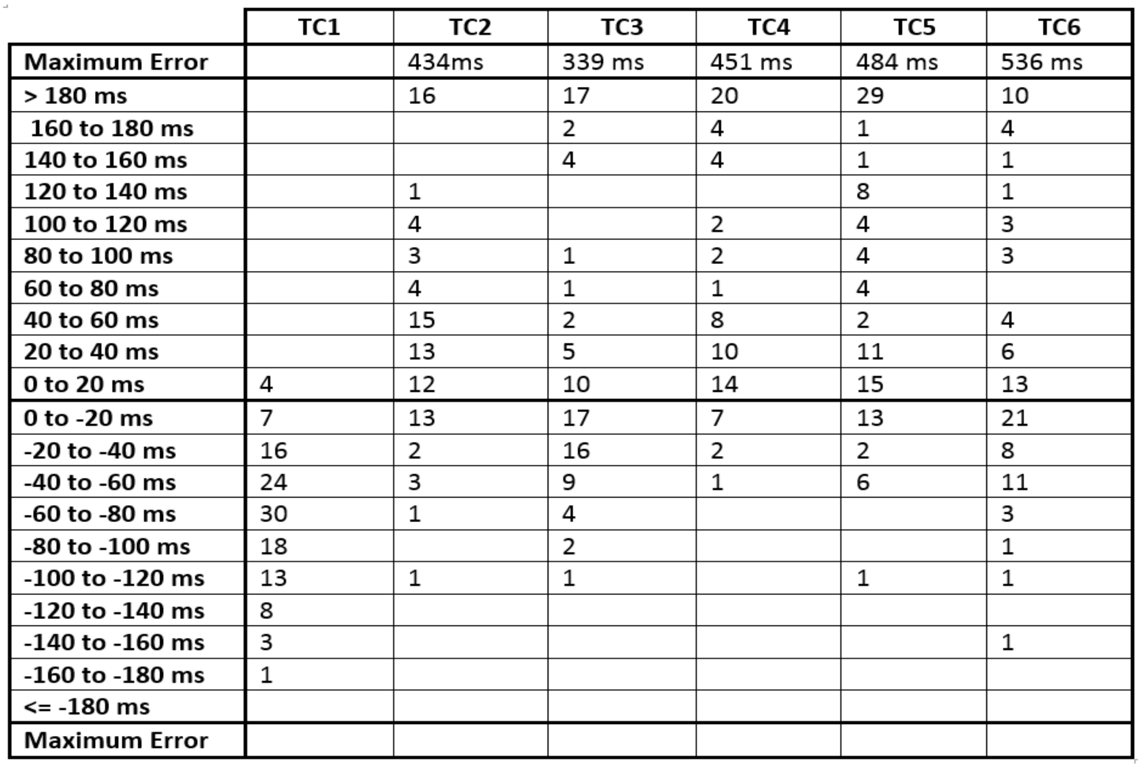
Table 5.3: The D$_{eterm}$Classifier: Silence Pause Detection, Insertion and Deletion Totals

either perceived or inserted, on the basis of duration alone.

### 5.3.2 Detailed Performance Analysis

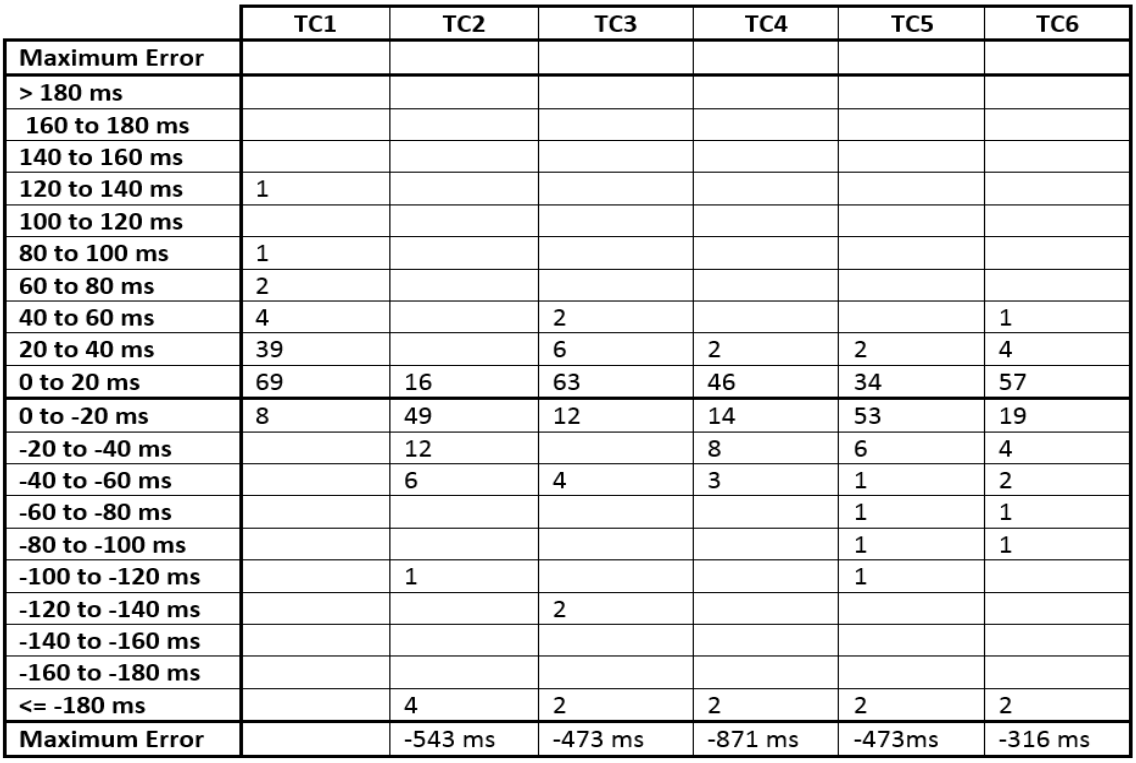
Table 5.4 shows the result of a comparison of the accuracy of the detected silence pause start times for the D$_{eterm}$Classifier, with the ground-truth; and Table 5.5 shows the result of a comparison of the accuracy of the detected silence pause end times for the D$_{eterm}$Classifier, also with the ground-truth. A negative error indicates early detection of the indicated event, and a positive error late detection. For each of the test-cases, the number of indicated events within the specified error bands are shown.

The data from Table 5.4 and 5.5 is also reproduced in the form of histograms of the temporal location of the computed silence pause starts/ends versus the ground-truth, in Figures 5.1 and 5.2. It is evident that for test-cases 2 to 6 inclusive, that although the silence pause start and end errors conform to a sort of distribution there is more than a scattering of outliers — particularly so for the silence pause start errors. The distributions for the test-cases is far from the ideal, which would be all errors within ±20$ms$ of zero; and so the D$_{eterm}$Classifier must be considered to be a high variance classifier, and this is confirmed by the sample standard deviations for the test-cases given in Figure 5.3.



| | TC1 | TC2 | TC3 | TC4 | TC5 | TC6 |
|---|---|---|---|---|---|---|
| **Maximum Error** | | 434ms | 339 ms | 451 ms | 484 ms | 536 ms |
| **> 180 ms** | | 16 | 17 | 20 | 29 | 10 |
| **160 to 180 ms** | | | 2 | 4 | 1 | 4 |
| **140 to 160 ms** | | | 4 | 4 | 1 | 1 |
| **120 to 140 ms** | | 1 | | | 8 | 1 |
| **100 to 120 ms** | | 4 | | 2 | 4 | 3 |
| **80 to 100 ms** | | 3 | 1 | 2 | 4 | 3 |
| **60 to 80 ms** | | 4 | 1 | 1 | 4 | |
| **40 to 60 ms** | | 15 | 2 | 8 | 2 | 4 |
| **20 to 40 ms** | | 13 | 5 | 10 | 11 | 6 |
| **0 to 20 ms** | 4 | 12 | 10 | 14 | 15 | 13 |
| **0 to -20 ms** | 7 | 13 | 17 | 7 | 13 | 21 |
| **-20 to -40 ms** | 16 | 2 | 16 | 2 | 2 | 8 |
| **-40 to -60 ms** | 24 | 3 | 9 | 1 | 6 | 11 |
| **-60 to -80 ms** | 30 | 1 | 4 | | | 3 |
| **-80 to -100 ms** | 18 | | 2 | | | 1 |
| **-100 to -120 ms** | 13 | 1 | 1 | | 1 | 1 |
| **-120 to -140 ms** | 8 | | | | | |
| **-140 to -160 ms** | 3 | | | | | 1 |
| **-160 to -180 ms** | 1 | | | | | |
| **<= -180 ms** | | | | | | |
| **Maximum Error** | | | | | | |

Table 5.4: The D$_{eterm}$Classifier: Silence Pause Start Error



| | TC1 | TC2 | TC3 | TC4 | TC5 | TC6 |
|---|---|---|---|---|---|---|
| Maximum Error | | | | | | |
| > 180 ms | | | | | | |
| 160 to 180 ms | | | | | | |
| 140 to 160 ms | | | | | | |
| 120 to 140 ms | 1 | | | | | |
| 100 to 120 ms | | | | | | |
| 80 to 100 ms | 1 | | | | | |
| 60 to 80 ms | 2 | | | | | |
| 40 to 60 ms | 4 | | 2 | | | 1 |
| 20 to 40 ms | 39 | | 6 | 2 | 2 | 4 |
| 0 to 20 ms | 69 | 16 | 63 | 46 | 34 | 57 |
| 0 to -20 ms | 8 | 49 | 12 | 14 | 53 | 19 |
| -20 to -40 ms | | 12 | | 8 | 6 | 4 |
| -40 to -60 ms | | 6 | 4 | 3 | 1 | 2 |
| -60 to -80 ms | | | | | 1 | 1 |
| -80 to -100 ms | | | | | 1 | 1 |
| -100 to -120 ms | | 1 | | | 1 | |
| -120 to -140 ms | | | 2 | | | |
| -140 to -160 ms | | | | | | |
| -160 to -180 ms | | | | | | |
| <= -180 ms | | 4 | 2 | 2 | 2 | 2 |
| Maximum Error | | -543 ms | -473 ms | -871 ms | -473ms | -316 ms |

Table 5.5: The D$_{eterm}$Classifier: Silence Pause End Error

Within the set of results, test-case 1 is anomalous in that the silence pause start and end errors conform more to the normal distribution, and there are no extreme outliers; and further, TC1 is the only test-case for which silence insertions were detected. Accordingly, for the TC1 data set, the D$_{eterm}$Classifier must be considered to be a low variance classifier and this is confirmed by the sample standard deviations for the test-cases given in Figure 5.3. However, the significant errors in the classification of speech as silence that were observed for TC1 (see Table 5.6), were not present in the classifications for TC2 to TC6, where all errors were in the classification of silence as speech. Accordingly for the TC1 data set only, the D$_{eterm}$Classifier must also be considered to be a low accuracy classifier.

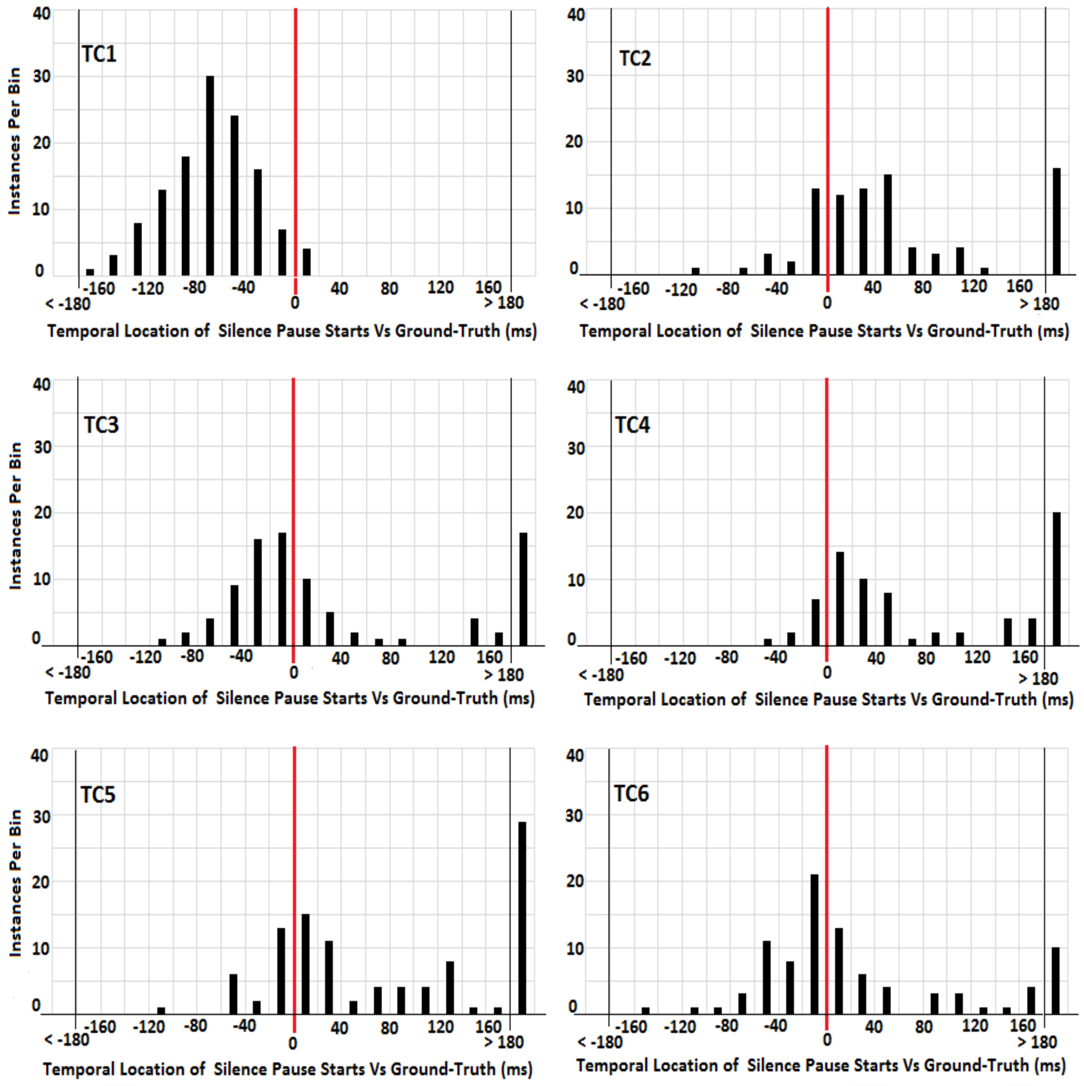


Figure 5.1: Histograms showing the Temporal Locations of All Computed Silence Pause Starts Vs the Ground-Truth, for the D$_{eterm}$Classifier:

The x-axis shows the temporal variation in the computed silence pause starts versus the ground-truth as allocated to 20ms bins, and the y-axis shows the number of instances of speech to silence transitions within each bin. The temporal locations of the speech to silence transitions from the ground-truth are all represented by the centre 0 co-ordinate on the x-axis.

### 5.3.3 Classification Failures

Classification failures are of three types. The first type are the errors that are perceptible with careful listening, when the Filtered Speech is compared with the Gated Speech (which is the filtered speech with all detected silence set to zero); the second type are the errors where noise during perceived silence pauses is incorrectly classified as speech , and the third type are when a substantial audible

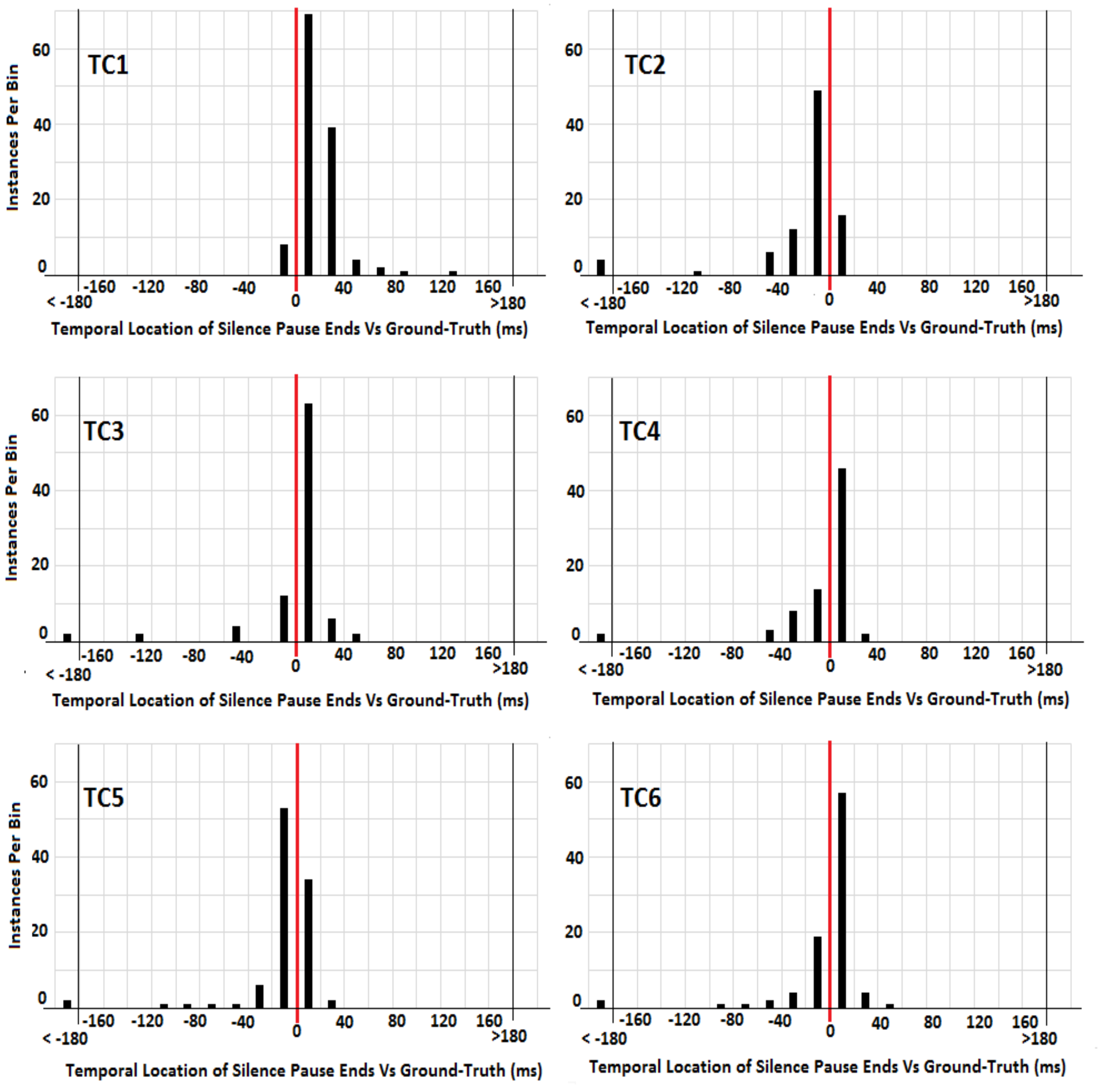


Figure 5.2: Histograms showing the Temporal Locations of All Computed Silence Pause Ends Vs the Ground-Truth, for the $D_{eterm}$Classifier:

The x-axis shows the temporal variation in the computed silence pause ends versus the ground-truth as allocated to 20ms bins, and the y-axis shows the number of instances of speech to silence transitions within each bin. The temporal locations of the speech to silence transitions from the ground-truth are all represented by the centre 0 co-ordinate on the x-axis.

non-speech event occurs during a silence pause.

To assess the effect of the first type of classification failure (where the effect of the error would be perceived by a listener), the location of all detected silence pauses were set to zero in a copy of the filtered speech waveform; and thereafter the two speech samples (the original filtered speech and the now gated speech) were compared. To add a degree of rigour to what is effectively a subjective process, all silence termination errors that were less than $20ms$ late were ignored,

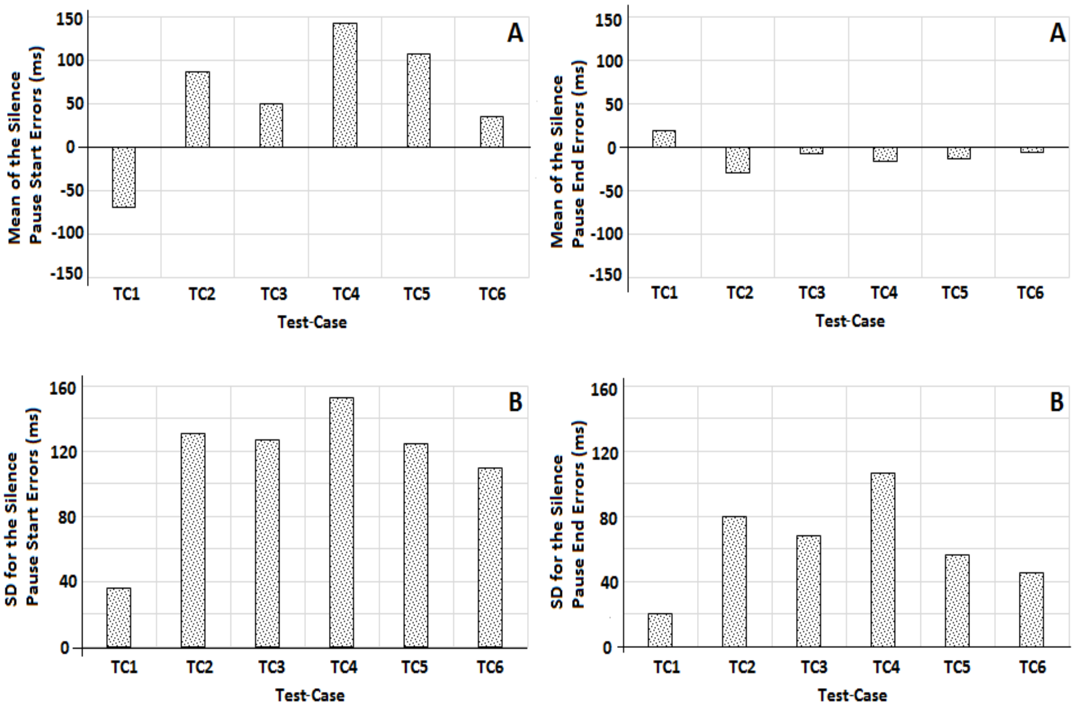


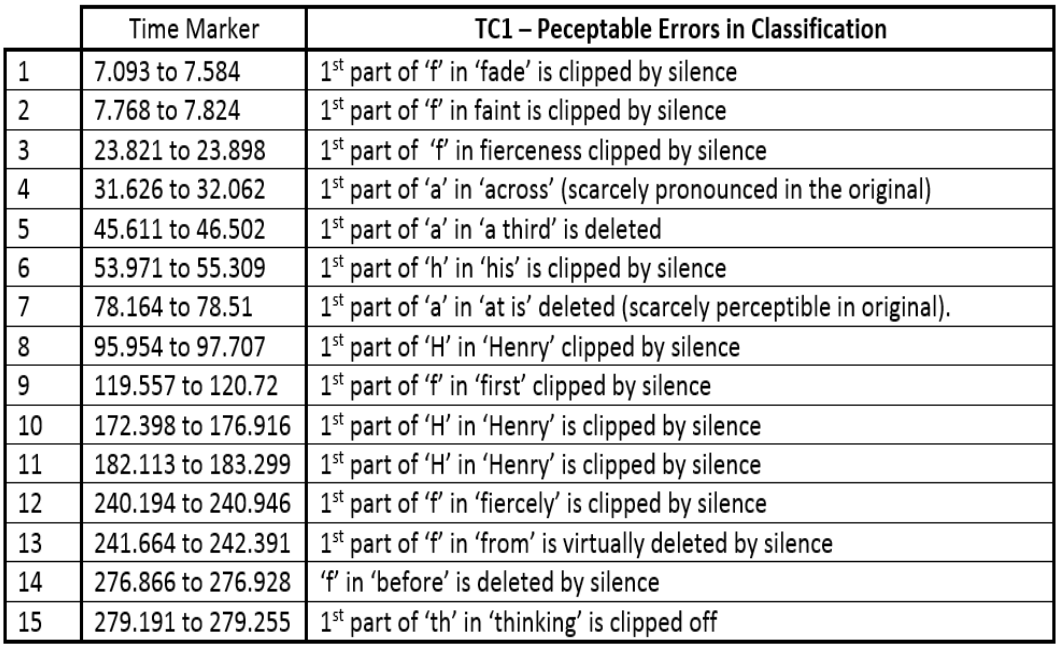
Figure 5.3: The Mean of, and SD for, the Errors about a Datum for the D$_{eterm}$Classifier

Figures 'A' above provide the Arithmetic Means of the silence pause errors for all test-cases, and Figures 'B' provide the sample Standard Deviations for all of the test-cases. For all Figures, the datum at '0' represents the ground-truth.

and for those more than $20ms$ late, short sections of the two samples were compared by being replayed alternately (when necessary). Table 5.6 provides details of the significant errors in classification.

TC1 was the only test-case of the six for which the first type of classification failures occurred; 14 of the 15 errors (see Table 5.6) taking the form of silence insertion at speech onset, and resulting in the clipping of speech. The remaining classification failure, the virtual deletion of the 'f' in 'before' (Entry 14 in Table 5.6), is scarcely perceptible.

The effect of the second type of classification failure (where the algorithm has failed to identify silence pauses, wholly or partially, because of noise during the silence is illustrated in Figures 5.4 and 5.5 — where the D$_{eterm}$Classifier has failed to identify the full extent of the silence pauses because of inhalation frication during the pause. This type of classification failure accounts for the majority of the outliers indicated in Figure 5.1 and Figure 5.2.



| | Time Marker | TC1 – Peceptable Errors in Classification |
|---|---|---|
| 1 | 7.093 to 7.584 | $1^{st}$ part of ‘f’ in ‘fade’ is clipped by silence |
| 2 | 7.768 to 7.824 | $1^{st}$ part of ‘f’ in faint is clipped by silence |
| 3 | 23.821 to 23.898 | $1^{st}$ part of ‘f’ in fierceness clipped by silence |
| 4 | 31.626 to 32.062 | $1^{st}$ part of ‘a’ in ‘across’ (scarcely pronounced in the original) |
| 5 | 45.611 to 46.502 | $1^{st}$ part of ‘a’ in ‘a third’ is deleted |
| 6 | 53.971 to 55.309 | $1^{st}$ part of ‘h’ in ‘his’ is clipped by silence |
| 7 | 78.164 to 78.51 | $1^{st}$ part of ‘a’ in ‘at is’ deleted (scarcely perceptible in original). |
| 8 | 95.954 to 97.707 | $1^{st}$ part of ‘H’ in ‘Henry’ clipped by silence |
| 9 | 119.557 to 120.72 | $1^{st}$ part of ‘f’ in ‘first’ clipped by silence |
| 10 | 172.398 to 176.916 | $1^{st}$ part of ‘H’ in ‘Henry’ is clipped by silence |
| 11 | 182.113 to 183.299 | $1^{st}$ part of ‘H’ in ‘Henry’ is clipped by silence |
| 12 | 240.194 to 240.946 | $1^{st}$ part of ‘f’ in ‘fiercely’ is clipped by silence |
| 13 | 241.664 to 242.391 | $1^{st}$ part of ‘f’ in ‘from’ is virtually deleted by silence |
| 14 | 276.866 to 276.928 | ‘f’ in ‘before’ is deleted by silence |
| 15 | 279.191 to 279.255 | $1^{st}$ part of ‘th’ in ‘thinking’ is clipped off |

Table 5.6: The D$_{eterm}$Classifier: TC1 Errors in Classification

A third form of classification error is when a substantial audible non-speech event occurs during a silence pause. Figure 5.6 illustrates one such example — the loud rustling of paper.

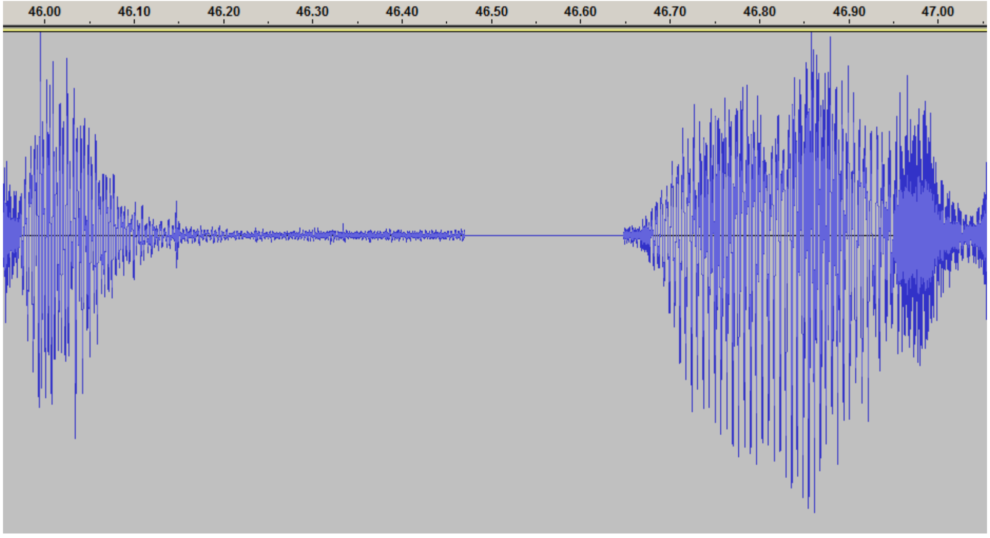


Figure 5.4: The D$_{eterm}$Classifier: Late Silence Pause Start Detection

Taken from TC2, this is the waveform with the temporal location of the silence pause as detected by the D$_{eterm}$Classifier set to true zero. The D$_{eterm}$Classifier has failed to reject as noise the inhalation fricative which immediately precedes the detected silence pause. As the detected silence pause is shorter than $250ms$, it will not be used for the generation of the silence or speech training data for the DNN.

The purpose of the initial classification is to establish a method for reliably identifying periods of speech and periods of silence so that the LogFB$_{dynamic}$ model

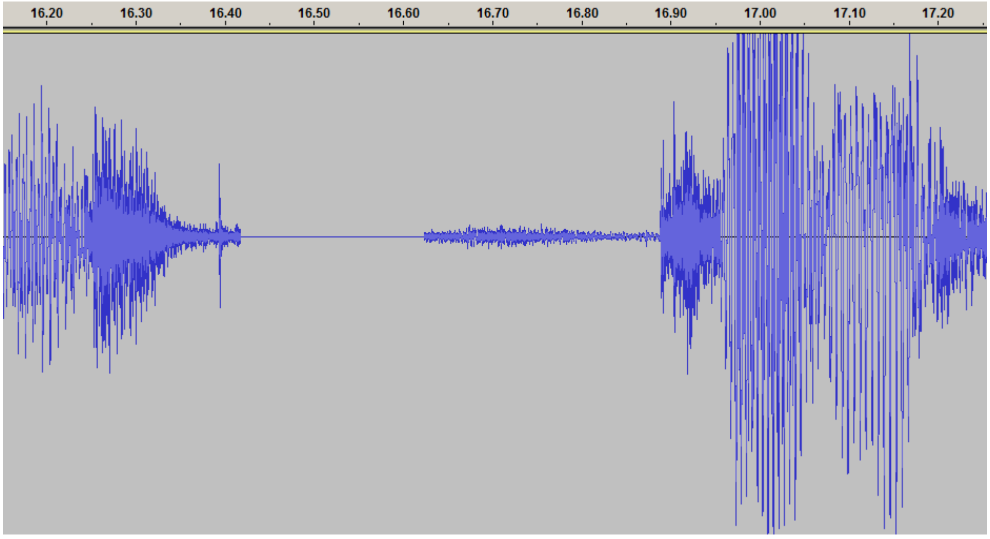


Figure 5.5: The D$_{eterm}$Classifier: Early Silence Pause End Detection

Taken from TC2, this is the waveform with the temporal location of the silence pause as detected by the D$_{eterm}$Classifier set to true zero. The D$_{eterm}$Classifier has failed to reject as noise the inhalation fricative which immediately follows the detected silence pause. As the detected silence pause is shorter than $250ms$, it will not be used for the generation of the silence or speech training data for the DNN.

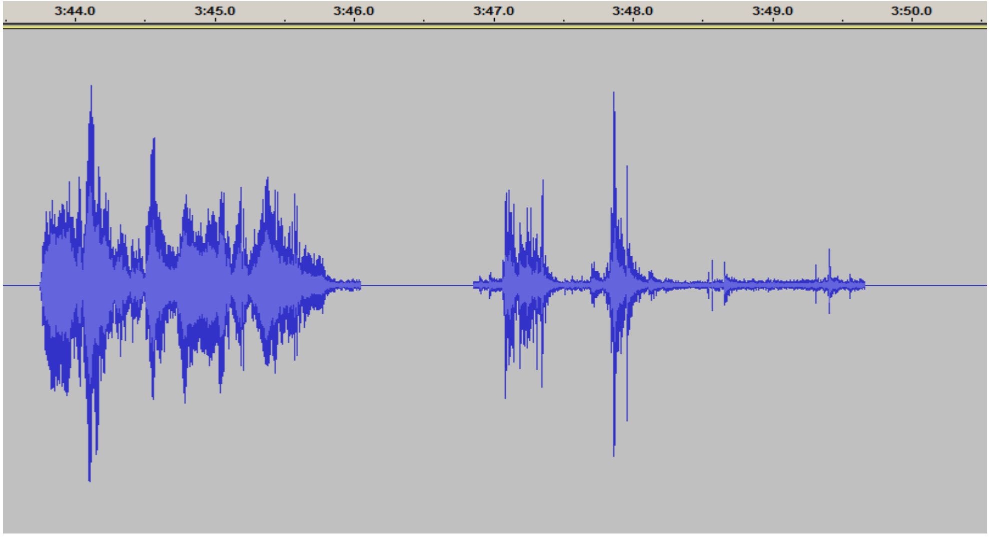
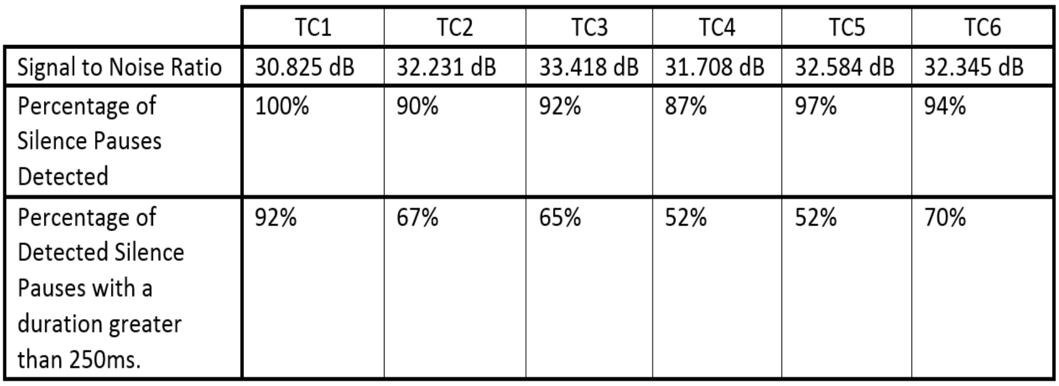


Figure 5.6: The D$_{eterm}$Classifier: Erroneous Detection of ‘Speech’

Taken from TC3, this shows the final block of text, “My anxiety was to gain real knowledge of the earth.” followed by a silence pause, and then the loud rustling of paper. The D$_{eterm}$Classifier has failed to reject this as noise because of the energy in the noise.

data segments can be attributed as either speech or as silence DNN training data (as shown in Figure 4.4). From Figures 5.4 and 5.5, it can be seen that under certain conditions the detected silence pauses do not provide a satisfactory classification, and to mitigate this, only Silence Pauses that are longer than $250ms$ are used in the generation of the DNN training Data (Table 5.7).

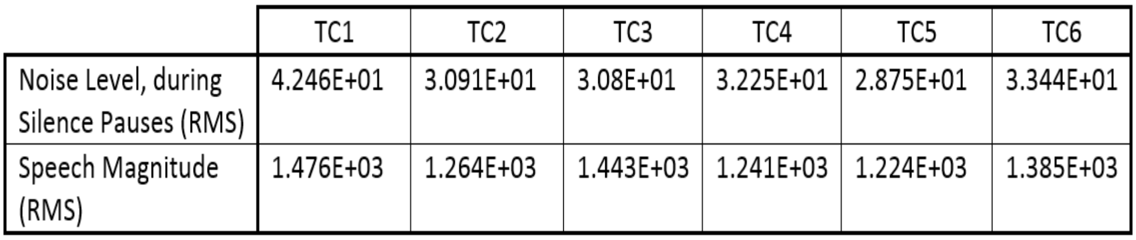


| | TC1 | TC2 | TC3 | TC4 | TC5 | TC6 |
|---|---|---|---|---|---|---|
| Signal to Noise Ratio | 30.825 dB | 32.231 dB | 33.418 dB | 31.708 dB | 32.584 dB | 32.345 dB |
| Percentage of Silence Pauses Detected | 100% | 90% | 92% | 87% | 97% | 94% |
| Percentage of Detected Silence Pauses with a duration greater than 250ms. | 92% | 67% | 65% | 52% | 52% | 70% |

For each of the test-cases, the percentage of Detected Silence Pauses with a duration of greater than $250ms$ is shown. Silence pauses with a duration of less than $250ms$ are not used in the creation of the silence and speech supervised training data (see Figure 4.4).

Table 5.7: The D$_{eterm}$Classifier: Detected Silence Pauses $>= 250ms$

It has previously been shown that the type of speech sample — whether narrative or conversational — can affect the accuracy of the estimate of the number of silence pauses from the punctuation in the text; and it can be seen from Table 5.8 that TC1 had the highest noise level during silence of the 6 test-cases. Within bounds though, the noise level during silence does not dominate the final speech silence threshold, because it is multiplied-up by the control loop at Step 7 of the D$_{eterm}$Classifier algorithm (see Table 5.2), until the pause count computed by the D$_{eterm}$Classifier is equal to or greater than the pause count computed from the punctuation in the text.



| | TC1 | TC2 | TC3 | TC4 | TC5 | TC6 |
|---|---|---|---|---|---|---|
| Noise Level, during Silence Pauses (RMS) | 4.246E+01 | 3.091E+01 | 3.08E+01 | 3.225E+01 | 2.875E+01 | 3.344E+01 |
| Speech Magnitude (RMS) | 1.476E+03 | 1.264E+03 | 1.443E+03 | 1.241E+03 | 1.224E+03 | 1.385E+03 |

For each of the test-cases, the RMS magnitudes measured for the signal to noise ratio calculation are shown. The D$_{eterm}$Classifier automatically selects the optimum speech/silence threshold for the noise level which prevails throughout the speech sample. Despite being recorded in the same environment with the same recording equipment — though at a different date and time — it can be seen that the ‘Noise Level, during Silence Pauses’ for TC1 is higher than for the other test-cases.

Table 5.8: Empirically Derived Silence and Speech RMS Magnitudes for Signal to Noise Ratio Calculation.

It is seemingly anomalous that for the test-case with the lowest signal to noise ratio and highest noise level during silence that all perceived pauses and 22 silence insertions are identified; but because both the amplitude of speech and the rate of change of amplitude of speech are modulated, then inevitably any increase in the silence/speech threshold will result in the detection of more silence. Thus increased noise during 'silence' in the speech signal may result in an apparent improvement to the level of silence identification, but also a reduction in the temporal accuracy of the $\mathrm{D}_{eterm}$Classifier.

### 5.3.4 Summary

For the six test-cases, the $\mathrm{D}_{eterm}$Classifier correctly identified 87% or more of the perceived silence pauses (Table 5.3). That up to 13% of pauses were not identified was largely due to the presence of inhalation fricatives during the silences resulting in periods of actual silence during the pauses that were too short to be recognised by the $\mathrm{D}_{eterm}$Classifier.

The question, "Is it possible to differentiate between silence and unvoiced speech based upon the energy in the waveform?", was posed at the start of Section 4. With a silence pause detection rate of at least 87%, the answer to this question is that it is possible so to do, at least to a useful extent. That is, whilst the $\mathrm{D}_{eterm}$Classifier would not be satisfactory as a speech/silence classifier it may be good enough to generate the supervised training data required for the DNN_Classifier. i.e. The sampling method used to build the silence and the speech training data for the DNN_Classifier requires knowledge of known silence pauses, not all silence pauses. This is because the sampling process is active only in and around the location of the known silence pauses (as described in Figure 4.4).

In listening tests, the speech reproduction of the filtered speech samples with detected silences reset to zero was found to be acceptable. For TC2–TC6 inclusive, most of the significant errors in classification involve either the late detection of the start of a silence pause or the early detection of the end of a silence pause and these have no effect on the quality of the speech. For TC1, which was the only test-case with speech erroneously classified as silence, there was at times just a hint of the choppy speech onsets that were experienced with some early examples of mobile phones.

An inadequacy in the method of obtaining the set point for the $\mathrm{D}_{eterm}$Classifier control loop was evident from the results, where the silence start and end error distributions for TC1 were significantly different to those of TC2 to TC6. Whether this will impact on the final classification will be evaluated in the following section.

The effect of the classification failures discussed in the Section 5.3.3 must result in some corruption of the training data where silence may be incorrectly classified as speech and vice versa. If the corrupted training data is only some tiny percentage of the total training data, then it should not present a problem, because the DNN_Classifier is configured to prevent over-fitting. That is, the effects of the bad classification data may be swamped by the weight of the correct classification data. (Over-fitting is where a classifier is trained to the point at which every nuance of the training data is accommodated; and may result in an excellent classification performance on the training data set and a poor classification performance on other data of the same class (da Silva et al. [2017])).

The $\mathrm{D}_{eterm}$Classifier, within the context of this work, is purposed only to provide speech and silence training data for the DNN_Classifier, and the extent to which it is successful in achieving this will be evaluated in the following section.

## 5.4 The DNN_Classifier

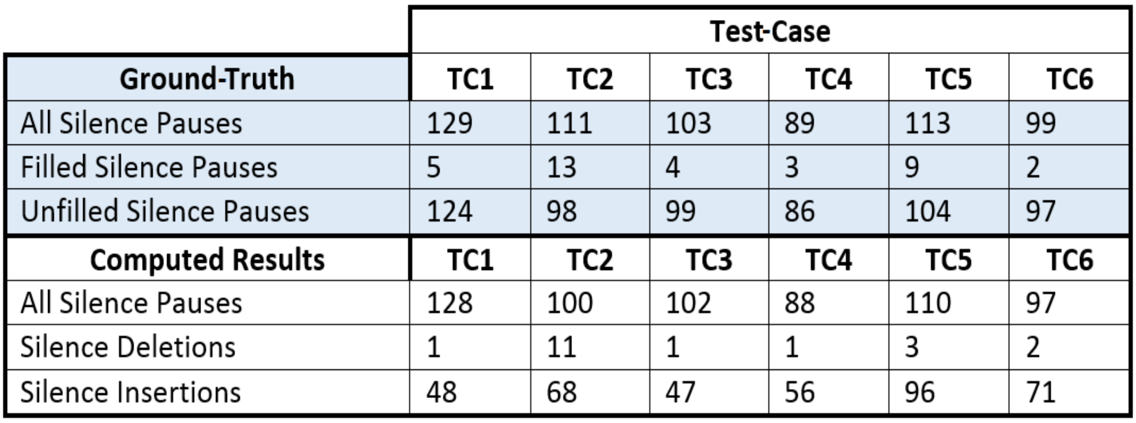
Following the rationale given in Section 4.8.1, the configuration for the DNN_Classifier was fixed with an input layer of 27 neurones, 2 hidden layers — each of 27 neurones, and a single output neurone.

As with the $\mathrm{D}_{eterm}$Classifier , a basic measure of the performance of the DNN_Classifier was obtained by comparing the Silence Pause counts with the ground-truth; and the results of this comparison are considered in Section 5.4.1. Thereafter, in Section 5.4.2 the consolidated results of a detailed comparison of the accuracy of the DNN_Classifier versus the ground-truth are considered; and in Section 5.4.3 classification failures are dealt with in detail. Note that the source data for the ground-truth, and the raw classification results are not reproduced herein.

### 5.4.1 Basic Performance

Table 5.9 shows the Silence Pause counts from the ground-truth Vs the Silence Pause counts from the computed results for the DNN_Classifier. Note that in Table 5.9 — *Silence Insertions*, a single count is used to represent what may be a cluster of short silence bursts in the same location.

For all test-cases 100% of the unfilled silence pauses as specified in the ground-truth were detected, and for 4 of the 6 test-cases some of the filled silence pauses were detected — though for some of these the indicated pauses were only a few milliseconds in length.



| | **Test-Case** | | | | | |
|---|---|---|---|---|---|---|
| **Ground-Truth** | **TC1** | **TC2** | **TC3** | **TC4** | **TC5** | **TC6** |
| All Silence Pauses | 129 | 111 | 103 | 89 | 113 | 99 |
| Filled Silence Pauses | 5 | 13 | 4 | 3 | 9 | 2 |
| Unfilled Silence Pauses | 124 | 98 | 99 | 86 | 104 | 97 |
| **Computed Results** | **TC1** | **TC2** | **TC3** | **TC4** | **TC5** | **TC6** |
| All Silence Pauses | 128 | 100 | 102 | 88 | 110 | 97 |
| Silence Deletions | 1 | 11 | 1 | 1 | 3 | 2 |
| Silence Insertions | 48 | 68 | 47 | 56 | 96 | 71 |

'Silence Insertions' are short periods of silence that may not be coincident with the Unfilled Silence Pauses. For example such as may occur at the 'Filled Silence Pauses' or at unvoiced stops, such as /p/, /t/, and /k/.

Table 5.9: The DNN_Classifier: Silence Pause Detection, Insertion and Deletion Totals

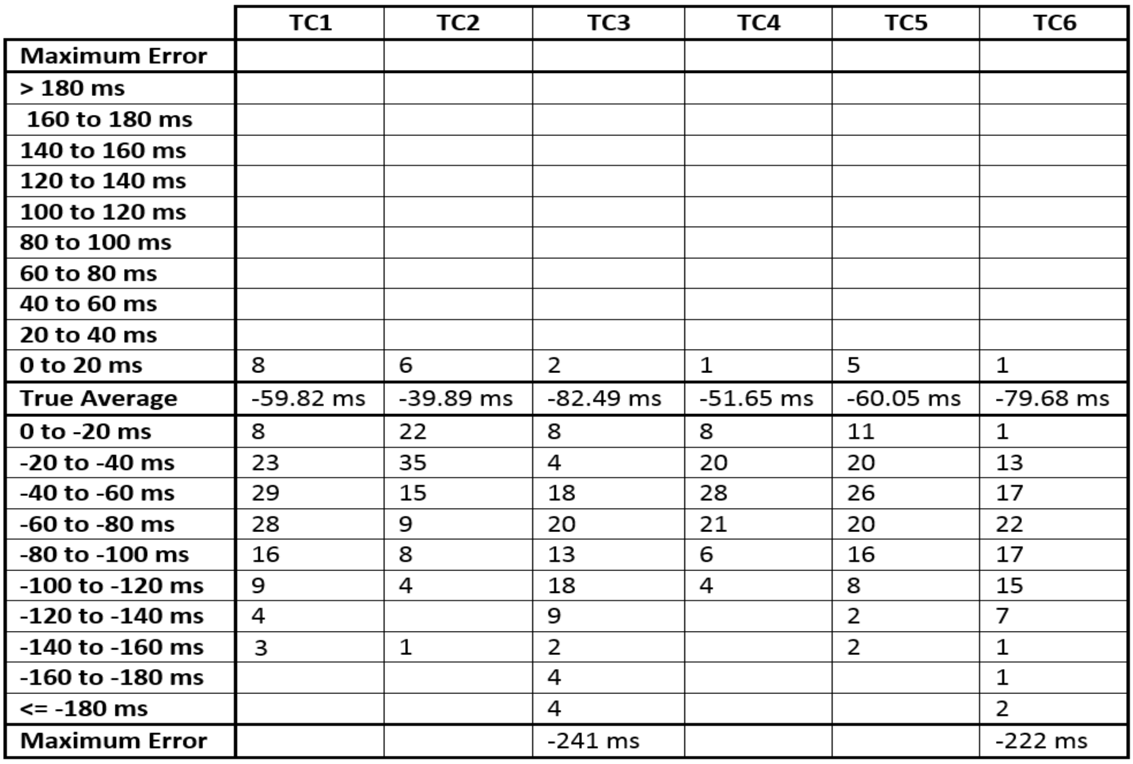
### 5.4.2 Detailed Performance Analysis

Table 5.10 shows the result of a comparison of the accuracy of the detected silence pause start times for the DNN_Classifier, with the ground-truth; and Table 5.11 shows the result of a comparison of the accuracy of the detected silence pause end times for the DNN_Classifier, also with the ground-truth. A negative error indicates early detection of the silence pause, and a positive error late detection. For clarity, the results are also provided as histograms (Figures 5.7 and 5.8 ). In addition, the Arithmetic Means, and Sample Standard Deviations for the distributions are shown in Figure 5.9.

A degree of systematic behaviour by the DNN_Classifier is evident for the distributions for the silence pause start and end errors shown in Figures 5.7 and 5.8, where the results are clustered with few outliers. It is evident that the main errors in detecting the silence pause starts are associated with early detection, and for silence pause ends are associated with late detection. That is, the DNN_Classifier is overestimating the duration of most of the silence pauses; but it is also evident that for all test-cases, that the distribution of the silence pause start and end errors is approximately normal. The errors in classification summarised in Table 5.13 are for the main part the result of this temporal inaccuracy.

From Figure 4.4, 'Speech and Silence Selection for Supervised Training' it may be expected that the ratio between the speech training data and the silence training data is of the order of 2 to 1, and Table 5.12, 'Duration of the Speech and Silence Training Data' indicates that the ratios are slightly greater than this — because a percentage of the silence pauses are shorter than $500ms$.

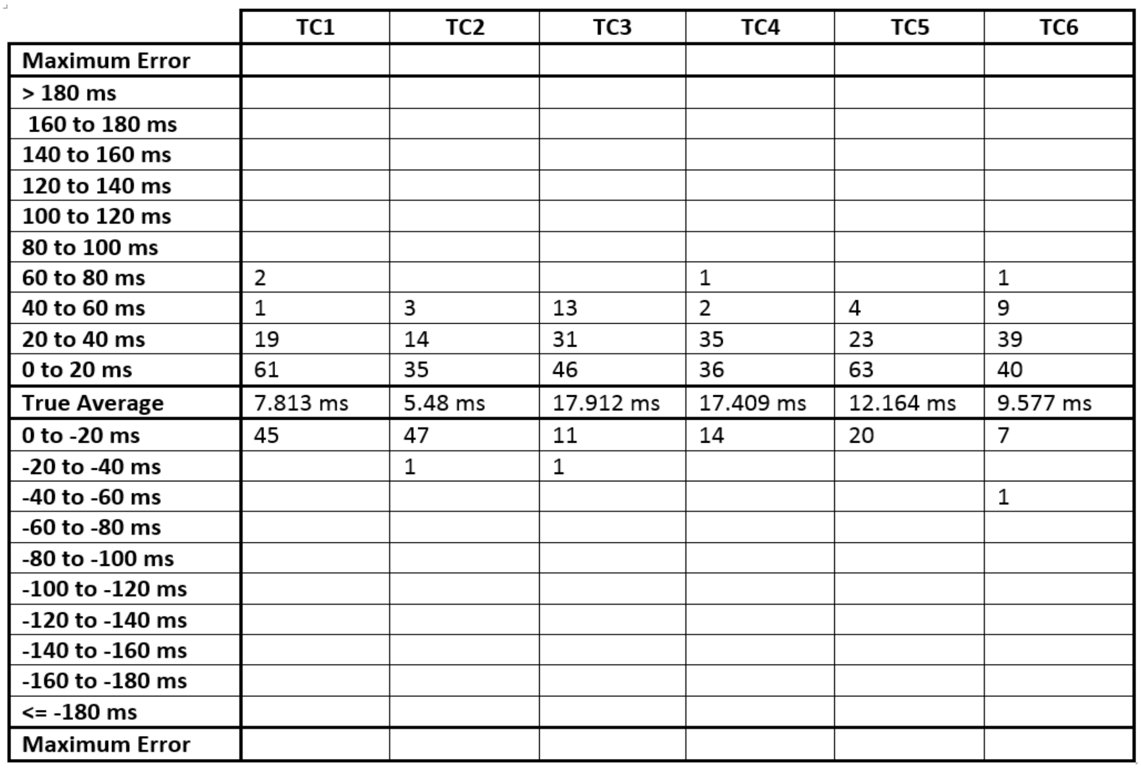
Figure 5.10 shows an area of detected silence in TC4, and this illustrates a disadvantage with the DNN_Classifier — where it becomes necessary to investigate short fragments of the waveform to assess whether they are speech or silence. This problem was addressed for the $\mathrm{D}_{eterm}$Classifier, where short fragments were compared with a clause prototype and either accepted as being speech, or rejected



| | TC1 | TC2 | TC3 | TC4 | TC5 | TC6 |
|---|---|---|---|---|---|---|
| Maximum Error | | | | | | |
| > 180 ms | | | | | | |
| 160 to 180 ms | | | | | | |
| 140 to 160 ms | | | | | | |
| 120 to 140 ms | | | | | | |
| 100 to 120 ms | | | | | | |
| 80 to 100 ms | | | | | | |
| 60 to 80 ms | | | | | | |
| 40 to 60 ms | | | | | | |
| 20 to 40 ms | | | | | | |
| 0 to 20 ms | 8 | 6 | 2 | 1 | 5 | 1 |
| True Average | -59.82 ms | -39.89 ms | -82.49 ms | -51.65 ms | -60.05 ms | -79.68 ms |
| 0 to -20 ms | 8 | 22 | 8 | 8 | 11 | 1 |
| -20 to -40 ms | 23 | 35 | 4 | 20 | 20 | 13 |
| -40 to -60 ms | 29 | 15 | 18 | 28 | 26 | 17 |
| -60 to -80 ms | 28 | 9 | 20 | 21 | 20 | 22 |
| -80 to -100 ms | 16 | 8 | 13 | 6 | 16 | 17 |
| -100 to -120 ms | 9 | 4 | 18 | 4 | 8 | 15 |
| -120 to -140 ms | 4 | | 9 | | 2 | 7 |
| -140 to -160 ms | 3 | 1 | 2 | | 2 | 1 |
| -160 to -180 ms | | | 4 | | | 1 |
| <= -180 ms | | | 4 | | | 2 |
| Maximum Error | | | -241 ms | | | -222 ms |

*The True Average is the average of the raw data, not of the binned data.

Table 5.10: Silence Pause Start Error for the DNN_Classifier.



| | TC1 | TC2 | TC3 | TC4 | TC5 | TC6 |
|---|---|---|---|---|---|---|
| Maximum Error | | | | | | |
| > 180 ms | | | | | | |
| 160 to 180 ms | | | | | | |
| 140 to 160 ms | | | | | | |
| 120 to 140 ms | | | | | | |
| 100 to 120 ms | | | | | | |
| 80 to 100 ms | | | | | | |
| 60 to 80 ms | 2 | | | 1 | | 1 |
| 40 to 60 ms | 1 | 3 | 13 | 2 | 4 | 9 |
| 20 to 40 ms | 19 | 14 | 31 | 35 | 23 | 39 |
| 0 to 20 ms | 61 | 35 | 46 | 36 | 63 | 40 |
| True Average | 7.813 ms | 5.48 ms | 17.912 ms | 17.409 ms | 12.164 ms | 9.577 ms |
| 0 to -20 ms | 45 | 47 | 11 | 14 | 20 | 7 |
| -20 to -40 ms | | 1 | 1 | | | |
| -40 to -60 ms | | | | | | 1 |
| -60 to -80 ms | | | | | | |
| -80 to -100 ms | | | | | | |
| -100 to -120 ms | | | | | | |
| -120 to -140 ms | | | | | | |
| -140 to -160 ms | | | | | | |
| -160 to -180 ms | | | | | | |
| <= -180 ms | | | | | | |
| Maximum Error | | | | | | |

*The True Average is the average of the raw data, not of the binned data.

Table 5.11: Silence Pause End Error for the DNN_Classifier.

as noise. However, that technique — ultimately based upon a ratio of powers — has been shown to be less than satisfactory; so an improved method must be adopted to correctly apportion these fragments following the speech/silence

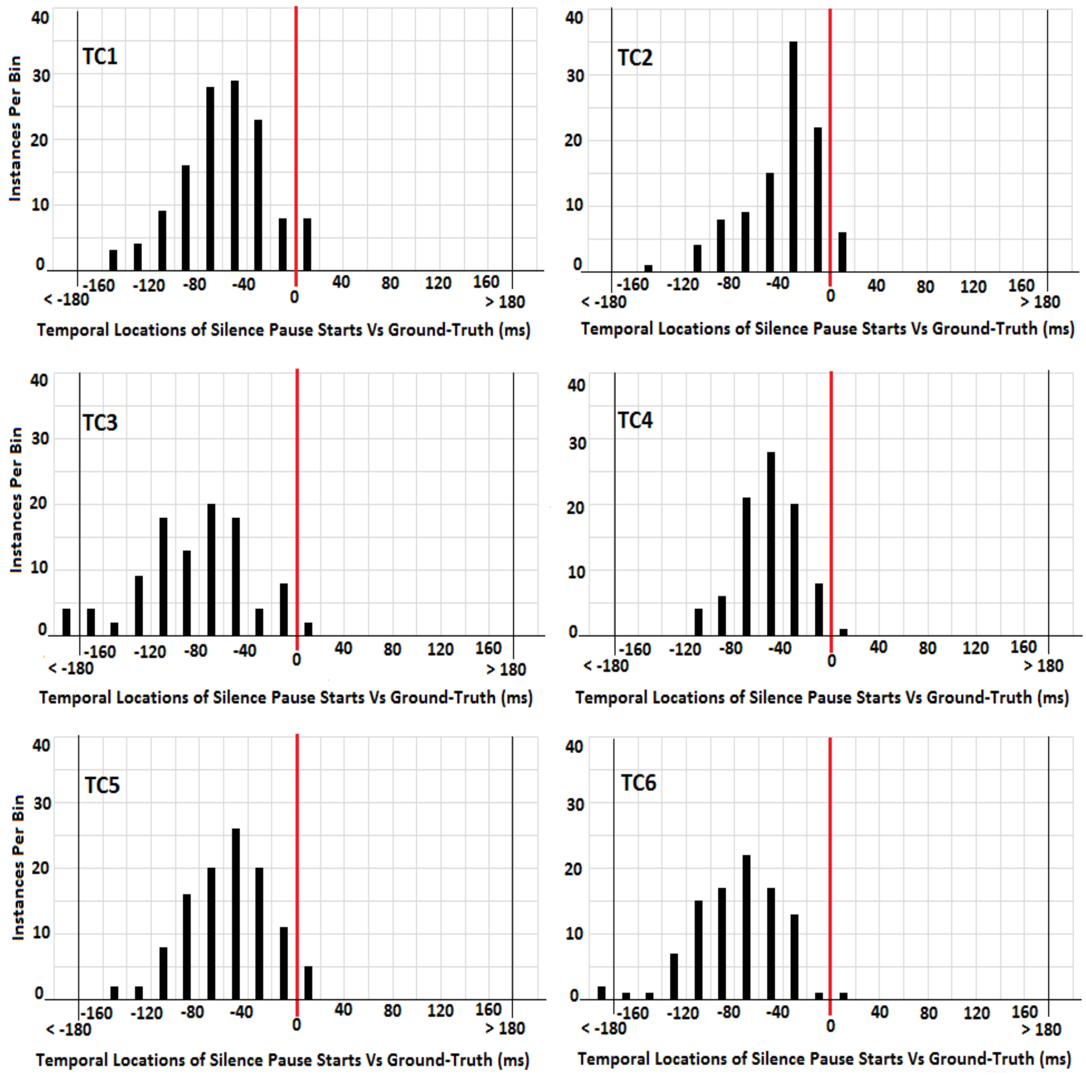


Figure 5.7: Histograms showing the Temporal Locations of All Computed Silence Pause Starts Vs the Ground-Truth, for the DNN_Classifier:

The x-axis shows the temporal variation in the computed silence pause starts versus the ground-truth as allocated to 20ms bins, and the y-axis shows the number of instances of speech to silence transitions within each bin. The temporal locations of the speech to silence transitions from the ground-truth are all represented by the centre '0' co-ordinate on the x-axis.

classification process. (As mentioned previously, Deng and O'Shaughnessy [2007] describe a different method for dealing with such fragments.)

For all of the test-cases, many silences additional to those perceived by the listener as silence pauses, were detected by the DNN_Classifier, and these insertions are consistently located in regions where one might expect to find them — for example at stops. On closer inspection a proportion of the insertions can be

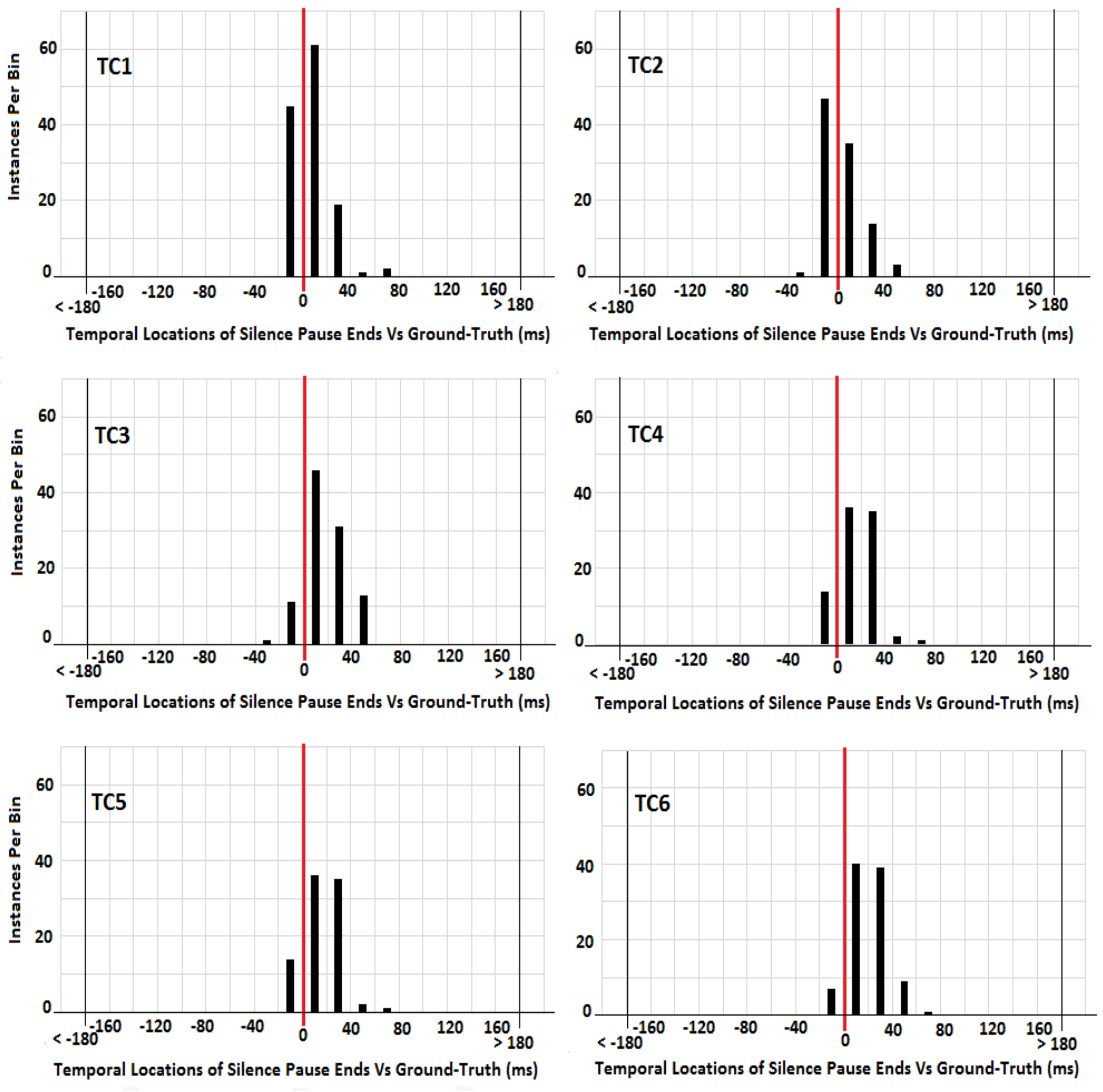
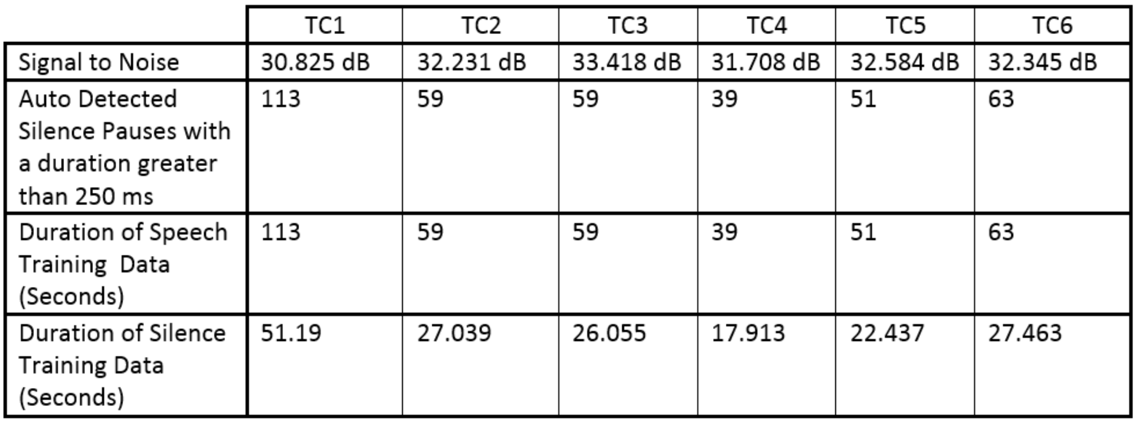


Figure 5.8: Histograms showing the Temporal Locations of All Computed Silence Pause Ends Vs the Ground-Truth, for the DNN_Classifier:

The x-axis shows the temporal variation in the computed silence pause ends versus the ground-truth as allocated to 20ms bins, and the y-axis shows the number of instances of speech to silence transitions within each bin. The temporal locations of the speech to silence transitions from the ground-truth are all represented by the centre ‘0’ co-ordinate on the x-axis.

seen to be positioned marginally earlier than might be expected (by up to $\sim 20ms$); though exactly why is not yet understood. It should be noted that despite the large number of insertions, silence is not detected for many of those speech elements which might be expected to have a brief moment of associated silence, and this is indicative of a natural variation in articulation. Examination of the pause duration data shows that the perceived pauses often had a much greater duration than the insertions — though not exclusively so. That is, the probability that a silence pause is a perceived pause declines — though not linearly — as the



| | TC1 | TC2 | TC3 | TC4 | TC5 | TC6 |
|---|---|---|---|---|---|---|
| Signal to Noise | 30.825 dB | 32.231 dB | 33.418 dB | 31.708 dB | 32.584 dB | 32.345 dB |
| Auto Detected Silence Pauses with a duration greater than 250 ms | 113 | 59 | 59 | 39 | 51 | 63 |
| Duration of Speech Training Data (Seconds) | 113 | 59 | 59 | 39 | 51 | 63 |
| Duration of Silence Training Data (Seconds) | 51.19 | 27.039 | 26.055 | 17.913 | 22.437 | 27.463 |

The $500ms$ before each silence pause and the $500ms$ after each silence pause are designated as speech training data resulting in 1 second of speech training data per pause; whereas either the first $500ms$ of each silence pause, or the length of the entire pause if it is shorter than $500ms$, is designated as silence training data. Notionally then, the number of auto detected silence pauses with a duration greater than $250ms$ should equal the total duration of the speech training data; and the total duration of the Silence training data should be half that of the speech training data. That the latter of these statements is not entirely correct is indicative that some of the silence pauses used for DNN training were shorter than $500ms$.

Table 5.12: Duration of the Speech and Silence Training Data

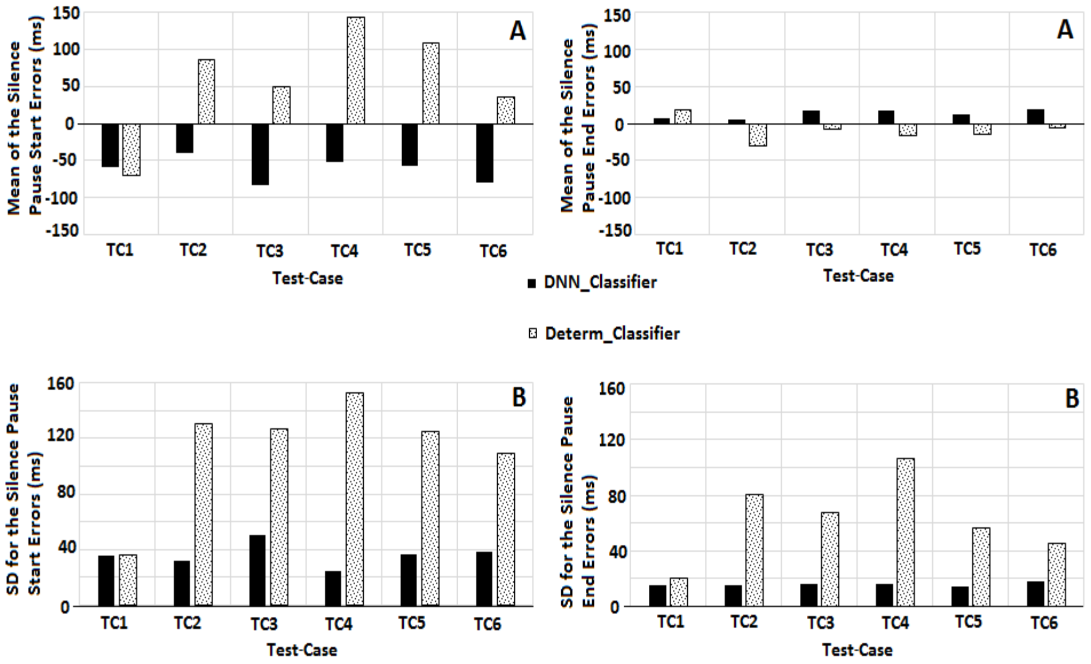


Figure 5.9: The Mean of, and SD for, the Errors about a Datum for the DNN_Classifier.

Figures ‘A’ provide the Arithmetic Means of the silence pause errors for all test-cases, and Figures ‘B’ provide the sample Standard Deviations for all of the test-cases. The results for the DNN_Classifier are indicated by the solid bars; and the datum at ‘0’ represents the ground-truth. For comparative purposes, the results for the D$_{eterm}$Classifier are also shown — as indicated by the patterned bars.

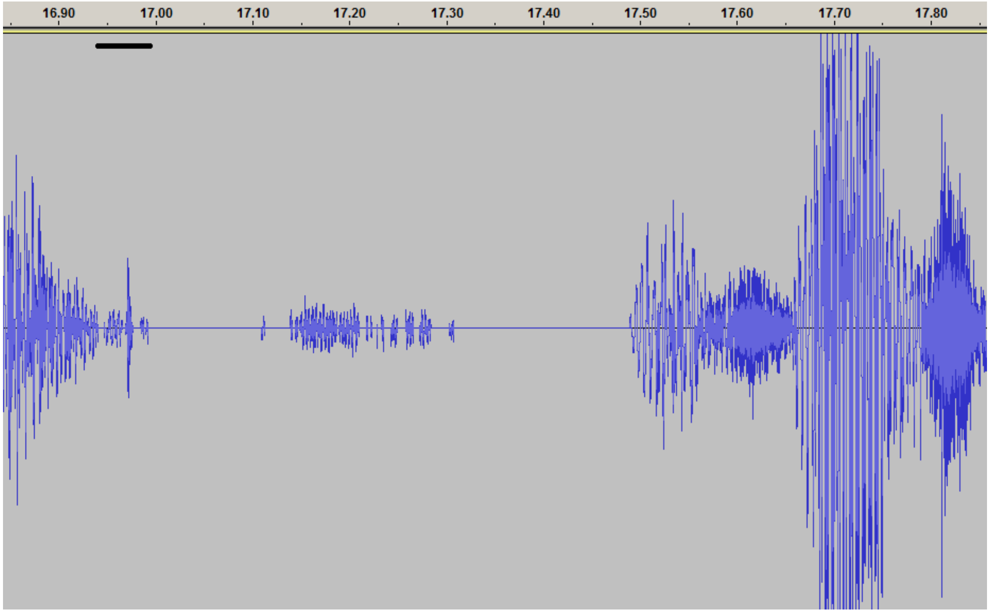


Figure 5.10: Fragmentation for the DNN_Classifier.

Taken from TC4, this shows the end of one block of text, the residue of an inhalation fricative and the start of the next block of text. It will be seen that several short silences are detected at the end of the first block — and this causes the classification issue of deciding where the first block should end. Any automated system must also be capable of assessing whether the inhalation fricative is silence, or speech.

duration of the pause reduces, such that it is not possible to make a highly accurate automatic classification of the very short pauses as either perceived or inserted, on the basis of duration alone.

From Figure 5.9, it can be seen that the start and end errors for TC1 for the DNN_Classifier do not exhibit the anomalous results obtained for TC1 for the $D_{eterm}$Classifier — the classification performance for TC1 is joint third for the silence pause start errors, and second for the silence pause end errors, i.e. for TC1, the degraded accuracy of the $D_{eterm}$Classifier did not propagate into the DNN_Classifier. This is confirmed to an extent by the data in Table 5.13, where the Perceived Classification Error Totals for the DNN_Classifier for TC1 are the lowest of all test-cases (although not all of the test-cases present an equal opportunity for classification errors).

In listening tests, the reproduction of the filtered speech samples with detected silences reset to zero — the gated speech — was acceptable; i.e. despite the significant errors in classification (Table 5.13), there is none of the 'choppy' speech onsets and endpoints that were experienced in the early days with some types of mobile phones. And although the start of silences is often detected earlier than may be expected, the effect of this is not audible in the gated speech.

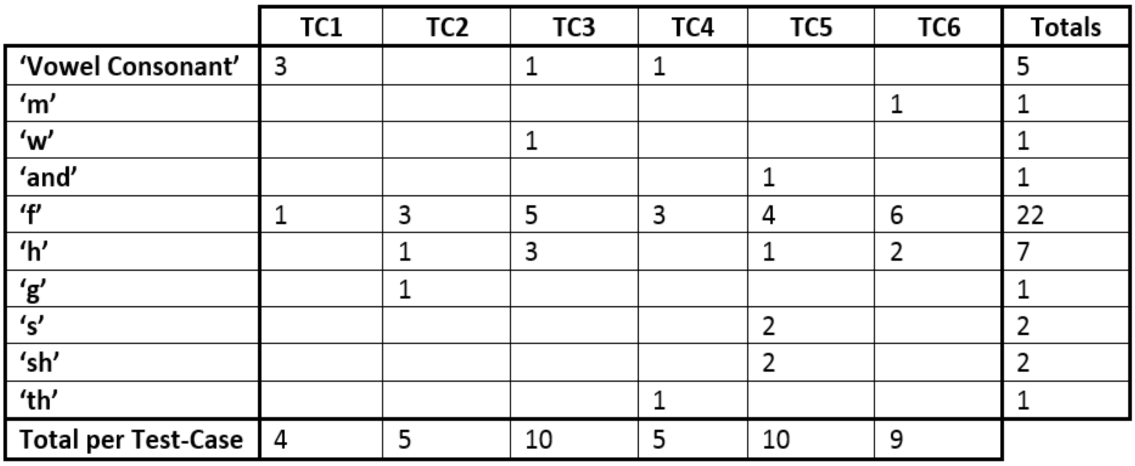
### 5.4.3 Classification Failures

One form of classification failure is where the effect of the error would be perceived by a listener, and these were assessed as described in Section 5.3.3.

Another form of error is where silence pauses, are wholly or partially not recognised as such, because of noise during the silence. Figure 5.11 shows where the DNN_Classifier has failed to identify the full extent of a silence pause because of inhalation frication during the pause.

A third form of classification error is when a substantial audible non-speech event occurs during a silence pause. Figure 5.12 shows the same 'rustling paper' event that was used to illustrate a non-audible speech event for the D$_{eterm}$Classifier. It should be noted that this is an extreme example, and is not typical of the silence pauses identified by the DNN_Classifier — which for the most part are clean.

The classification errors for the DNN_Classifier are collated in Table 5.13 — a total of forty three for the six test-cases, and the detail of the errors is provided in Tables 5.14 to 5.19 and Figures 5.13 to 5.24. Twenty two of these classification errors involved the unvoiced fricative 'f', and eight the unvoiced fricative 'h'. It may be that these fricatives were confused in the classification process with inhalation fricatives — some of which were partially classified as speech as can be seen in Figure 5.10. It is also relevant that by no means all instances of 'f' and 'h' were classified incorrectly, the more powerful examples being less susceptible to misclassification. Of the remainder of the classification errors, for 'g' and 'th' the situation is similar to that for 'f' and 'h', whereas the five vowel/consonant, the two 's' and the two 'sh' failures, and the 'm' and 'w' failures indicate speech like events during silence which are similar to events which must also be present to a very limited extent in speech. There was a suggestion in Section 4.6, that the bandwidth of the model could be reduced from 21 to 19 Bark bands, but fricatives and sibilants have a high-frequency component, and since the classification of /s/ and /f/ is already less than perfect it is not clear what effect (if any) the reduction of bandwidth would have on the detection of these consonants.



| | TC1 | TC2 | TC3 | TC4 | TC5 | TC6 | Totals |
|---|---|---|---|---|---|---|---|
| 'Vowel Consonant' | 3 | | 1 | 1 | | | 5 |
| 'm' | | | | | | 1 | 1 |
| 'w' | | | 1 | | | | 1 |
| 'and' | | | | | 1 | | 1 |
| 'f' | 1 | 3 | 5 | 3 | 4 | 6 | 22 |
| 'h' | | 1 | 3 | | 1 | 2 | 7 |
| 'g' | | 1 | | | | | 1 |
| 's' | | | | | 2 | | 2 |
| 'sh' | | | | | 2 | | 2 |
| 'th' | | | | 1 | | | 1 |
| Total per Test-Case | 4 | 5 | 10 | 5 | 10 | 9 | |

The 'Vowel Consonant' errors comprise short silence insertions in, 'ar', 'gl' and 'igh' for TC1, and in 'i' for TC3 and TC4. The 'and' that is deleted is scarcely audible in the original speech file.

Table 5.13: Perceived Classification Error Totals for the DNN_Classifier.

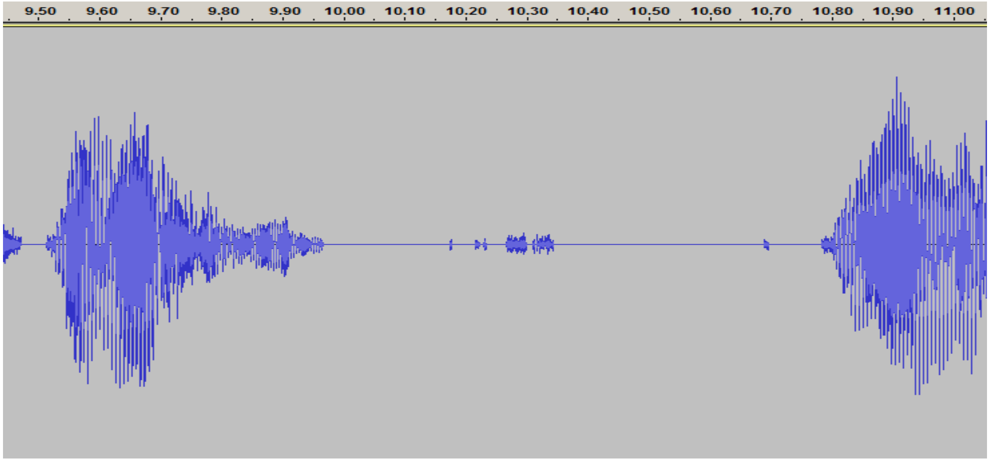


Figure 5.11: Failure of Silence Detection for the DNN_Classifier.

An inhalation fricative during the silence between two blocks of text has erroneously been classified by the DNN_Classifier as 'speech' (From TC4).

The waveforms for a selection of the errors are shown in Figures 5.14 to 5.24. The errors are represented in the waveforms by setting the signal magnitude to true zero; and these are highlighted by a bar that is drawn parallel to and time

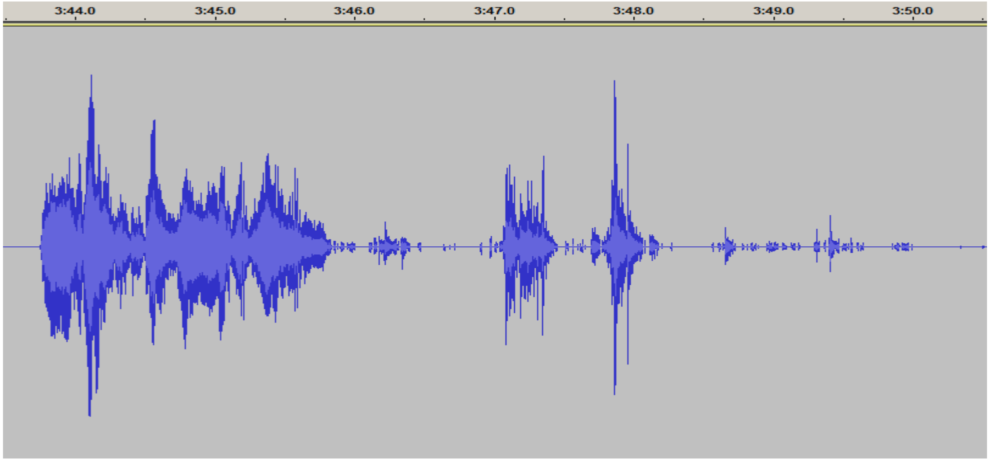
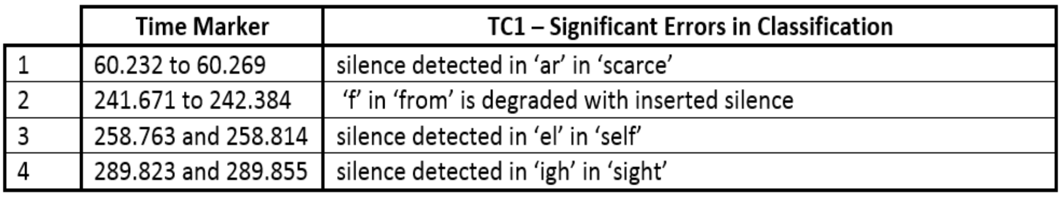


Figure 5.12: Erroneous Detection of ‘Speech’ for the DNN_Classifier.

Taken from TC3, this shows the final block of text, “My anxiety was to gain real knowledge of the earth.” followed by a ‘silence pause’, and then the loud rustling of paper. This is not typical of the silence pauses throughout the recording — which for the most part are free of incorrectly classified noise.

co-incident with the erroneous silence insertions. The start of those of the errors which occurred at speech onset, was obtained by a visual inspection of the filtered waveform.

### TC1 Classification Errors:



| | Time Marker | TC1 – Significant Errors in Classification |
|---|---|---|
| 1 | 60.232 to 60.269 | silence detected in ‘ar’ in ‘scarce’ |
| 2 | 241.671 to 242.384 | ‘f’ in ‘from’ is degraded with inserted silence |
| 3 | 258.763 and 258.814 | silence detected in ‘el’ in ‘self’ |
| 4 | 289.823 and 289.855 | silence detected in ‘igh’ in ‘sight’ |

Table 5.14: TC1 Errors in Classification for the DNN_Classifier.

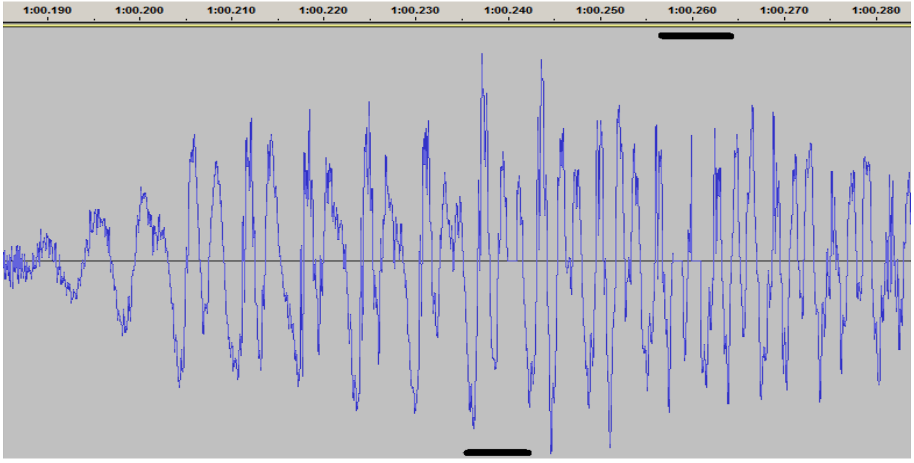


Figure 5.13: TC1 Classification Error 1 — During ‘ar’ in ‘scarce’

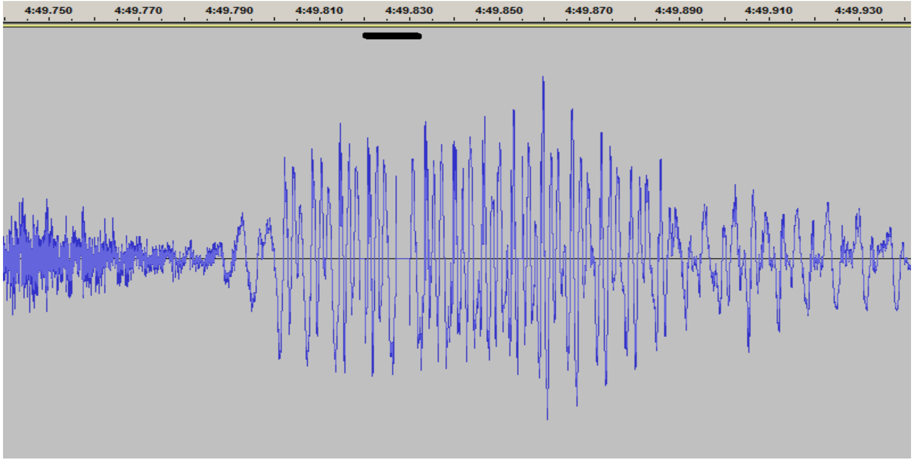
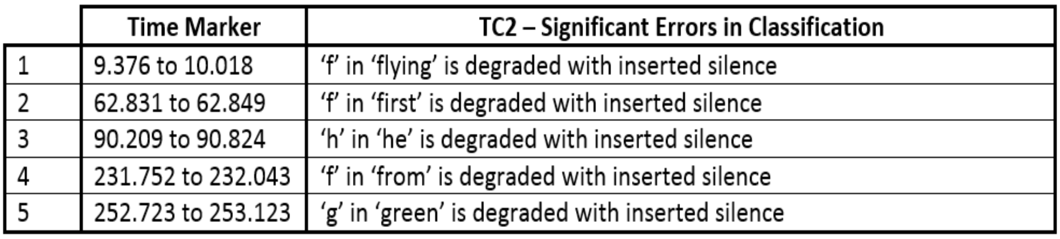


Figure 5.14: TC1 Classification Error 4 — During ‘igh’ in ‘sight’

**TC2 Classification Errors:**



|   | Time Marker | TC2 – Significant Errors in Classification |
|---|---|---|
| 1 | 9.376 to 10.018 | ‘f’ in ‘flying’ is degraded with inserted silence |
| 2 | 62.831 to 62.849 | ‘f’ in ‘first’ is degraded with inserted silence |
| 3 | 90.209 to 90.824 | ‘h’ in ‘he’ is degraded with inserted silence |
| 4 | 231.752 to 232.043 | ‘f’ in ‘from’ is degraded with inserted silence |
| 5 | 252.723 to 253.123 | ‘g’ in ‘green’ is degraded with inserted silence |

Table 5.15: TC2 Errors in Classification for the DNN_Classifier.

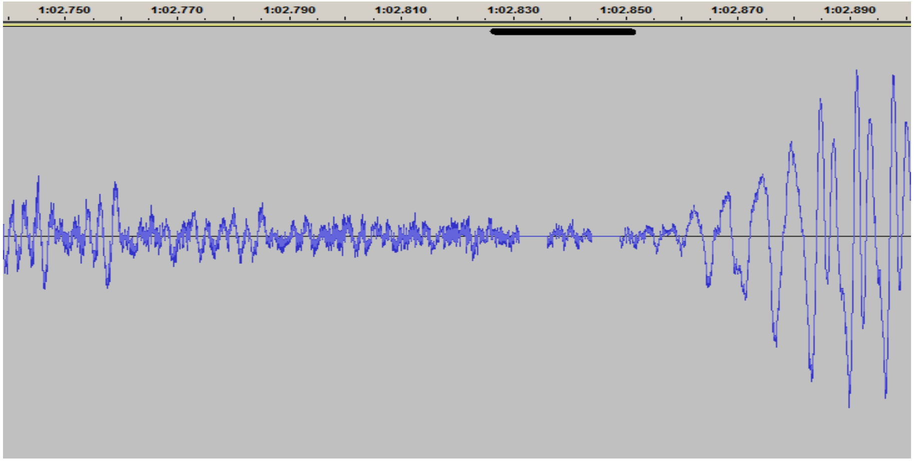


Figure 5.15: TC2 Classification Error 2 — During ‘f’ in ‘first’

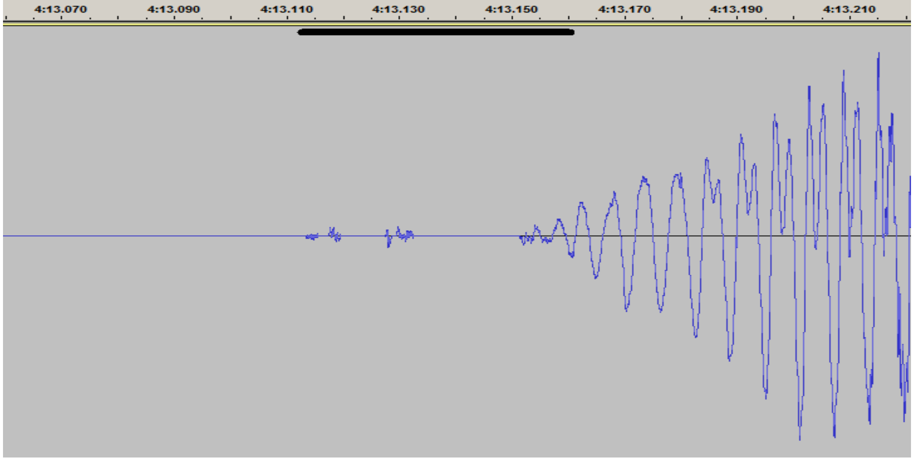
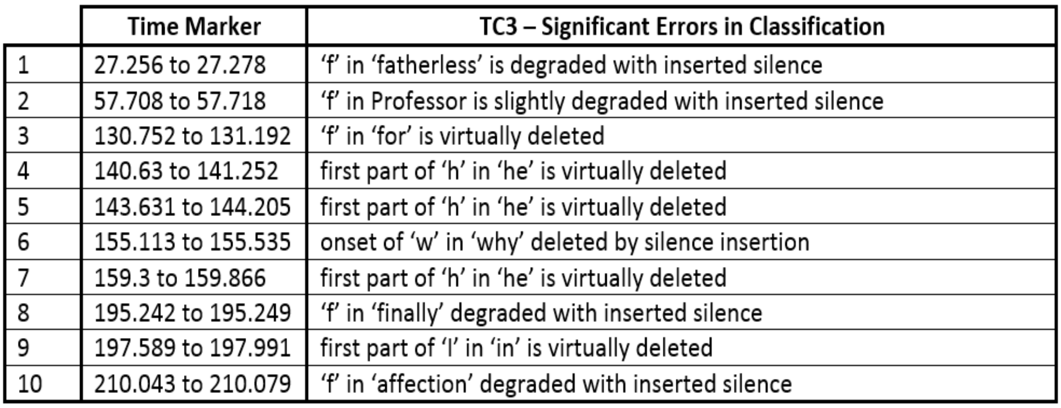


Figure 5.16: TC2 Classification Error 5 — During ‘g’ in ‘green’

### TC3 Classification Errors:



| | Time Marker | TC3 – Significant Errors in Classification |
|---|---|---|
| 1 | 27.256 to 27.278 | ‘f’ in ‘fatherless’ is degraded with inserted silence |
| 2 | 57.708 to 57.718 | ‘f’ in Professor is slightly degraded with inserted silence |
| 3 | 130.752 to 131.192 | ‘f’ in ‘for’ is virtually deleted |
| 4 | 140.63 to 141.252 | first part of ‘h’ in ‘he’ is virtually deleted |
| 5 | 143.631 to 144.205 | first part of ‘h’ in ‘he’ is virtually deleted |
| 6 | 155.113 to 155.535 | onset of ‘w’ in ‘why’ deleted by silence insertion |
| 7 | 159.3 to 159.866 | first part of ‘h’ in ‘he’ is virtually deleted |
| 8 | 195.242 to 195.249 | ‘f’ in ‘finally’ degraded with inserted silence |
| 9 | 197.589 to 197.991 | first part of ‘I’ in ‘in’ is virtually deleted |
| 10 | 210.043 to 210.079 | ‘f’ in ‘affection’ degraded with inserted silence |

Table 5.16: TC3 Errors in Classification for the DNN_Classifier.

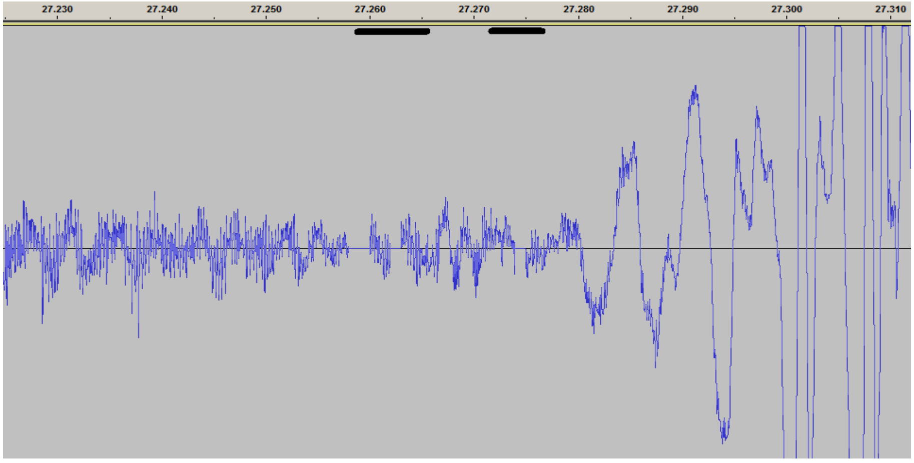


Figure 5.17: TC3 Classification Error 1 — During ‘f’ in ‘fatherless’

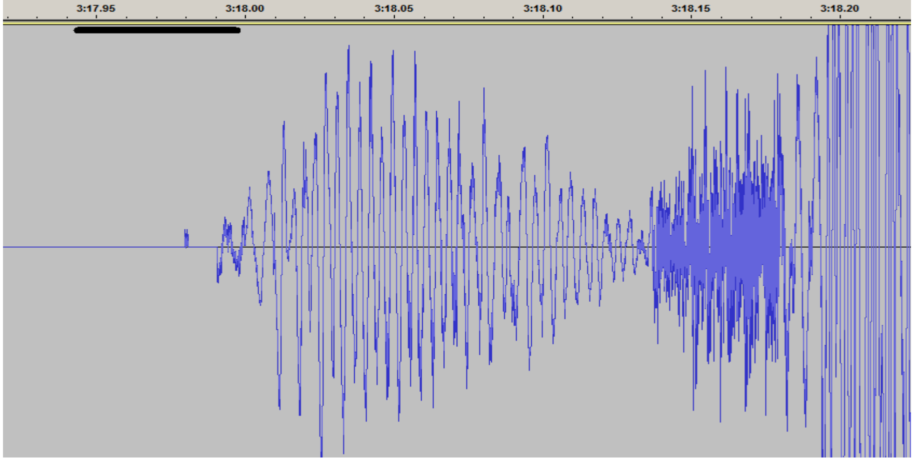
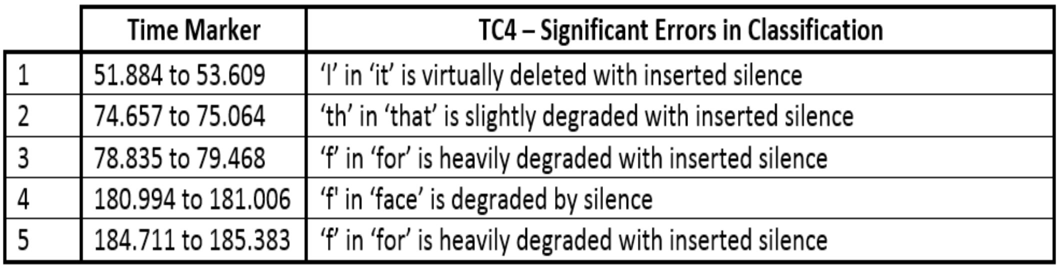


Figure 5.18: TC3 Classification Error 10 — During ‘i’ in ‘in’

### TC4 Classification Errors:



| | Time Marker | TC4 – Significant Errors in Classification |
|---|---|---|
| 1 | 51.884 to 53.609 | ‘I’ in ‘it’ is virtually deleted with inserted silence |
| 2 | 74.657 to 75.064 | ‘th’ in ‘that’ is slightly degraded with inserted silence |
| 3 | 78.835 to 79.468 | ‘f’ in ‘for’ is heavily degraded with inserted silence |
| 4 | 180.994 to 181.006 | ‘f' in ‘face’ is degraded by silence |
| 5 | 184.711 to 185.383 | ‘f’ in ‘for’ is heavily degraded with inserted silence |

Table 5.17: TC4 Errors in Classification for the DNN_Classifier.

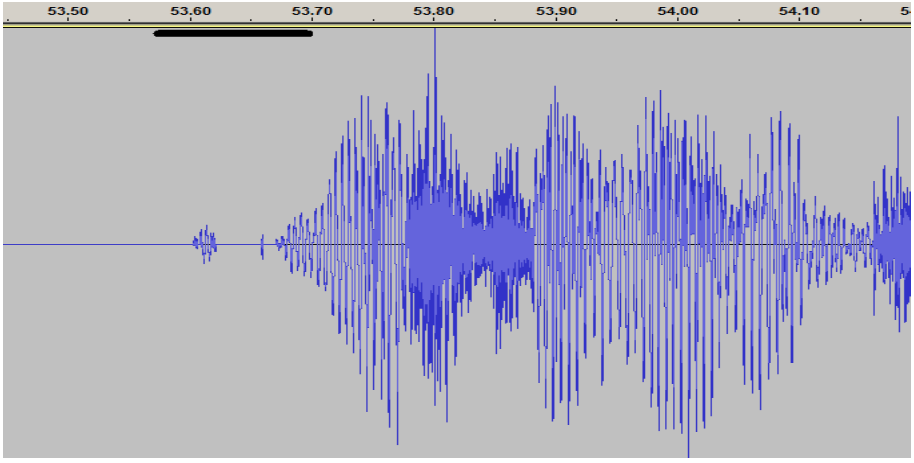


Figure 5.19: TC4 Classification Error 1 — During ‘i’ in ‘it’

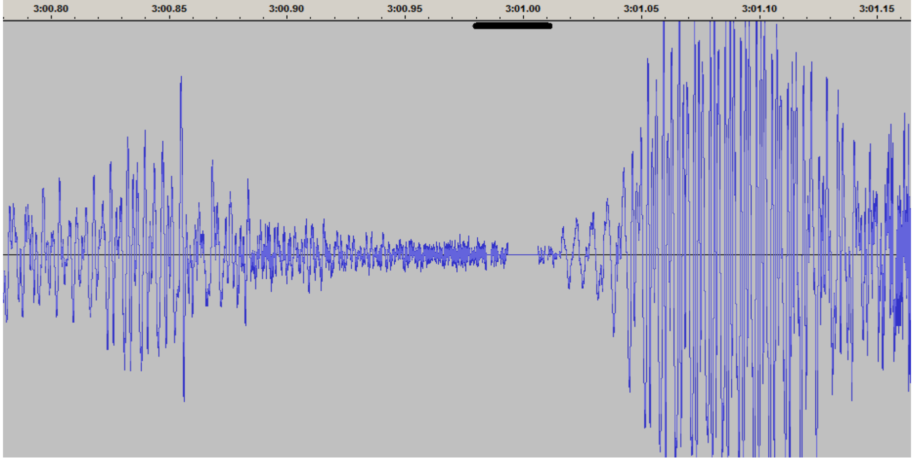
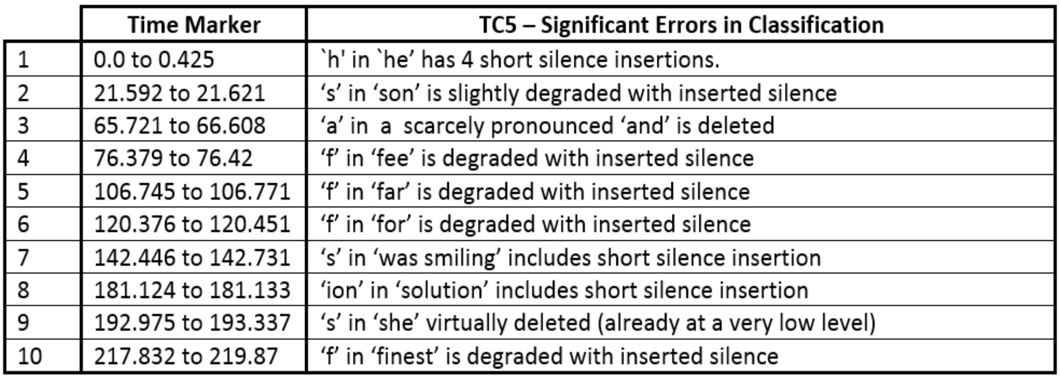


Figure 5.20: TC4 Classification Error 4 — During ‘f’ in ‘face’

### TC5 Classification Errors:



| | Time Marker | TC5 – Significant Errors in Classification |
|---|---|---|
| 1 | 0.0 to 0.425 | `h' in `he’ has 4 short silence insertions. |
| 2 | 21.592 to 21.621 | ‘s’ in ‘son’ is slightly degraded with inserted silence |
| 3 | 65.721 to 66.608 | ‘a’ in a scarcely pronounced ‘and’ is deleted |
| 4 | 76.379 to 76.42 | ‘f’ in ‘fee’ is degraded with inserted silence |
| 5 | 106.745 to 106.771 | ‘f’ in ‘far’ is degraded with inserted silence |
| 6 | 120.376 to 120.451 | ‘f’ in ‘for’ is degraded with inserted silence |
| 7 | 142.446 to 142.731 | ‘s’ in ‘was smiling’ includes short silence insertion |
| 8 | 181.124 to 181.133 | ‘ion’ in ‘solution’ includes short silence insertion |
| 9 | 192.975 to 193.337 | ‘s’ in ‘she’ virtually deleted (already at a very low level) |
| 10 | 217.832 to 219.87 | ‘f’ in ‘finest’ is degraded with inserted silence |

Table 5.18: TC5 Errors in Classification for the DNN_Classifier.

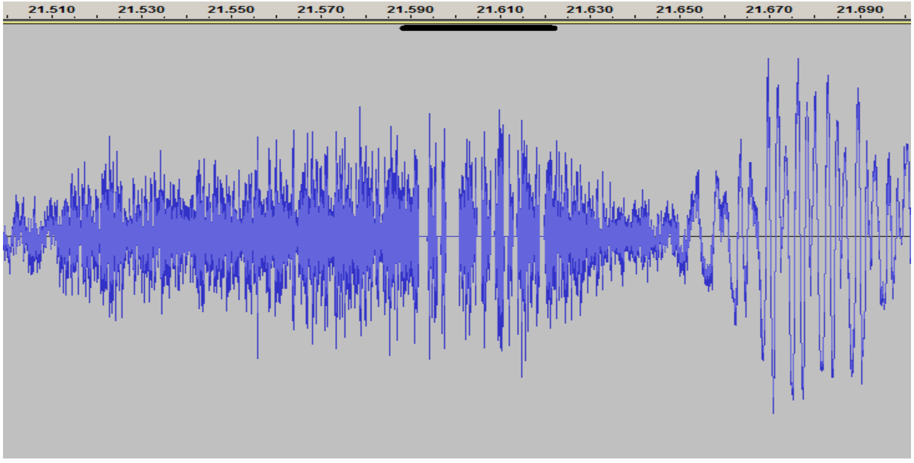


Figure 5.21: TC5 Classification Error 2 — During ‘s’ in ‘son’

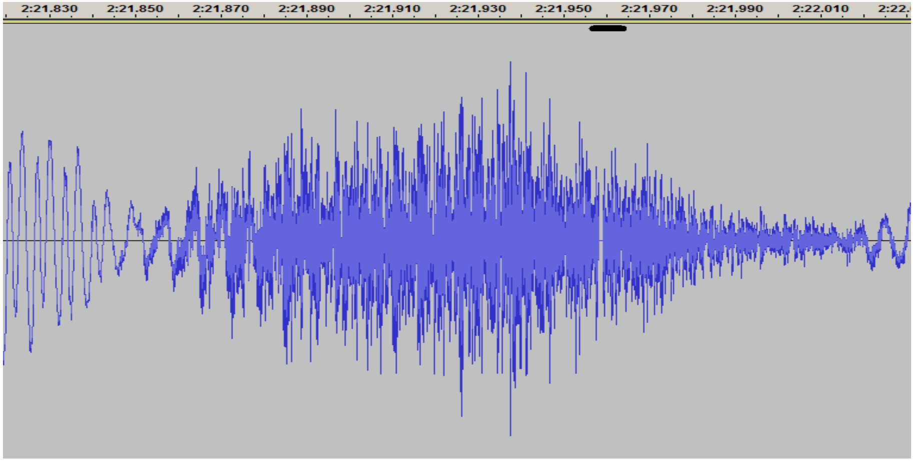
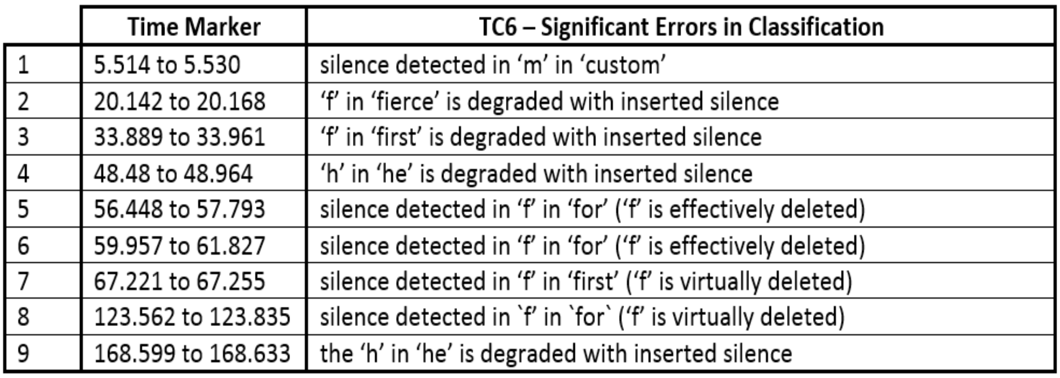


Figure 5.22: TC5 Classification Error 7 — During ‘s’ in ‘smiling’

### TC6 Classification Errors:



| | Time Marker | TC6 – Significant Errors in Classification |
|---|---|---|
| 1 | 5.514 to 5.530 | silence detected in ‘m’ in ‘custom’ |
| 2 | 20.142 to 20.168 | ‘f’ in ‘fierce’ is degraded with inserted silence |
| 3 | 33.889 to 33.961 | ‘f’ in ‘first’ is degraded with inserted silence |
| 4 | 48.48 to 48.964 | ‘h’ in ‘he’ is degraded with inserted silence |
| 5 | 56.448 to 57.793 | silence detected in ‘f’ in ‘for’ (‘f’ is effectively deleted) |
| 6 | 59.957 to 61.827 | silence detected in ‘f’ in ‘for’ (‘f’ is effectively deleted) |
| 7 | 67.221 to 67.255 | silence detected in ‘f’ in ‘first’ (‘f’ is virtually deleted) |
| 8 | 123.562 to 123.835 | silence detected in `f’ in `for` (‘f’ is virtually deleted) |
| 9 | 168.599 to 168.633 | the ‘h’ in ‘he’ is degraded with inserted silence |

Table 5.19: TC6 Errors in Classification for the DNN_Classifier.

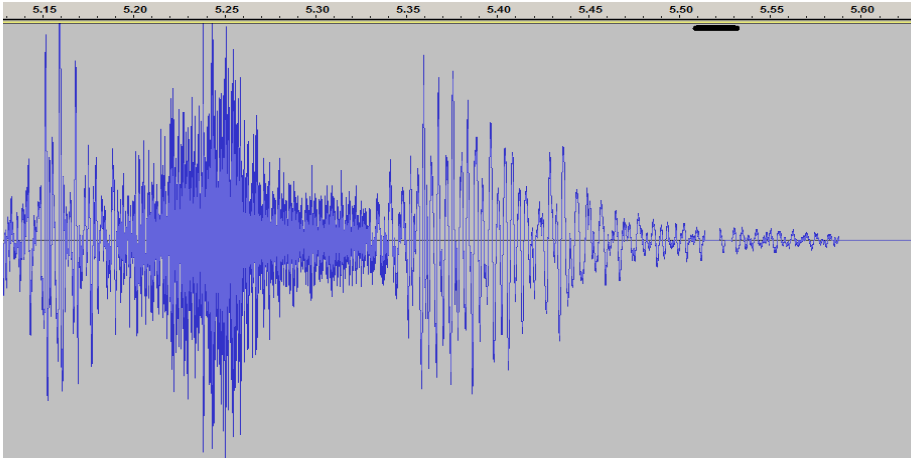


Figure 5.23: TC6 Classification Error 1 — During ‘m’ in ‘custom’

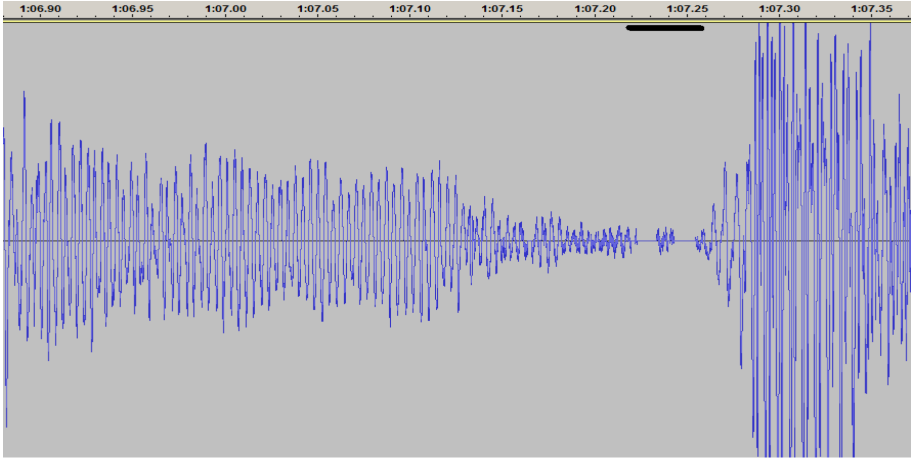


Figure 5.24: TC6 Classification Error 7 — During ‘f’ in ‘first’

Figure 5.25 shows the summation of the durations for the classification errors for the 6 test-cases expressed as a percentage of the duration of the test data sets (Table 4.2) as derived from the data provided in Tables 5.14 through 5.19, and also shows the summation of the durations for the classification errors for the $D_{eterm}$Classifier, for test-case TC1 as derived from the data in Table 5.6, again expressed as a percentage of the duration of the test data set. It is evident that the percentage of speech incorrectly classified by the DNN_Classifier as silence is substantially less for TC1 than for the other test-cases; and this despite the relatively poor DNN training data for TC1 that was obtained with the $D_{eterm}$Classifier. A possible cause of this is that TC1 was provided with approximately 1.8 times as much training data as the other test-cases (see Table 5.12), and this, together with the result in Figure 5.25, suggests that for training data, bulk may be more important than detailed accuracy. That is, if enough of the training data is accurate, then this will swamp the effects of the inaccurate training data; because of the resistance to over-fitting (a feature of the split sample training technique). However this is conjecture, and further empirical work is required to establish exactly why the test-case with the most inaccurate training data seemingly provides the better final classification.

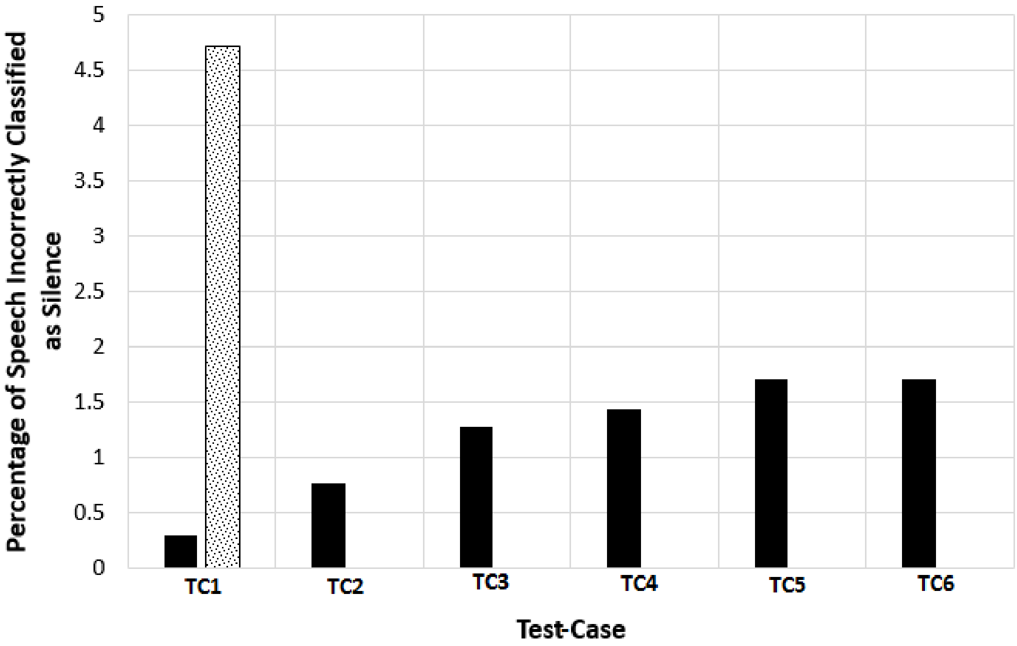


Figure 5.25: Speech Incorrectly Classified as Silence for the DNN_Classifier.

The black bars indicate the percentage of speech incorrectly classified as silence by the DNN_Classifier for each of the test-cases. This data is derived from the temporal durations of the errors given in Tables 5.14 through 5.19. Also indicated (by the patterned bar) is the percentage of speech incorrectly classified as silence for TC1 for the D$_{eterm}$Classifier — as derived from the temporal durations of the errors given in Table 5.6.
Unexpectedly, for TC1, the test-case with the lowest quality classification of all the test-cases for the D$_{eterm}$Classifier, the amount of speech incorrectly classified as silence by the DNN_Classifier is the lowest for all of the test-cases.

### 5.4.4 Summary

That the results show that the DNN_Classifier located 100% of the unfilled silence pauses identified in the ground-truth implicitly supports the first hypothesis and directly supports the second.

The DNN_Classifier showed no particular sensitivity to corrupt training data; and this was largely due to the training data selection strategy which was purposed to select training data only at known good speech/silence and silence/speech transitions.

As noted in Section 5.3.4, the performance of the $\mathrm{D}_{eterm}$Classifier could be improved, but the results for the DNN_Classifier show that in its present form the $\mathrm{D}_{eterm}$Classifier is capable of providing an acceptable quality of training data; and that the two stage binary speech/silence classification system of Figure 1.2, is feasible.

The silence start and end error distributions for the $\mathrm{D}_{eterm}$Classifier for TC1 were significantly different to those for TC2 to TC6, but this had no discernible effect on the operation of the DNN_Classifier. A legitimate question which follows from this is that since the operation of the DNN_Classifier was not compromised by the poor results from the $\mathrm{D}_{eterm}$Classifier for TC1 (these the result of an erroneously high auto-selection of the speech/silence linear separation threshold), then is the complexity of the $\mathrm{D}_{eterm}$Classifier justified?

Of the classification failures for the DNN classifier where short segments of speech were incorrectly classified as silence, 79% were located in unvoiced speech sounds, and the remaining 21% were in voiced speech sounds. This confirms that even for the DNN Classifier that the detection of voiced speech is considerably easier, than the discrimination between unvoiced speech and silence.

Overall, the two stage binary speech/silence classification system might be described as a virtually faultless, low variation, low-granularity filled silence-pause detector, with medium accuracy temporal registration of the silence pause starts and ends, and with a tendency to slightly over-estimate pause durations.

## 5.5 Consistency of the DNN Training Process:

The Train/Validate/Test technique is purposed to find the optimal network, and the results of the testing designed to ascertain the extent to which the particular implementation of the Train/Validate/Test process as used for this work is successful in this, is considered in Section 5.5.1: ‘Variation as a $f_{\text{(Initialisation)}}$’.

Following that, in Section 5.5.2: ‘Variation as a $f_{\text{(Depth of Neural Network)}}$’, is a consideration of the results of an investigation purposed to establishing whether any correlation exists between the depth of the ANN/DNN and the classification performance.

Because the network is trained using a heuristic algorithm, and the training setup includes the initialisation of the weights for all of the artificial neurones with a different bounded random number; then no two trained networks will be the same. Accordingly, there will inevitably be differences in the classification process, and if a network is repeatedly trained and tested with the same training data, then a distribution of the classification results must form. The ‘Optimally Trained Neural Network’ (hereafter the ‘optimal neural network’) is here defined to be any network which provides a result similar to those other networks which provide results where the density of results in the distribution of the classification results is at a maximum (i.e. the most likely solution). If the optimal neural network does not subsequently provide a satisfactory classification, then the problem lies with the quality of the training data; and improvements to that would result in a different — and improved — optimal neural network.

### 5.5.1 Variation as a $f_{\textbf{(Initialisation)}}$

To establish a measure of the consistency of the DNN training process, for each test-case, a combined DNN ‘Training/Execute’ Classification cycle (hereafter, the train/exec. cycle) was repeated 100 times — with a fixed Training, Validation and

Test Data Set. Figure 5.26 illustrates the resultant variability in the classification, as a set of histograms, one for each test-case; where the result of each of the tests is the detected duration of the speech, and each of the histogram bins gives the number of durations which fall within each $200ms$ band. Each histogram is a record of all 100 durations, and all of the histograms are to the same relative, though not absolute, scale. TC5 is the test-case with the least variation $(14 * 200ms = 2.8$ Seconds), and TC1 is the test-case with the greatest variation $(42 * 200ms = 8.4$ Seconds).

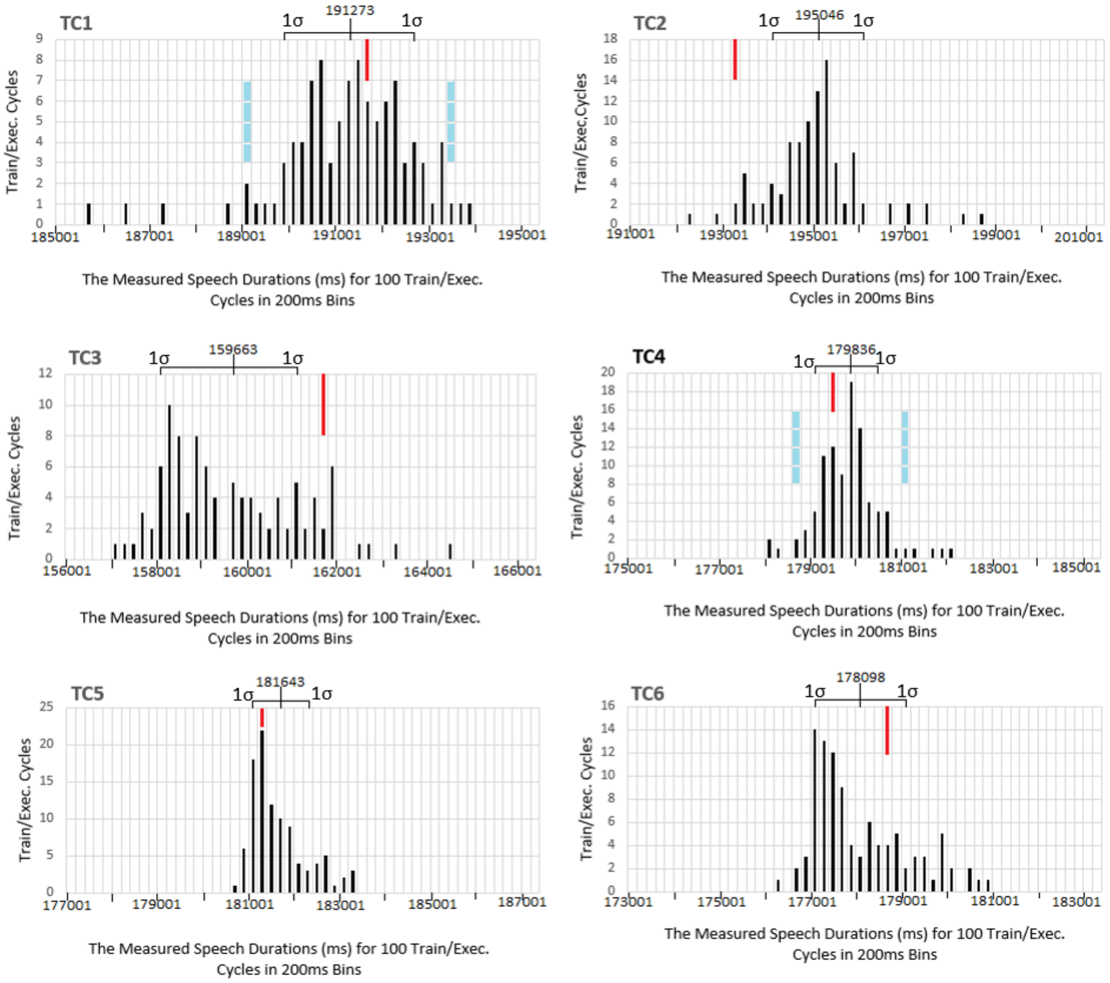


Figure 5.26: Variability introduced into Classification by the DNN_Classifier

A histogram for each of the 6 test-cases is shown, where each bin on the x-axis represents $200ms$ and where the number on the y-axis is the number of train/exec. cycles which result in a speech duration which falls within each bin. The red drop down bars indicate the positions in the distributions of the train/exec. cycles that were used for the detailed performance evaluation in Section 5.3, 'The D$_{eterm}$Classifier', and Section 5.4, 'The DNN_Classifier'; and the turquoise dashed drop down bars on the graphs for TC1 and TC4 indicate the positions in the distributions of the training/execution cycles that were used for the quantitative assessment of the performance of outlier trained DNNs.

Figure 5.26 has been annotated with the mean values, and the sample Standard Deviations ($\sigma$); and also indicates the positions with red drop-down lines in the distributions for the six test-cases used for the detailed performance evaluation in Section 5.3, 'The $D_{eterm}$Classifier', and Section 5.4, 'The DNN_Classifier', and with turquoise dashed drop-down lines for the two test-cases (TC1 and TC4) used for the quantitative assessment of the effect on the performance of the classifier of using a non-optimal DNN. It is interesting that for the detailed analysis, the only test-case which came close to using the optimum Neural Network was TC4, and that for four of the test-cases, TC1, TC4, TC5, and TC6, the performance of the Neural Network was within one SSD of the mean, and that for the remaining two test-cases, TC2 and TC3 the performance of the Neural Network was within two SSD of the mean.

The histograms shown in Figure 5.26 are also a clear representation of the density of the results — integrated over $200ms$ intervals. Were the distributions Gaussian, then the mode of the distribution (the most commonly occurring results) would be a valid indication of central tendency, but it can be seen that the data for TC1 is bimodal, and it is necessary to refer to the mean of the distribution to identify the significant mode of the data. Figure 5.26 shows that for TC1, that the mean of the distribution is a good indication of central tendency.

Figure 5.27 shows the Cumulative Sample Standard Deviation for two data series for up to 100 train/exec. cycles, for each of the test-cases. The two data series per test-case consist of the same data set, but ordered differently; and it can be seen that the number of train/exec. cycles required before the standard deviation achieves some measure of consistency varies considerably between the test-cases, as does the final standard deviation. At worst, up to 60 train/exec. cycles may be necessary before the standard deviation (and by implication the distribution) becomes fully representative of the population — allowing that is, a variability in the standard deviation of up to $200ms$. Such a variability is negligible when considered in the context of the number of silences — the ground-truth — in the Test Data sets. For example, from Figure 5.1 it can be seen that the test-case with the least number of silences is TC4 with 86 unfilled silence pauses; and assuming an equal distribution throughout, a variation of $200ms$

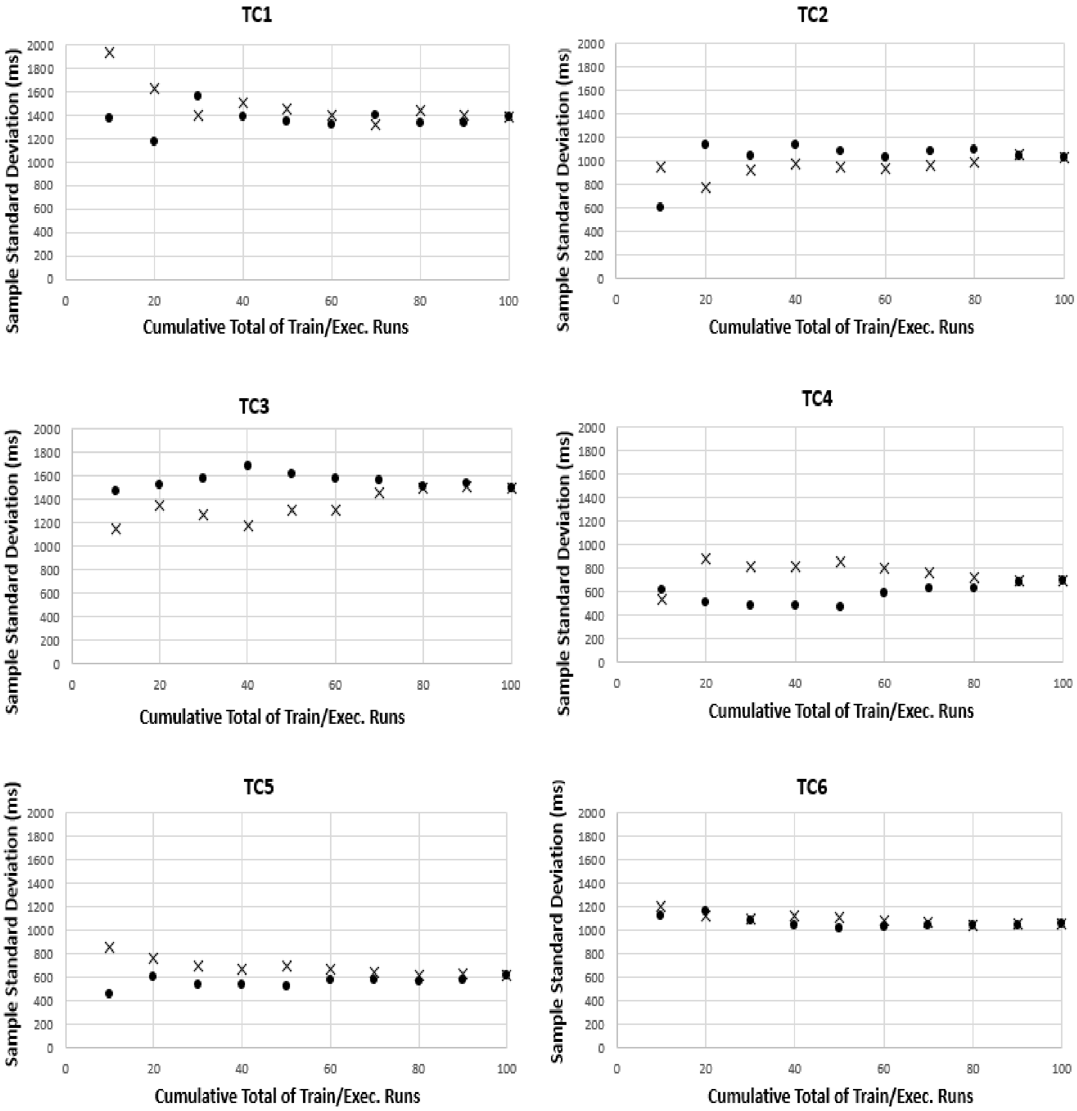
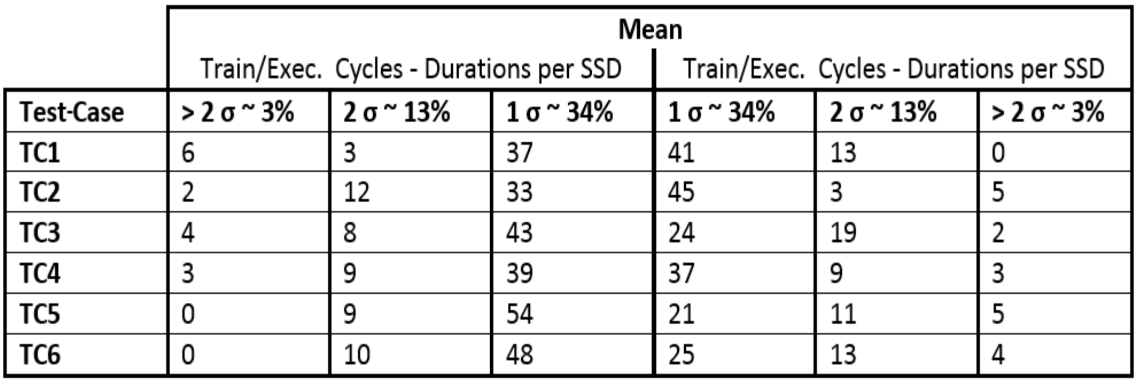


Figure 5.27: Cumulative Standard Deviation — for up to 100 Train/Exec. Cycles.

The x-axis indicates the number of train/exec. cycles used for each of the sample standard deviation calculations, and the y-axis the standard deviation in milliseconds for that particular number of cycles. The two data series for each graph are the same data set, but ordered differently.

would result in a change to each silence duration of less than $2.5ms$.

Table 5.20 provides the totals for the training/execution cycles which fall within each Standard Deviation; and since there were 100 training/execution runs, these numbers are also percentages, and if divided by 100, probabilities. Also shown is the percentage of training/execution cycles which may be expected to fall within each Standard Deviation, were the distributions to comply with the Gaussian Norm. Whilst the distributions cannot be said to be typically Gaussian, there is sufficient clustering of the results within 1 SSD around the mean, to suggest that the optimum Neural Network may be one which provides a result which is within a few hundred milliseconds of the mean for the distribution. The results in Section 5.4 for TC5 — which it can be seen from Figure 5.26 used the preferred/optimal DNN — give no indication of over-fitting (which is where the neural network will operate well with the training data, but less well with new instances of the same class); and a similar situation obtained for TC4 which used a DNN which was near to optimal.



| | Mean | | | | | |
|---|---|---|---|---|---|---|
| | Train/Exec. Cycles - Durations per SSD | | | Train/Exec. Cycles - Durations per SSD | | |
| **Test-Case** | **> 2 σ ~ 3%** | **2 σ ~ 13%** | **1 σ ~ 34%** | **1 σ ~ 34%** | **2 σ ~ 13%** | **> 2 σ ~ 3%** |
| **TC1** | 6 | 3 | 37 | 41 | 13 | 0 |
| **TC2** | 2 | 12 | 33 | 45 | 3 | 5 |
| **TC3** | 4 | 8 | 43 | 24 | 19 | 2 |
| **TC4** | 3 | 9 | 39 | 37 | 9 | 3 |
| **TC5** | 0 | 9 | 54 | 21 | 11 | 5 |
| **TC6** | 0 | 10 | 48 | 25 | 13 | 4 |

Table 5.20: The Results of 100 Train/Exec. Cycles as a Function of Sample Standard Deviation

It can be seen by comparing the data on the RMS speech magnitudes given in Table 5.8, with the standard deviations in Figure 5.27, that they largely correlate. That is, the greater the RMS speech magnitude, the greater the standard deviation. The exception to this is TC1 and TC3 where the RMS speech magnitude for TC1 is marginally higher than that for TC3 whereas for TC1 the standard deviation is lower than for TC3 by approximately $100ms$. Other than that, the correlation suggests, that the operation of the classifier may be some function of the magnitude of the source signal. Conversely, there is no indication of a correlation between the noise during silences and the standard deviations or between the Signal to Noise ratios and the standard deviations. Whether or not the magnitude of the standard deviation is a function of the magnitude of the source signal remains to be seen, but Figure 5.26 shows that the variation in the range of the distributions is such that any individual training cycle may return a less than optimal DNN. So it would seem advantageous to include in the classifier sufficient training runs with the same training, validation and test data, to obtain the distribution — and then to select the optimal DNN as one from where the density of the results is at its greatest.

TC1 and TC4 were selected for a quantitative assessment of the effect on the performance of the speech/silence classifier of using a non-optimal DNN. TC1 was chosen because the distribution of the results was found to be less Gaussian with a higher standard deviation, and TC4 because the distribution of the results was found to be more Gaussian with a lower standard deviation. Two trained DNNs — one each side of the mean — were selected for both TC1 and TC4 (see Figure 5.26), each with a classification performance of within $1\sigma$ and $2\sigma$ of the mean of the distribution. To implement the assessment, the auto-detected silence pauses for each of the networks were inserted in the Filtered speech by forcing the corresponding signal magnitude to true zero, and the first minute of the resultant audio for the two test-cases was compared with the results previously obtained for the performance evaluation ( see Section 5.4: 'The DNN_Classifier').

**TC1_189016ms:** In the first minute of the audio, there were 32 separate instances of the erroneous detection of silence in fricatives amounting to $452ms$. The silence pause starts were typically slightly delayed amounting to $275ms$ over 28 pauses, and the silence pause ends were typically slightly earlier, amounting to $-71ms$ over 28 pauses. There were also 4 additional insertions — each of a few milliseconds — between words, and 3 additional insertions at unvoiced stops. None of the silence pauses were erroneously classified as speech.

**TC1_193423ms:** In the first minute of the audio, there were no instances of the erroneous detection of silence in fricatives. The silence pause starts were typically slightly delayed amounting to $194ms$ over 28 pauses, and the silence pause ends were typically slightly earlier, amounting to $-76ms$ over 28 pauses. None of the silence pauses were erroneously classified as speech.

**TC4_178750ms:** In the first minute of the audio there were no instances of the erroneous detection of silence in fricatives, and variations in the detected pause starts and pause ends were not significant; but $430ms$ of silence was incorrectly classified as speech, compared with the $535ms$ of silence that was incorrectly classified as speech during the performance evaluation.

**TC4_181127ms:** In the first minute of the audio there were no instances of the erroneous detection of silence in fricatives, and variations in the detected pause starts and pause ends were not significant, but 770$ms$ of silence was incorrectly classified as speech, compared with the 535$ms$ of silence that was incorrectly classified as speech during the performance evaluation, and there were 7 extra silence insertions.

From the above, it can be seen that a degraded classification performance — particularly evident in the results for TC1_189016ms and TC4_181127ms — can be the outcome if a network is chosen which returns a result that is on the periphery of the distribution. So again, it may be concluded that for the implementation of the train/validate/test technique used for this work, an additional step of establishing the distribution of results returned for multiple train/exec. cycles for the same training, validate and test data set is desirable, although, from the data in Table 5.20 it can be seen that the probability of any particular train/exec. cycle returning a result out-with the mean ±1 Standard Deviation is relatively low.

### 5.5.2 Variation as a $f_{\text{(Depth of Neural Network)}}$

The contention explored in the previous section was the performance of a DNN is some function of the random initialisation of the weights within the network, and that there exists an optimally trained DNN that can be selected from the results of a set of train/exec. cycles, such that the optimally trained DNN is from that part of the distribution where the density of results is at a maximum. The results of limited testing suggest that this is the case, but don't address the associated question of whether a DNN with 2 Hidden Layers is the optimal configuration. The preliminary informal testing purposed to identifying the optimal ANN configuration (Section 4.8.1) was able to identify unsuitable ANN configurations but was less successful at identifying the optimal configuration. To investigate whether the DNN with 2 Hidden Layers is the optimal network, the experimental procedure that was used to establish 'Variation as a $f_{\text{(Initialisation)}}$' was repeated for test-case 3 only, with Artificial Neural Networks with 1, 4 and 6 hidden layers. Because the purpose of this work, was to investigate the link between the depth of the neural network and its performance, the width of the hidden layers was held constant at 27 artificial neurones — although clearly, many different NN configurations exist.

Figure 5.28 shows a rescaled version of the histogram for TC3 for the DNN with two hidden layers (as already discussed in Section 5.5.1 — 'Variation as a $f_{\text{(Initialisation)}}$'), plus histograms for an ANN with 1 hidden layer, a DNN with 4 hidden layers and a DNN with six hidden layers. The same scaling is used for the x-axis for all 4 histograms, so the histograms are directly comparable.

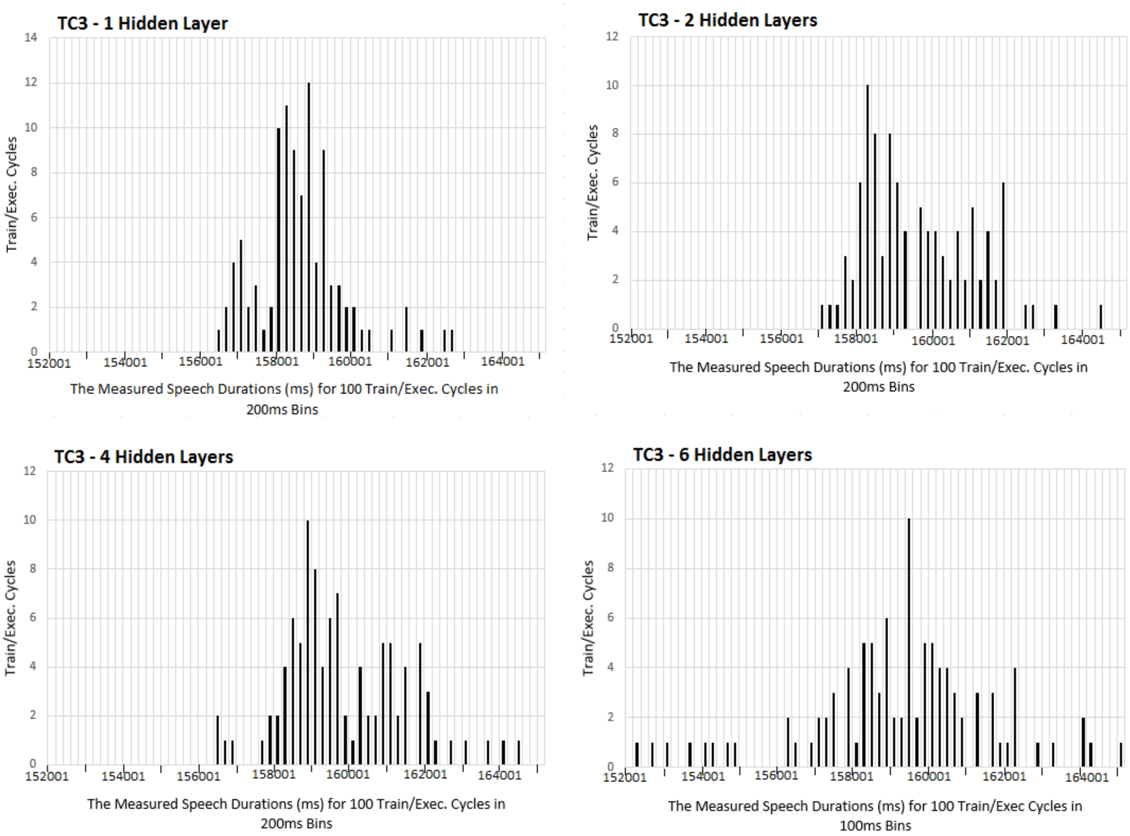


Figure 5.28: Variability in Classification as a $f_{\text{(Depth of Neural Network)}}$ for TC3.

A histogram for each of the 4 ANN configurations is shown, where each bin on the x-axis represents $200ms$ and where the number on the y-axis is the number of train/exec. cycles which result in a speech duration which falls within each bin. The data, 'TC3 – 2 Hidden Layers', is a rescaled version of that already shown in Figure 5.26 in Section 5.5.1: 'Variation as a $f_{\text{(Initialisation)}}$'.

From Figure 5.28, it can be seen that the network with the distribution with the lowest dispersion is the ANN with 1 hidden layer, that the distributions for the DNNs with 2 and 4 hidden layers are very similar — though greater than those for the network with 1 hidden layer, and the network with the greatest dispersion is the DNN with 6 hidden layers. The Modes for the distributions for the three DNNs indicate that for the 'preferred solution', that the duration of detected speech increases as the depth of the network is increased.

Figure 5.29 shows the Cumulative Sample Standard Deviation for the 4 networks, derived as described in Section 5.5.1 — Variation as a $f_{\text{(Initialisation)}}$.

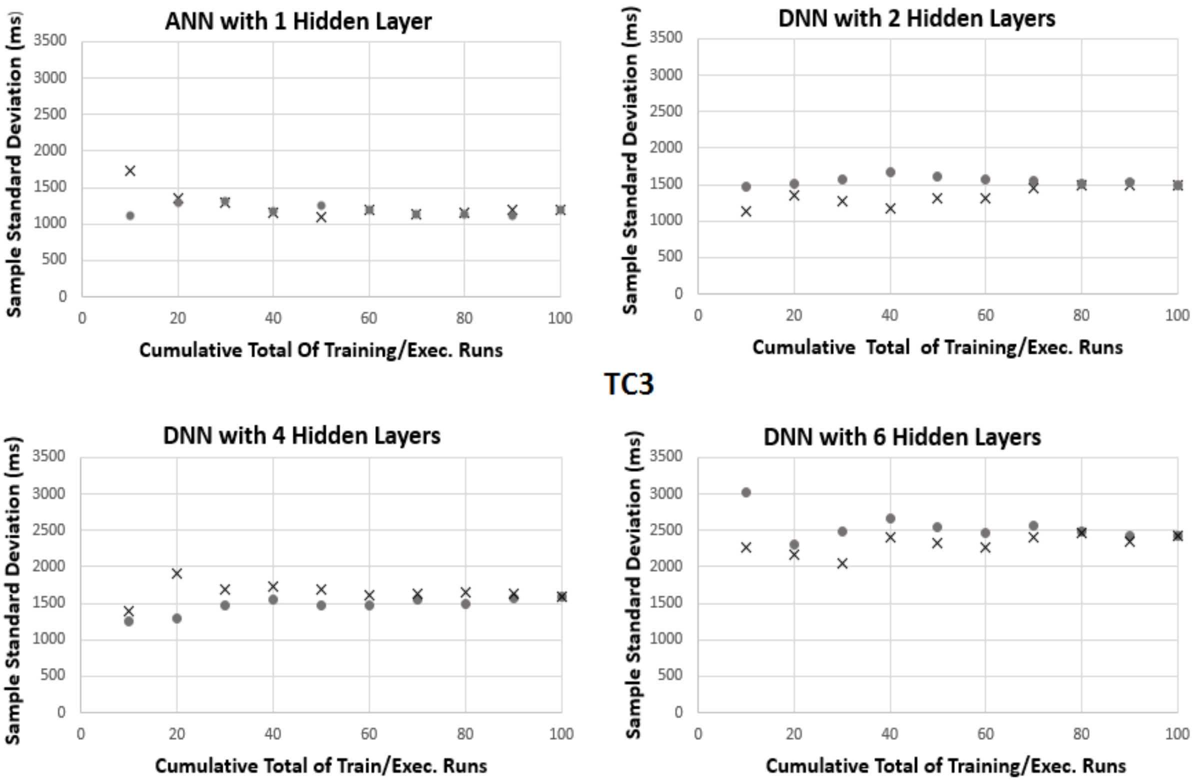


Figure 5.29: Cumulative Sample Standard Deviation as a $f_{\text{(Depth of Neural Network)}}$ for TC3.

The x-axis indicates the number of train/exec. cycles used for each of the sample standard deviation calculations, and the y-axis the standard deviation in milliseconds for that particular number of cycles. The two data series for each graph are the same data set, but ordered differently. Note that the graph for the 'DNN with 2 Hidden Layers' is a rescaled version of the graph for TC3 shown in Figure 5.27 in Section 5.5.1: 'Variation as a $f_{\text{(Initialisation)}}$'.

The standard deviation for 100 train/exec. cycles can be seen to be a function of the depth of the network, increasing as the depth of the network is increased. In absolute terms the standard deviation varies from 1.183 seconds for the ANN with 1 hidden layer to 2.426 seconds for the DNN with 6 hidden layers; and for the DNNs with 2 and 4 hidden layers, the Standard deviations are 1.493 and 1.584 respectively.

From Figure 5.28, it can be seen — with the exception of the network with 1 hidden layer — that the Modes of the Distributions are a good indicator of the centre of the distributions. That said, there is sufficient variation in the centre of the distributions to warrant an assessment of classification differences by the 4 networks. Accordingly, for each of the 4 networks, a train/exec. cycle was chosen for detailed analysis from the distribution where the duration of the detected speech was within the bin where the density of results was at (or near) a maximum.

The data for analysis comprised four files: ANN_1HidLyr_158385ms.wav, DNN_2HidLyrs_158218ms.wav, DNN_4HidLyrs_158804ms.wav, DNN_6HidLyrs_159667ms.wav — constructed as before, by inserting true zero into the Filtered speech files at all auto-detected silence locations. The investigation was in two parts: the first a detailed analysis of the first minute of the data files, and the second an investigation — in response to the outcome of the analysis of the first part — as to why the disparity in the durations of detected speech for the different networks, as exemplified by the 1.449 seconds disparity between the DNN_2HidLyrs_158218ms.wav and the DNN_6HidLyrs_159667ms.wav.

The detailed analysis of the first minute of the data sets versus the ground-truth showed that the performance of the four networks to be similar. That is, there was little difference between the silence starts and end points between the 4 networks, with a consequential similar disparity between those measurements and the ground-truth. The most significant difference between the 4 data sets was between the number of Insertions — auto-detected silence without an analogue in the ground-truth — where for the network with 1 hidden layer there were a total of 38 Insertions, the network with 2 hidden layers 39 Insertions, with 4 hidden layers 34 Insertions and with 6 hidden layers 30 Insertions. On the basis of Insertions alone, it might seem that the network with 6 hidden layers provided a slight performance improvement; but the results discussed so far don't address the question of why the optimum network for the 6 layer network detects 1.449 seconds more speech than the 2 layer network. To put this in context, the TC3 test data set has 99 unfilled silences (from the ground-truth) which — with an even distribution of the additional 1.449 seconds of speech throughout the recording — would result in a delta per pause of $\frac{1.449}{99} = 11.3ms$, or a variation per speech/silence and silence/speech transition of $5.65ms$.

A variation of the order of $5.6ms$ in the auto-detected location of speech end-points and onsets is negligible, given the precision for this work as specified in Section 5.3; provided that is, that the deltas are evenly distributed throughout. To assess the distribution, the temporal duration of the locations for all of the speech insertions in the silence pauses for the 2 layer DNN was compared with that for the 6 layer DNN, and a significant disparity was measured. Overall, a difference of approximately $600ms$ was identified for the speech insertions, $318ms$ of which was associated with a loud noise event (the noisy rustling of paper) that followed the speech. This reduces the deltas attributable to each silence/speech and speech/silence transitions to a little less than $\frac{1.449-0.6}{99*2} \sim 4.3ms$, given that the 6 layer DNN will re-introduce a small amount of speech commensurate with the reduced number of silence insertions.

There are two key points in the previous paragraphs. Firstly, although the 6 Layer DNN appears to provide a small performance advantage when detecting silence, it provides a small disadvantage in that it also increases the amount of speech (noise in fact) that is erroneously inserted in the detected silences (most usually in bulking-up existing noise insertions). Secondly — the loud noise event at the end of the TC3 test-data set aside — the time deltas between the speech durations for the four networks that were chosen to be representative of their aggregated density distributions are evenly spread throughout the recordings.

Although the activity of assessing the impact of the depth of the DNN on classification performance has failed to reach a definitive conclusion, this is not an entirely negative outcome. It is implicit to the hypotheses, which focus on the speech model and the training data selection method, that these dictate the performance of the classifier, rather than the configuration of the ANN/DNN; and the results of this section confirm this by showing that the performance of the classifier is largely independent of its depth (within the bounds explored).

Aside: It must be observed that the noisy rustling of paper at the end of the recording were not included in the calculation of the signal to noise ratios given in Table 4.1, as the raw data for the calculations was derived using the 45 seconds of the speech that followed after the first minute of the recording. That is, the signal to noise ratios given in Table 4.1 are very much an approximation. However, the rustling sounds did appear as a single 'speech' segment to the D$_{eterm}$Classifier and as a consequence were a source of corruption to the 'speech' training data for the DNN_Classifier.

### 5.5.3 Summary

Note that the points below refer to the specific embodiments of the techniques described in this work.

**Variation as a $f_{\textbf{(Initialisation)}}$**

- The training and classification performance of the DNN is a function of the initialisation — with bounded random values — of the neurone inputs weights.

- If a DNN is subjected to multiple train/exec. cycles with the same Training/Validation/Test data set, then the collected results form a distribution with a clear central tendency.

- An optimally trained DNN exists — which lies in the vicinity of the mean of the multiple train/exec. cycles results distribution.

- The Training/Validation/Test technique does not necessarily identify the optimally trained DNN.

- To ensure the best classification network, when training a DNN, the optimally trained DNN should be selected from the vicinity of the mean of a distribution of train/exec. results for the same Training/Validation/Test data set.

- If a train/exec. cycle is used which lies on the periphery of the results distribution, then the classification performance may be adversely affected, and the risk of this being a problem increases as the dispersion of the results distribution increases.

- If the optimally trained DNN in subsequent use fails to provide a satisfactory classification, then the problem is with the quality of the training data.

- If the training data is changed then a new optimally trained DNN should be identified.

- There is no particular correlation between the noise during silence or the signal to noise ratios, and the magnitude of the standard deviation; but there is a correlation between the RMS speech magnitude and the magnitude of the sample standard deviation — which is worthy of further investigation.

### Variation as a $f_{\textbf{(Depth of Neural Network)}}$

- The Sample Standard Deviation increases in magnitude as the depth of the network is increased.

- For the three DNNs, the mode of the distribution for multiple train/exec. cycles with the same training/validation/test data set increases in absolute value as the depth of network is increased.

- It becomes increasingly important to identify the optimally trained DNN DNN as the depth of the network is increased because of the increased dispersion.

- The speech/silence classification performance for the three DNNs and the ANN is very similar; but an increasing amount of noise during silence is classified as speech as the depth of the DNN is increased.

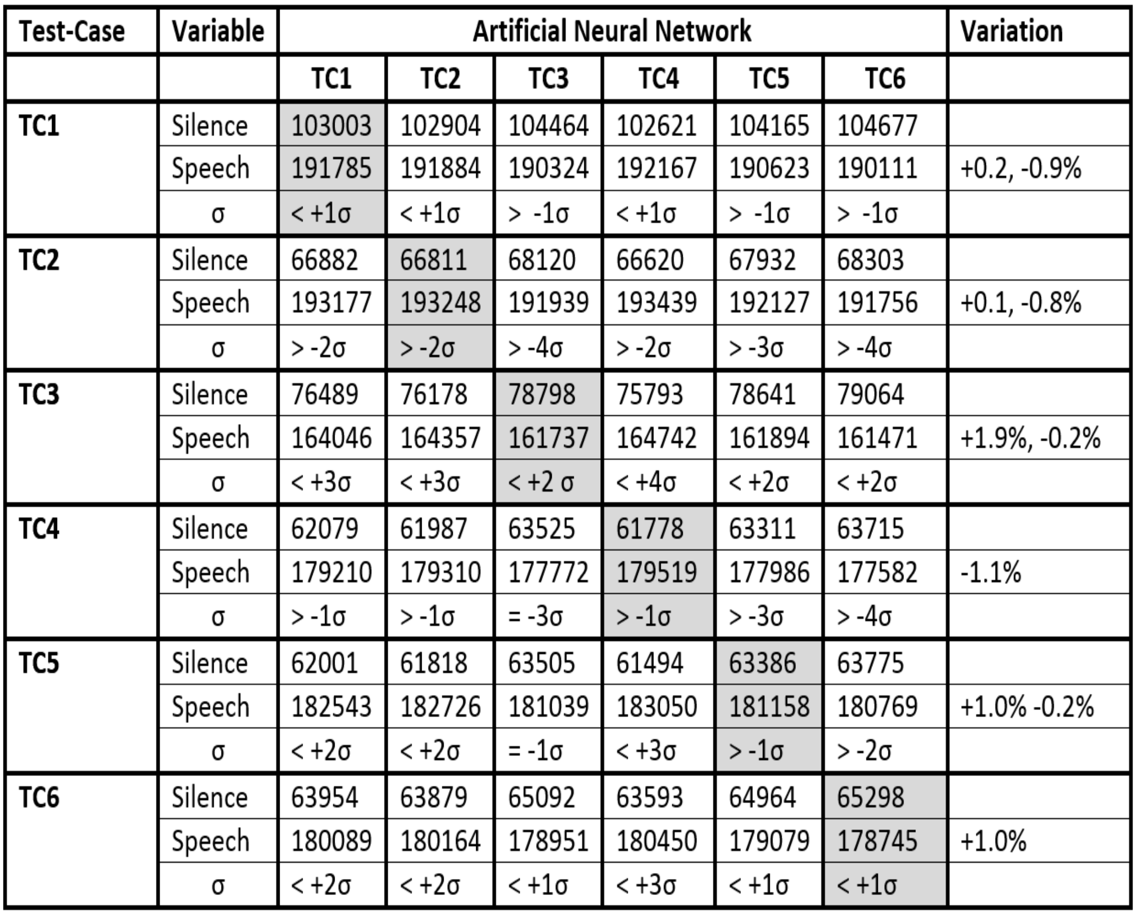
- The results support the implication of the hypotheses that the classification is more a function of the quality of the training data and the the audio model, than the network configuration (within bounds).

## 5.6 Generalisability of the Solution:

Since the recordings which constitute the speech corpus were made in the same environment, using the same equipment and are in the same voice, it can be argued that the separate neural nets for each of the test-cases should be similar to one another — and from this it may be inferred that any of the trained DNNs will provide a satisfactory performance with all of the test-cases.

Table 5.21, shows the result of the classification into speech and silence for all of the test data sets by each of the trained DNNs. The variation is the maximum difference between the trained DNN for the particular test-case (highlighted in the diagonal) and the other 5 DNNs. In absolute terms, the duration of the detected speech varied from 1.683 seconds, for the neural network trained with the TC2 data set to 3.271 seconds, for the neural network trained with the TC3 data set. A variation in duration of 1.683 seconds equates to $\sim 17ms$ per silence pause ($\sim 8.5ms$ per speech/silence and silence/speech transition), whereas 3.271 seconds equates to $\sim 33ms$ per silence pause ($\sim 16.5ms$ per speech/silence and silence/speech transition). These figures assume that the variations in total speech duration are evenly distributed throughout the test audio, and may be artificially high because the effect of the bulking up of existing noise insertions during silence pauses, as described in section 5.5.1, is ignored. Whatever, the worst case figure of $16.5ms$ per transition is within the error margin of $20ms$ as specified in Section 4.3, the 'Ground-Truth'.

The standard deviation measures are included so that the extent to which the measured speech durations are within the bounds of the distributions already established for each of the test-cases shown in Figure 5.26 can be assessed. That there are 4 instances of absolute measured speech durations between $3\sigma$ and $4\sigma$ of the mean for the distributions is indicative that the six trained DNNs do not all provide the same classification performance.



| Test-Case | Variable | Artificial Neural Network | | | | | | Variation |
|---|---|---|---|---|---|---|---|---|
| | | TC1 | TC2 | TC3 | TC4 | TC5 | TC6 | |
| TC1 | Silence | 103003 | 102904 | 104464 | 102621 | 104165 | 104677 | |
| | Speech | 191785 | 191884 | 190324 | 192167 | 190623 | 190111 | +0.2, -0.9% |
| | σ | < +1σ | < +1σ | > -1σ | < +1σ | > -1σ | > -1σ | |
| TC2 | Silence | 66882 | 66811 | 68120 | 66620 | 67932 | 68303 | |
| | Speech | 193177 | 193248 | 191939 | 193439 | 192127 | 191756 | +0.1, -0.8% |
| | σ | > -2σ | > -2σ | > -4σ | > -2σ | > -3σ | > -4σ | |
| TC3 | Silence | 76489 | 76178 | 78798 | 75793 | 78641 | 79064 | |
| | Speech | 164046 | 164357 | 161737 | 164742 | 161894 | 161471 | +1.9%, -0.2% |
| | σ | < +3σ | < +3σ | < +2 σ | < +4σ | < +2σ | < +2σ | |
| TC4 | Silence | 62079 | 61987 | 63525 | 61778 | 63311 | 63715 | |
| | Speech | 179210 | 179310 | 177772 | 179519 | 177986 | 177582 | -1.1% |
| | σ | > -1σ | > -1σ | = -3σ | > -1σ | > -3σ | > -4σ | |
| TC5 | Silence | 62001 | 61818 | 63505 | 61494 | 63386 | 63775 | |
| | Speech | 182543 | 182726 | 181039 | 183050 | 181158 | 180769 | +1.0% -0.2% |
| | σ | < +2σ | < +2σ | = -1σ | < +3σ | > -1σ | > -2σ | |
| TC6 | Silence | 63954 | 63879 | 65092 | 63593 | 64964 | 65298 | |
| | Speech | 180089 | 180164 | 178951 | 180450 | 179079 | 178745 | +1.0% |
| | σ | < +2σ | < +2σ | < +1σ | < +3σ | < +1σ | < +1σ | |

The standard deviation measures provide an indication of the locations for the speech durations on the distributions shown in Figure 5.26. Three of the test-cases, TC2, TC3 and TC4 include speech durations with an absolute location between $3\sigma$ and $4\sigma$ of the mean for the distributions.

Table 5.21: Generalisability of the Trained DNNs

Magnitude normalisation of the model data is described in Section 4.7, and is where the audio model data is scaled to be within the range of 0.0 to 1.0 before DNN training or processing. This normalisation is an important contributory factor to the consistency of the results across the test-cases shown in Table 5.21. A second important contributory factor is the implicit frequency normalisation that follows when all of the test-cases are in the same voice. Clearly, were different voices to be used in the testing, then frequency normalisation would be required to achieve similar results. (The basic assumptions which underpin this are that the essence of speech is constant for all, and that the normalised model as described herein is capable of capturing this essence of speech.)

In summary, the results indicate that the six trained neural networks provide a similar, though not identical, classification performance; and the magnitude of the differences in detected speech durations — at $< 16.5ms$ per speech/silence and silence/speech transition — is within the margin of error specified for this work in Section 4.3, the 'Ground-Truth'.

## 5.7 Experiments using an MP3 Speech Corpus

During the development phase of this work, a test corpus comprising eight extracts from 'talking books' by eight different individuals (four female and four male) was used, and results with this corpus are described in Appendix C: "Experiments with MP3 Compressed Speech". The key differences between the tests in Appendix C and the main tests are for the work in Appendix C that :-

- The recordings were in the MP3 Lossy Compression format, and so had already been subjected to psychoacoustics processing.
- The ANN was configured with an input layer of 27 neurons, one hidden layer with 54 neurons, and an output layer of 1 neuron.
- The rules for selecting the Artificial Neural Network Training data were different (Figure 5.30).

The experiments on the MP3 corpus yielded some systematic results that were similar to those of the main experiments. Specifically, whereas the D$_{eterm}$Classifier identified only between 81.5% and 96.5% of the perceived unfilled pauses, the ANN_Classifier identified better than 98% of the pauses, and though the performance of the D$_{eterm}$Classifier degraded as the noise level increased, this was not the case for the ANN_Classifier. Caution is necessary when correlating the results of two unrelated speech corpora in this way, but common to both are that the speech is intelligible, and that the post-recording processing system is similar.

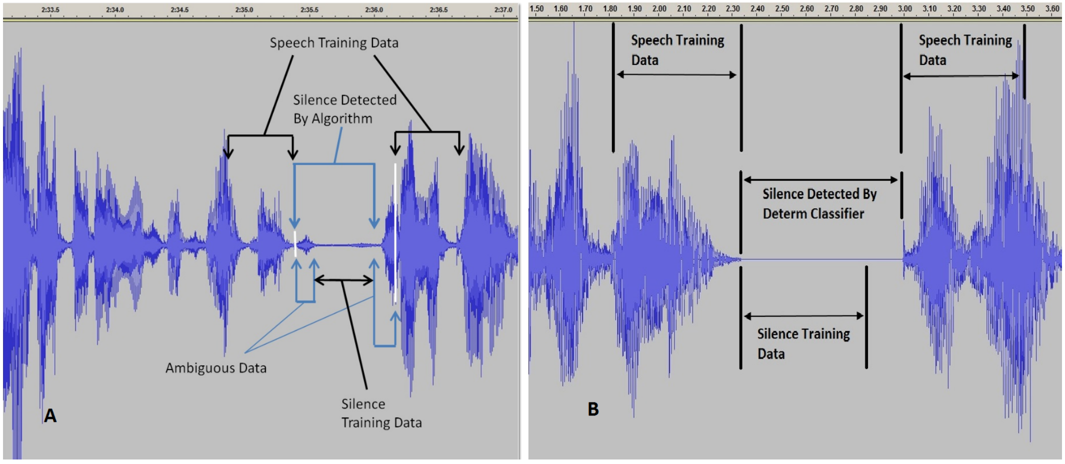


Figure 5.30: Alternative Rules for the Selection of the Speech and Silence Training Data

Part **A** shows the rules for the selection of training data for the MP3 tests, and Part **B** the rules for the selection of training data used for the main part of this work. Counter intuitively the rules shown in Part **B** seem to provide the better classification performance; though the evidence for this is tenuous. A similar result was found many years earlier by Ghiselli-Crippa and El-Jaroudi [1991]. The rules for the selection of training data for the MP3 tests are fully described in Appendix C.3, and the rules for the selection of the training data used for the main part of this work in Section 4.7.

The most consistent failure mode of the $D_{eterm}$Classification process was the deletion of weak consonants at the end of blocks of speech; and this is indicative of a possible problem with comparing the RMS magnitude of a 'prototype' clause with the RMS magnitude of short potential clauses. That is, the consonants at the end of speech blocks can be preceded by an imperceptible silence or near silence, and if the $D_{eterm}$Classifier recognises this as silence, then the consonant is 'separate' from the block of speech, is tested as a potential clause, and may be rejected. So it is possible to explain how weak consonants can be deleted by the $D_{eterm}$Classifier.

The most common failure mode for the DNN_Classifier was the detection of silences during voiced speech. With knowledge of earlier work by Ghiselli-Crippa and El-Jaroudi [1991] — who found that the performance of their system for voiced/unvoiced/silence discrimination improved when they included the transitional silence/speech data in their training set — it was surmised that these failures were also possibly due to the training data strategy, where the ambiguous data around the speech/silence decision point was excluded from the training data set (Figure 5.30A). In a comparison of the test results of the main body of this work — where the ambiguous data around the silence/speech decision points was included in the training data (Figure 5.30B) — with the results of the experiment described in Appendix C, it was found that the incidence of erroneous silence detection during voiced speech was substantially lower for the former than for the latter; and this tends to support the findings by Ghiselli-Crippa and El-Jaroudi [1991], that the ambiguous data at the speech/silence and silence/speech transitions is better included in the training data set.

## 5.8 Discussion

This work differs with most other work in the field in three respects. The first is the feature set (or speech model), the second is the two step classification process, and the third is the intent — fully automated silence detection. One paper which atypically addresses all of these is by Deng and O'Shaughnessy [2007], where the authors achieved a voiced-unvoiced-silence detection accuracy that was greater than 91.15 %, and a VAD accuracy that was greater than 97.45 %. Note that Deng and O'Shaughnessy used unsupervised learning from the actual data being classified, and so did not require a training data set.

Works by Wei and Yanpu [2005] and Ghiselli-Crippa and El-Jaroudi [1991] also has some similarities with the work herein. Wei and Yanpu [2005], selected the harmonics with the greatest signal to noise ratio from the noisy speech input, and then regenerated the speech audio using only those harmonics. They took no cognizance of Bark bands and masking, so their audio model is not the perceptual model. Additionally they expressed their results in the form of improvements in the signal to noise ratio, hence there is no basis for a comparison of results. Ghiselli-Crippa and El-Jaroudi [1991] obtained with their tri-state ANN classifier, an accuracy of between 95.31 to 96.63%. That said, the authors were concerned particularly with the training algorithm, and manually built their training data. Their best performance was with a training data set which included transitional frames (those difficult to classify), and the results herein duplicate Ghiselli-Crippa and El-Jaroudi's findings, in this respect.

# Chapter 6: Conclusions

The purpose in this work was to research and develop methods for identifying the silence pauses in pre-recorded continuous speech, and the early work involved an investigation into a deterministic speech/silence classifier, the $\mathrm{D}_{eterm}$Classifier. The results of this investigation — presented in Chapter 5 — confirmed the findings of other researchers that there is an issue with the discrimination between unvoiced speech and silence. So, although the $\mathrm{D}_{eterm}$Classifier correctly identified typically more than 86% of the unfilled perceived silence pauses it did not compare well with other speech/silence classifiers, except in one important respect: though it would only detect a percentage of silence pauses, it did so with certainty; and this was deemed enough to facilitate the generation of training data for a second speech/silence classifier.

Reasoning that we as individuals are very capable speech/silence discriminators, it was decided that the psychoacoustics model of speech provided the greatest possibilities. So the $\mathrm{LogFB}_{dynamic}$ speech model was devised, and in keeping with the human perception paradigm, it was decided to use a Supervised Artificial Neural Network as the second speech/silence classifier.

Whereas the $\mathrm{D}_{eterm}$Classifier was perhaps more the result of development than research, the $\mathrm{LogFB}_{dynamic}$ evolved from a synthesis of ideas. Fletcher [1940] described how the frequency response of the cochlea (the mechanical to neurological transducer within the ear) is fixed, and Egan and Hake [1950] confirmed that the masking effect tracks the instantaneously most powerful harmonics. Although later Zwicker [1961] was a little ambiguous on whether the critical bands might be statically bounded — and no evidence of such bounds has so far been found — his formalisation of the idea of critical bands included the Bark scale; and from this it may be supposed that the minimum number of critical bands required to effectively span the audio spectrum is the number of Bark bands as defined by Zwicker. There is some empirical evidence that this might be the case, in that the partitions for the perceptual model for the successful MP3 Decoder were selected to be, “Roughly equivalent to the critical bands of human hearing” — Brandenburg [1999]. Accordingly two hypotheses were formed as follows:

## 6.1 Hypotheses

- First Hypothesis: An audio model based upon an interpretation of the psychoacoustics model of hearing will include sufficient information to facilitate the recognition of silence in speech.
- Second Hypothesis: A deterministic speech/silence binary classifier can provide enough information to facilitate the generation of accurate speech and silence training data, such that the errors in classification by the deterministic classifier can be eliminated by a subsequent supervised classification process which uses a psychoacoustics audio model.

## 6.2 Contribution

The results (see Chapter 5) showed that the DNN_Classifier — when trained with data generated by the $\mathrm{D}_{eterm}$Classifier — correctly identified the silence pauses that the $\mathrm{D}_{eterm}$Classifier had incorrectly classified as speech, and support both hypotheses; as do the results presented in Appendix C.

The first contribution — and one that is central to this work — is the $\mathrm{LogFB}_{dynamic}$ speech representation. This model differs in key respects with existing speech models in that it:

- was conceived as a unified dynamic model which could operate with a higher temporal resolution than other unified models, such as the MFCCs family of speech models.

- used a variation of the log filter-bank which did not use overlapping filters.
- incorporated an approximation of simultaneous dynamic masking.

The second, and the lesser of the two contributions, was the refinement of a two stage speech silence classifier so that it comprised a deterministic classification followed by a supervised classification. In fact the deterministic classifier itself included separate processes to identify voiced and unvoiced speech, so conceptually — though not in the realisation — had some similarities with the work of Qi et al. [2004]. A key characteristic of the D$_{eterm}$Classifier is that it can reliably detect a high percentage of pauses; and this is of importance because certain knowledge of the temporal locations of the silence pauses is essential when generating the training data for any supervised pattern matching system. (It is self evident that the time before and after known silence pauses is not silence, and this provides a means of accurately selecting the training data.)

Not only did the combination of the D$_{eterm}$Classifier with the LogFB$_{dynamic}$ and the DNN_Classifier automatically identify the location of unfilled silence pauses, it also identified the location of some of the filled silence pauses plus many brief moments of silence associated with stop consonants. Unfortunately the silences associated with the filled pauses are very short; and within the context of this work there is no automated method for identifying whether a short silence is at a stop consonant or at a filled pause. (The identification of stop consonants falls within the area of automatic speech recognition, and the identification of filled pauses is a research area in its own right; and neither of these falls within the scope of this thesis.)

In consideration of the specific application areas identified in the introduction, with some further work the speech/silence classifier might be suitable for data compression, speech analysis, and creating the silence model for forced alignment, but is less suitable for the 'ASR automatic insertion of punctuation' activity.

The LogFB$_{dynamic}$ speech model, although a departure from much of the work in the Literature, performed well enough to justify its use for the recognition of silence in speech; so the research question on whether an alternative to the existing methods of discriminating between speech and silence exists has been answered in the affirmative.

## 6.3 Future Work

One future activity might be to change the LogFB$_{dynamic}$ model by incorporating frequency normalisation as well as magnitude normalisation. The purpose in this would be to investigate whether the model would then contain enough of those elements of speech that are common to all speakers, to facilitate a more general classification capability.

A second activity might be to research whether the noise immunity, and hence performance of the system shown in Figure 1.2 could be improved by adding the capability for the selection of the 'optimal' trained DNN.

A third activity might be to investigate whether the combination of the D$_{eterm}$Classifier with the DNN_Classifier would be suitable for the detection of other 'Robust Categorical Features' of speech (Lea [1986]) — such as sibilants. Ultimately this might support a technique for synchronising recorded speech with text by using the detected silence to delineate phrases, and the robust categorical features to identify specific words.

# Appendix A: Bark Band Ada Specification

This Ada specification defines the harmonics used in coding the Bark bands and illustrates that with a 30$ms$ DFT window the first 5 bands are constructed from three harmonics each. Although this imposes a limit on the accuracy for the lower Bark bands, there remains some scope for implementing masking.

The harmonic frequency is the harmonic number multiplied by $33.33_{rec}$ Hz.

```
SAMPLES_IN_10mS : constant := 441;
SAMPLES_IN_30mS : constant := SAMPLES_IN_10mS * 3 ;

type Valid_Harmonics_Index_Type
   is range 0..((SAMPLES_IN_30mS)/2 - 1);  (Shannon [1949])

type AudioFeatureType is
  (Bark_B1,  Bark_B2,  Bark_B3,  Bark_B4,
   Bark_B5,  Bark_B6,  Bark_B7,  Bark_B8,
   Bark_B9,  Bark_B10, Bark_B11, Bark_B12,
   Bark_B13, Bark_B14, Bark_B15, Bark_B16,
   Bark_B17, Bark_B18, Bark_B19, Bark_B20,
   Bark_B21, Bark_B22, Bark_B23, Bark_B24);

Harmonic_Rnge_Frst : Constant Valid_Harmonics_Index_Type := 1;
Harmonic_Rnge_Last : Constant Valid_Harmonics_Index_Type := 2;

type Valid_Harmonics_Range_Type
   is array (Harmonic_Rnge_Frst..Harmonic_Rnge_Last)
      of Valid_Harmonics_Index_Type;

type Valid_Harmonics_Range_Arr_Type
   is array(AudioFeatureType)
      of Valid_Harmonics_Range_Type;
```

```
Valid_Harmonics_Range_Constants : Valid_Harmonics_Range_Arr_Type :=
   (Bark_B1  =>(Harmonic_Rnge_Frst => 1,  Harmonic_Rnge_Last => 3),
    Bark_B2  =>(Harmonic_Rnge_Frst => 4,  Harmonic_Rnge_Last => 6),
    Bark_B3  =>(Harmonic_Rnge_Frst => 7,  Harmonic_Rnge_Last => 9),
    Bark_B4  =>(Harmonic_Rnge_Frst => 10, Harmonic_Rnge_Last => 12),
    Bark_B5  =>(Harmonic_Rnge_Frst => 13, Harmonic_Rnge_Last => 15),
    Bark_B6  =>(Harmonic_Rnge_Frst => 16, Harmonic_Rnge_Last => 19),
    Bark_B7  =>(Harmonic_Rnge_Frst => 20, Harmonic_Rnge_Last => 23),
    Bark_B8  =>(Harmonic_Rnge_Frst => 24, Harmonic_Rnge_Last => 27),
    Bark_B9  =>(Harmonic_Rnge_Frst => 28, Harmonic_Rnge_Last => 32),
    Bark_B10 =>(Harmonic_Rnge_Frst => 33, Harmonic_Rnge_Last => 38),
    Bark_B11 =>(Harmonic_Rnge_Frst => 39, Harmonic_Rnge_Last => 44),
    Bark_B12 =>(Harmonic_Rnge_Frst => 45, Harmonic_Rnge_Last => 52),
    Bark_B13 =>(Harmonic_Rnge_Frst => 53, Harmonic_Rnge_Last => 60),
    Bark_B14 =>(Harmonic_Rnge_Frst => 61, Harmonic_Rnge_Last => 70),
    Bark_B15 =>(Harmonic_Rnge_Frst => 71, Harmonic_Rnge_Last => 81),
    Bark_B16 =>(Harmonic_Rnge_Frst => 82, Harmonic_Rnge_Last => 95),
    Bark_B17 =>(Harmonic_Rnge_Frst => 96, Harmonic_Rnge_Last => 111),
    Bark_B18 =>(Harmonic_Rnge_Frst => 112,Harmonic_Rnge_Last => 132),
    Bark_B19 =>(Harmonic_Rnge_Frst => 133,Harmonic_Rnge_Last => 159),
    Bark_B20 =>(Harmonic_Rnge_Frst => 160,Harmonic_Rnge_Last => 192),
    Bark_B21 =>(Harmonic_Rnge_Frst => 193,Harmonic_Rnge_Last => 231),
    Bark_B22 =>(Harmonic_Rnge_Frst => 232,Harmonic_Rnge_Last => 285),
    Bark_B23 =>(Harmonic_Rnge_Frst => 286,Harmonic_Rnge_Last => 360),
    Bark_B24 =>(Harmonic_Rnge_Frst => 361,Harmonic_Rnge_Last => 465));
```

# Appendix B: An Experiment with Simultaneous Masking

## Introduction

To investigate masking, Egan and Hake [1950] employed a rigorous approach that was intended to substantiate the laws of perception; and though their results clearly demonstrated simultaneous masking, they also showed the mechanisms of perception to be variable. Rather than repeat Egan and Hake's method of masking a tone with a narrow band of noise, here a simplified method of measuring the effect of a powerful pure tone on a nearby less powerful pure tone was evaluated.

The rationale for the test was that the level of masking at any particular frequency would be similar to the level of the unmasked waveform (i.e unaffected by the Masker) plus the amount of boost that must be provided to the masked waveform to maintain the same datum — which is the level at which the potentially masked waveform just ceases to be audible.

## Method

The experiment used a masker sine-wave of 570 Hz (Bark band 6 centre frequency) of magnitude of $\sim$ -5.0 dB and a (potentially masked) sine-wave which was variable in magnitude in steps of between -20 and -52 dB. (The reference level of 0dB was arbitrarily chosen to be 2ˆ13.) The masker/masked sine waves comprised approximately six seconds of the masker, with the potentially masked waveform added into the middle two seconds. Thus should the lesser of the two tones not be masked, then a transition and change in the sound would be perceived in the middle of the test.

The experiment was mechanised by having multiple copies of the masker/masked waveform, and then listening to them in order of -20dB down to -52dB to the level at which the potentially masked waveform was no longer perceptible, and recording the level above this. This entire process was repeated for all potential masked frequencies in steps of 10 Hz between 360 Hz and 840 Hz.

### Results

Figure B.1 shows the extent for which it was necessary to boost each of the frequencies such that they were each just perceptible.

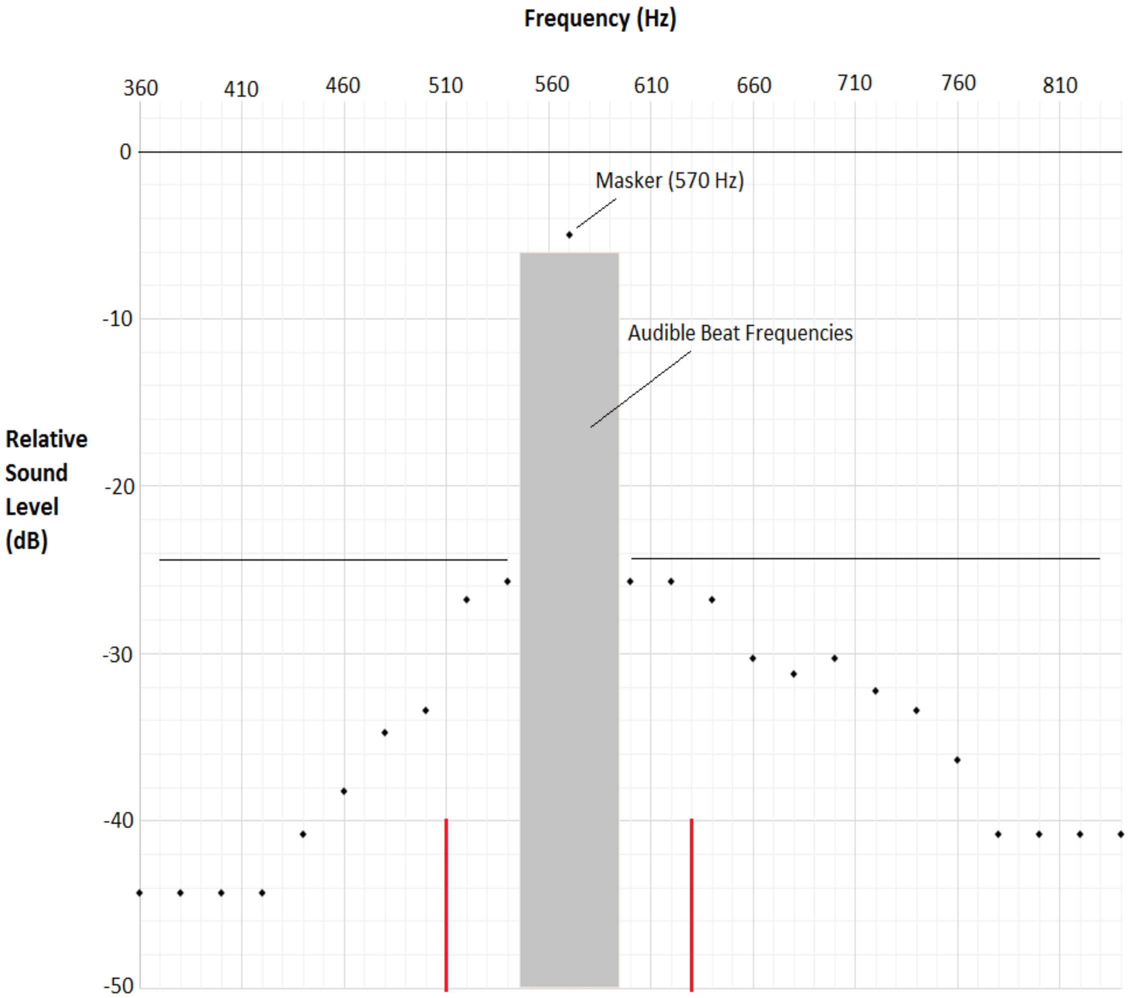


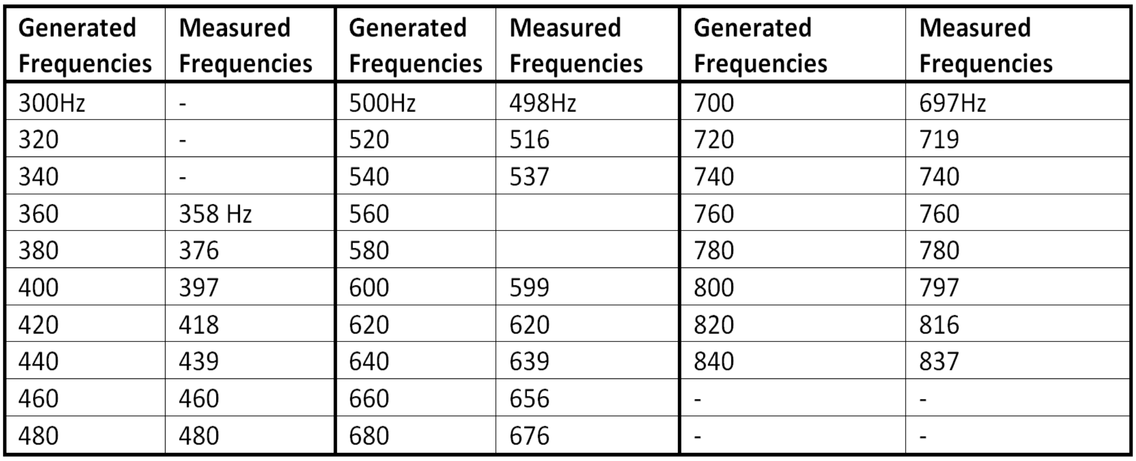
Figure B.1: Simultaneous Masking About 570 Hz Centre Frequency.

The boost required to frequencies about a Bark band centre frequency of 570 Hz (the masker), with concomitant Bark band limits of 510 Hz and 630 Hz indicated by the red vertical bars. That there is a step change in the amount of boost that must be applied at 510 Hz and 650 Hz (6 dB and -3.8 dB respectively) provides some evidence of the Bark (critical) band. Note though that the overall masking effect extends well beyond 510 Hz to 630 Hz.

### Limitations of the Experiment

With this test, when the potentially masked sine-waves were nearer than about 15 Hz to the masker frequency then a beat frequency was clearly audible. So the test cannot be used to ascertain the effects of masking when the frequency of the potentially masked tones is near to that of the masker.

The correctness of the test frequencies was verified with a spectrum analysis of the mixed frequency part of the test audio (see Table B.1), but otherwise the test system was not calibrated, nor was an independent method of measuring the sound level used. So the results are ad hoc, and can only be used to add a measure of support to the work of others.



| Generated Frequencies | Measured Frequencies | Generated Frequencies | Measured Frequencies | Generated Frequencies | Measured Frequencies |
|---|---|---|---|---|---|
| 300Hz | - | 500Hz | 498Hz | 700 | 697Hz |
| 320 | - | 520 | 516 | 720 | 719 |
| 340 | - | 540 | 537 | 740 | 740 |
| 360 | 358 Hz | 560 | | 760 | 760 |
| 380 | 376 | 580 | | 780 | 780 |
| 400 | 397 | 600 | 599 | 800 | 797 |
| 420 | 418 | 620 | 620 | 820 | 816 |
| 440 | 439 | 640 | 639 | 840 | 837 |
| 460 | 460 | 660 | 656 | - | - |
| 480 | 480 | 680 | 676 | - | - |

The performance of the frequency generator was verified by using the Audacity® [2014] Spectrum Analyser automatic peak tracker on the second greatest peak in the spectrum, using an FFT window size of 4096 samples.(It was not possible to discriminate the spectral peaks at frequencies of 560Hz and 580Hz from the masker peak of 570Hz.)

Table B.1: Masking Test Frequencies as Verified by Spectrum Analysis

### Discussion

The psychoacoustic effect of simultaneous masking is clearly observed in the results, where the magnitude of the potentially masked frequency must be increased if it is to be just barely audible, as the frequency of the potentially masked waveform is stepped nearer to that of the masker.

When the potentially masked frequency is within a few Hz of the masker a beat frequency is audible, and thus it is not possible to estimate the level of masking. However although the potentially masked frequency may or may not be masked, the beat frequency is perceived by the subject and so the potentially masked frequency still maintains a presence in the perceived audio.

It is not always possible to know whether the combination of the masker with the just perceptible masked frequency is correct. i.e. Perceiving whether or not the masker carries a second frequency is far less difficult than perceiving the actual nature of the second frequency.

That there was a step change in the amount of boost required at the lower and near the upper Bark band limits of 510 Hz and 630 Hz (510 Hz and 650 Hz as measured) provides some evidence of the Bark (critical) band; but the full extent of the upper frequency masking encompasses half of band 6 plus most of band 7, whereas the lower masking encompasses half of band 6 plus most of band 5.

In conclusion then, the psychoacoustic effect of masking can be demonstrated to an extent using just two pure tones, but it is not possible with this method of forming a total picture of just what is going on. The results of Egan and Hake [1950] (Figure 1) show the extent of masking is not only a function of frequency but is also a function of magnitude, and since these are both linearly variable then there must exist an infinite number of masking scenarios — even with just two tones.

# Appendix C: Experiments with MP3 Compressed Speech

## C.1 Introduction

This work differs from that presented in the main body of the thesis in several key respects:

- The speech corpus was encoded in the MP3 Lossy Compression Audio Format. A requirement of the relevant standard (MP3-Standard [1995]) is that MP3 encoded audio is capable of being correctly decoded by the Standardised Decoder; and although the Standard provides descriptions of suitable psychoacoustics models there is no requirement to use these models. This means that the configuration of the psychoacoustics models for any given MP3 audio source is unknown. The LogFB$_{dynamic}$ implements a greatly simplified psychoacoustics model, and though it seems intuitively correct that it is merely duplicating part of the MP3 psychoacoustics processing, the extent to which the processes conflict is unknown.

- A different paradigm was used for selecting the speech and silence training data.

- A 3 layer supervised ANN was used as the final speech/silence binary classifier.

- Natural logarithms were used throughout for generating the log energies (except for the Signal to Noise ratio calculations).

## C.2 Selection Of Corpus

The test corpus comprised 8 extracts from ‘talking’ books; and these were read by four adult females and four adult males. The eight extracts — each with a different signal to noise ratio — were all of about 15 minutes duration, and were sectioned into train, validate and test data sets, to support split sample testing. (Table C.1)

| Test-Case | ∼Noise Level | Train | Validate | Test |
|---|---|---|---|---|
| TC1-N (F) | 37.6 dB | Pt1 305.931 | Pt3 342.234 | Pt2 285.261 |
| TC2-C (M) | 37.9 dB | Pt1 303.725 | Pt2 296.489 | Pt3 304.596 |
| TC3-J (M) | 43.6 dB | Pt2 300.566 | Pt3 312.322 | Pt1 299.401 |
| TC4-L (M) | 41.6 dB | Pt1 301.355 | Pt2 299.473 | Pt3 306.693 |
| TC5-D (F) | 34.8 dB | Pt3 313.737 | Pt1 301.024 | Pt2 300.892 |
| TC6-T (M) | 35.6 dB | Pt3 334.680 | Pt1 301.209 | Pt2 300.888 |
| TC7-W (F) | 58.1 dB | Pt1 301.662 | Pt3 435.211 | Pt2 306.383 |
| TC8-JE (F) | 25.2 dB | Pt3 320.457 | Pt2 303.489 | Pt1 297.339 |

The texts and the audio were divided into three approximately equal parts and then the parts were assigned to be one of Training, Validate or Test data.
The number following the speech part identifier is the duration of that part in seconds. (The ordering of the parts as training data, validation data and test data was varied to negate the effect of any systematic variation in the speech as the recording progressed.)

Table C.1: Partitioning of the Test-Cases into the Training, Validation and Test Data Sets

### Format of the Speech Files

To ensure that the software processing would not artificially limit the bandwidth or otherwise compromise the quality of the recordings, the D$_{eterm}$Classifier, the LogFB$_{dynamic}$ and the ANN_Classifier were coded to accept only recordings in the 44100 sps, 16 bit signed, uncompressed WAVE Format. To accommodate this, all of the MP3 speech recordings were re-sampled to the WAVE Format (without recourse to the acoustic domain), using a DP004 digital recorder (Tascam™ [2017]). This method was used in preference to using some proprietary conversion utility because it eliminated any direct digital to digital conversion with attendant potential for aliasing or corruption of the recording.

## C.3 ANN Training Data Preparation

There are areas of ambiguity around the start and end of all silences identified by the D$_{eterm}$Classifier. and Figure C.1 identifies this ambiguous data — which is excluded from the training data set — and the speech and silence training data. The ambiguous data blocks are each arbitrarily allocated a duration of $90ms$, and the speech training data blocks each a duration of $500ms$. The duration of the silence training blocks are equal to the durations measured by the Non-Statistical Silence/Speech Classifier minus the duration of one ambiguous data block. The ambiguous data is not used for training.

The purpose of this approach is to reduce to some extent the conflict which will arise when silence like data appears in the speech training data and vice versa.

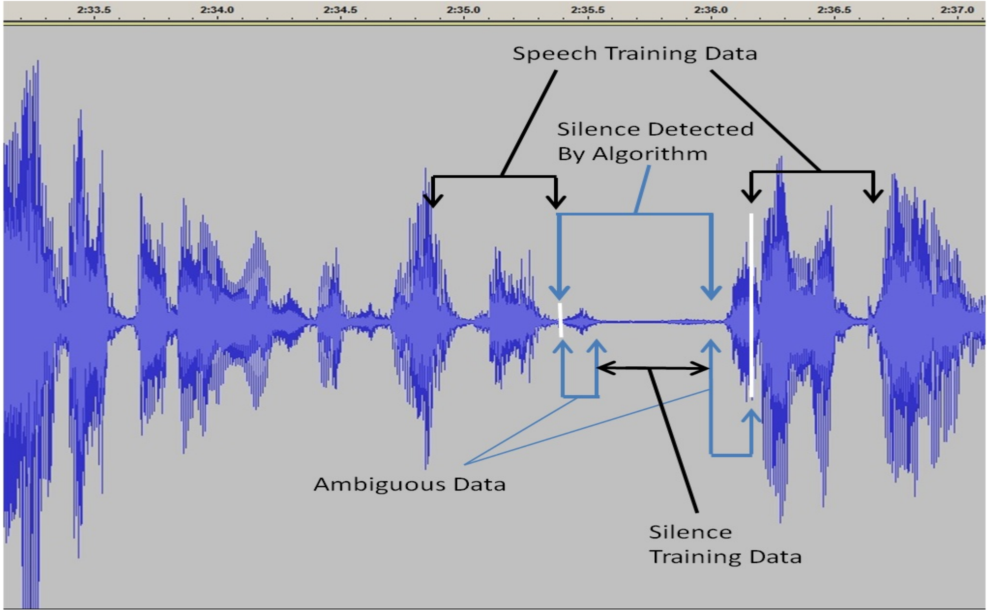


Figure C.1: Speech and Silence Qualifiers for Supervised Training.

### C.3.1 Configuration

The ANN was configured with 27 input neurons, 1 hidden layer of 54 neurons with the sigmoid activation function (range 0.0–1.0), and 1 output neuron; and trained using the Resilient Propagation (RPROP) back-propagation algorithm (Reidmiller and Braun [1993]). The configuration of the input layer was dictated by the number of terms in the speech model, and of the output layer by the requirement for an unambiguous speech/silence binary decision. For the hidden layer five potential configurations were tested; and the configuration that was adopted — 1 hidden layer of 54 neurones — provided the most consistent classification results (i.e. duration of detected silence) for 4 complete training/execution cycles with the TC1-N data-set. The other configurations tested comprised, 1 hidden layer of 13 neurones, 1 hidden layer of 27 neurones, 1 hidden layer of 81 neurones, and 2 hidden layers each of 27 neurones.

The speech endpoint ambiguity span and the speech onset ambiguity span were both fixed at 90 milliseconds, and the silence example span, and clause example span were both set to a maximum of 500 milliseconds.

### C.3.2 Test Method

With the exception of how the speech and silence training data was defined, and excluding the generalisability and consistency tests, the test method was as described in Section 4.8.2.

## C.4 Results

### C.4.1 The D$_{eterm}$Classifier

There were no instances of false clause starts for any of the test-cases. That is, all instances of early clause onsets culminated in speech segments. Thus all of the silence pauses for all test-cases — when set to true zero on the speech recordings — were free of noise.

Where a test-case is marked to be non-viable as a speech recording, this means that part of a word, or many trailing consonants are missing, such that a listener would notice the errors.

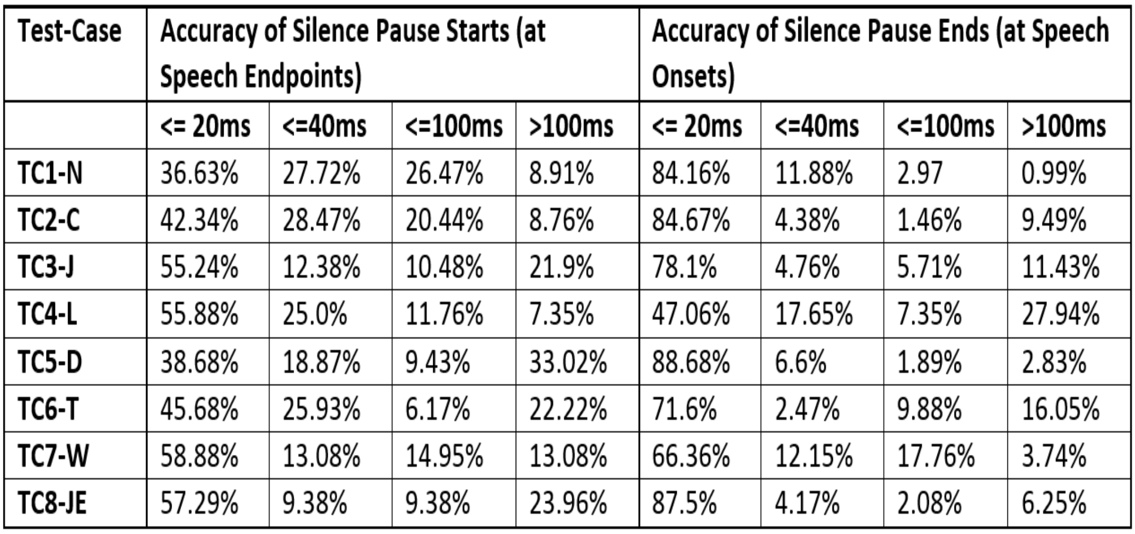
#### Auto-Detected Silence Pauses Vs Perceived Silence Pauses: Insertions and Deletions; For the D$_{eterm}$Classifier:

Table C.2 compares the Silence Pause total from the D$_{eterm}$Classifier with the Perceived Silence Pauses, and also identifies the total of extra silence detections: that is silences that are not perceived as such by the listener such as those occurring at stops.

| Test-Case | Perceived Pauses (Ground-Truth) | Auto Detected Silence Pauses | Filled Silence Pauses | Silence Pause Deletions | Silence Pause Insertions |
|---|---|---|---|---|---|
| **TC1-N** | 110 | 101 | 2 | 7 (6.36%) | 14 |
| **TC2-C** | 143 | 137 | 4 | 2 (3.5%) | 24 |
| **TC3-J** | 120 | 105 | 1 | 14 (11.67%) | 34 |
| **TC4-L** | 90 | 68 | 0 | 22 (24.5%) | 2 |
| **TC5-D** | 118 | 106 | 0 | 12 (9.3%) | 22 |
| **TC6-T** | 99 | 81 | 0 | 18 (19.2%) | 44 |
| **TC7-W** | 114 | 107 | 0 | 7 (5.2%) | 15 |
| **TC8-JE** | 106 | 96 | 0 | 10 (9.5%) | 48 |

Table C.2: Silence Insertions and Deletions for the D$_{eterm}$Classifier

**Auto Detected Silence Pauses Vs Perceived Silence Pauses: Pauses Start and End temporal Accuracy; For the D$_{eterm}$Classifier**



| Test-Case | Accuracy of Silence Pause Starts (at Speech Endpoints) | | | | Accuracy of Silence Pause Ends (at Speech Onsets) | | | |
|---|---|---|---|---|---|---|---|---|
| | <= 20ms | <=40ms | <=100ms | >100ms | <= 20ms | <=40ms | <=100ms | >100ms |
| TC1-N | 36.63% | 27.72% | 26.47% | 8.91% | 84.16% | 11.88% | 2.97 | 0.99% |
| TC2-C | 42.34% | 28.47% | 20.44% | 8.76% | 84.67% | 4.38% | 1.46% | 9.49% |
| TC3-J | 55.24% | 12.38% | 10.48% | 21.9% | 78.1% | 4.76% | 5.71% | 11.43% |
| TC4-L | 55.88% | 25.0% | 11.76% | 7.35% | 47.06% | 17.65% | 7.35% | 27.94% |
| TC5-D | 38.68% | 18.87% | 9.43% | 33.02% | 88.68% | 6.6% | 1.89% | 2.83% |
| TC6-T | 45.68% | 25.93% | 6.17% | 22.22% | 71.6% | 2.47% | 9.88% | 16.05% |
| TC7-W | 58.88% | 13.08% | 14.95% | 13.08% | 66.36% | 12.15% | 17.76% | 3.74% |
| TC8-JE | 57.29% | 9.38% | 9.38% | 23.96% | 87.5% | 4.17% | 2.08% | 6.25% |

Table C.3: Silence Pause Start and End Measurement Accuracy: D$_{eterm}$Classifier.

Table C.3 is the result of a comparison of the silence pause start and end times as detected by the D$_{eterm}$Classifier with those of the perceived silence pauses.

**Classification Failures for the D$_{eterm}$Classifier:**

The failures of classification for each of the test-cases are described below, and the failure data is collated in Table C.4.

1. **TC1-N:**
   Trailing Consonant Deletions : 1 – /t/

2. **TC2-C:**
   Trailing Consonant Deletions : 10 – /t/, /t/, /k/, /k/, /k/, /t/, /s/, /k/, /ge/, /ge/.

3. **TC3-J:**
   Trailing Consonant Deletions : 6 – /d/, /p/, /k/ (i.e. /c/ in arithmetic), /t/, /d/, /t/.

‘Ting’ missing at the end of the phrase, ‘To be kept waiting.’

‘Ter missing at the end of the phrase, ‘A glass of water.’

4. **TC4-L:**

   —

5. **TC5-D:**

   Trailing Consonant Deletions : /ts/, /ps/, /t/, /tch/, /ps/.

6. **TC6-T:**

   Trailing Consonant Deletions : 1 – /p/.

   ‘And saying there was the sort of man’ missing at the end of the recording.

7. **TC7-W:**

   —

8. **TC8-JE:**

   Trailing Consonant Deletions : /tt/, /s/, /ce/, /k/, /s/, /sa/, /s/, /s/, /ce/, /x/, /ch/.

   ‘Night’ missing at the end of the phrase, ‘Early that night.’

   ‘It’ missing mid phrase in , ‘Liked it better.’

   ‘Ted’ missing at the end of the phrase, ‘Where I knew she was not wanted.’

   ‘Nts he’ missing mid phrase in , ‘What presents he brought her.’

   ‘Sake’ missing at the end of the phrase, ‘Parting keepsake.’

   ‘Piece’ missing at the end of the phrase, ‘two on the mantelpiece.’

| **Test-Case** | **Deleted Consonants** | **Deleted Syllables or Words** | **Deleted Clauses** | **Viable?** |
|---|---|---|---|---|
| TC1-N (F) | 1 | 0 | 0 | Yes |
| TC2-C (M) | 10 | 0 | 0 | Yes |
| TC3-J (M) | 6 | 2 | 0 | No |
| TC4-L (M) | 0 | 0 | 0 | Yes |
| TC5-D (F) | 5 | 0 | 0 | No |
| TC6-T (M) | 1 | 0 | 1 | No |
| TC7-W (F) | 0 | 0 | 0 | Yes |
| TC8-JE (F) | 11 | 6 | 0 | No |

Table C.4: Errors for the D$_{eterm}$Classifier.

### C.4.2 ANN_Classifier

#### Auto-Detected Silence Pauses Vs Perceived Silence Pauses: Insertions and Deletions. For the ANN_Classifier

Table C.5 shows the Perceived Silence Pauses Vs the Silence Pauses automatically detected by the Supervised ANN_Classifier. Note that in Table C.5 — *Silence Insertions*, a single count is used to represent what may be a cluster of short silence bursts in the same location.

| Test-Case | Perceived Pauses (Ground-Truth) | Auto Detected Silence Pauses | Filled Silence Pauses | Silence Pause Deletions | Silence Pause Insertions |
|---|---|---|---|---|---|
| TC1-N (F) | 110 | 107 | 2 | 1 (0.91%) | 253 |
| TC2-C (M) | 143 | 138 | 4 | 1 (0.7%) | 148 |
| TC3-J (M) | 120 | 119 | 1 | 0 | 196 |
| TC4-L (M) | 90 | 89 | 0 | 1 (1.11%) | 214 |
| TC5-D (F) | 118 | 118 | 0 | 0 | 490 |
| TC6-T (M) | 99 | 99 | 0 | 0 | 276 |
| TC7-W (F) | 114 | 112 | 0 | 2 (1.75%) | 224 |
| TC8-JE (F) | 106 | 106 | 0 | 0 | 540 |

Table C.5: Silence Insertions and Deletions for the ANN_Classifier

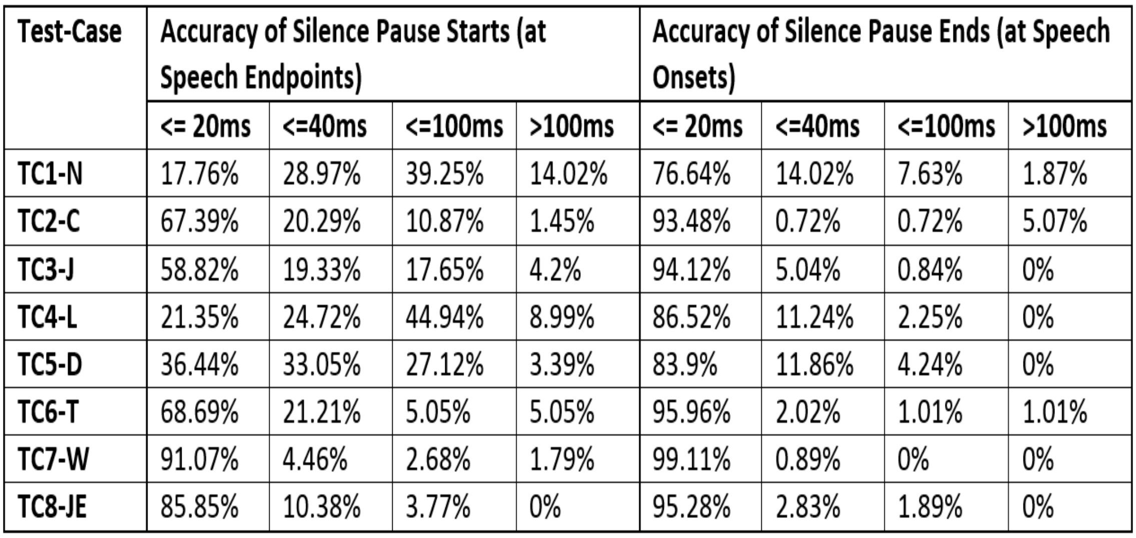
#### Auto-Detected Silence Pauses Vs Perceived Silence Pauses: Pauses Start and End temporal Accuracy. For the ANN_Classifier:

Table C.6 indicates a measure of the accuracy of the silence-pause starts (speech endpoint detection) and silence pause ends (speech onset detection) for the Supervised ANN_Classifier.

Tables C.4 and C.7 are the results of a qualitative assessment of the speech.

#### Classification Failures for the ANN_Classifier

For the Supervised ANN Binary Classifier (Table C.7) although there was only one deleted consonant, for TC1-N, TC2-C, and TC4-L there were a few degraded trailing /s/ consonants, and particularly for TC6-T and TC8-J there was uncertainty in the location of silence associated with the fricative consonant /f/.



| Test-Case | Accuracy of Silence Pause Starts (at Speech Endpoints) | | | | Accuracy of Silence Pause Ends (at Speech Onsets) | | | |
|---|---|---|---|---|---|---|---|---|
| | <= 20ms | <=40ms | <=100ms | >100ms | <= 20ms | <=40ms | <=100ms | >100ms |
| TC1-N | 17.76% | 28.97% | 39.25% | 14.02% | 76.64% | 14.02% | 7.63% | 1.87% |
| TC2-C | 67.39% | 20.29% | 10.87% | 1.45% | 93.48% | 0.72% | 0.72% | 5.07% |
| TC3-J | 58.82% | 19.33% | 17.65% | 4.2% | 94.12% | 5.04% | 0.84% | 0% |
| TC4-L | 21.35% | 24.72% | 44.94% | 8.99% | 86.52% | 11.24% | 2.25% | 0% |
| TC5-D | 36.44% | 33.05% | 27.12% | 3.39% | 83.9% | 11.86% | 4.24% | 0% |
| TC6-T | 68.69% | 21.21% | 5.05% | 5.05% | 95.96% | 2.02% | 1.01% | 1.01% |
| TC7-W | 91.07% | 4.46% | 2.68% | 1.79% | 99.11% | 0.89% | 0% | 0% |
| TC8-JE | 85.85% | 10.38% | 3.77% | 0% | 95.28% | 2.83% | 1.89% | 0% |

Table C.6: Silence Pause Start and End Measurement Accuracy for the ANN_Classifier.

Additionally, for all test-cases, there were a few erroneous silence insertions in what was clearly speech. These silence insertions, most with a duration of less than 5 mS — and some as short as 1 mS — are audible. There were also inaudible silence insertions in the lower energy sections of dwindling weak unstressed sounds at the end of clauses.

1. **TC1-N:**
   16$ms$ burst of silence insertions in the word 'He'.
   16$ms$ silence insertion in the word 'Easy'.
   8$ms$ silence insertion in the word 'Dear'.
   1$ms$ silence insertion in the word 'Beg'.
   1$ms$ silence insertion in the word 'Afternoon'.

2. **TC2-C:**
   1$ms$ silence insertion in the words 'I have'.
   1$ms$ silence insertion in the word 'Afternoon'.
   13$ms$ burst of silence insertions in the words 'A happy'.

3. **TC3-J:**
   1$ms$ silence insertion in the word 'Harry'.

4. **TC4-L:**
   1$ms$ and 2ms silence insertions in the word 'No other'.

5. **TC5-D:**

   28$ms$ burst of silence insertions in the word ‘That’.

6. **TC6-T:**

   None.

7. **TC7-W:**

   12$ms$ silence insertion in the words ‘Get’.

   23$ms$ burst of silence insertions in the word ‘Were’.

   1$ms$ silence insertion in the word ‘If’.

8. **TC8-JE:**

   None.

| **Test-Case** | **Deleted Consonants** | **Deleted Clauses** | **Viable?** |
|---|---|---|---|
| TC7-W (F) | 0 | 0 | Yes |
| TC3-J (M) | 0 | 0 | Yes |
| TC4-L (M) | 1 | 0 | Yes |
| TC2-C (M) | 0 | 0 | Yes |
| TC1-N (F) | 0 | 0 | Yes |
| TC6-T (M) | 0 | 0 | Yes |
| TC5-D (F) | 0 | 0 | Yes |
| TC8-JE (F) | 0 | 0 | Yes |

Table C.7: Errors for the ANN_Classifier

## C.5 Conclusion

Against all measures, the performance of the ANN_Classifier exceeded that of the D$_{eterm}$Classifier; and in absolute terms, the ANN_Classifier correctly identified better than 98% of the silence pauses for all eight test-cases. However, the original source for the audio recordings have available many thousands of different works, read by several thousand individuals; and so with only eight recordings read by eight individuals the results are not statistically significant. What is of significance is that of the two part classification process, it is the D$_{eterm}$Classifier which suffers increasing error with increasing noise, whereas the D$_{eterm}$Classifier/ANN_Classifier exhibits no particular sensitivity to noise. That is, for the latter case, there seems to be no correlation between the classification accuracy and the signal to noise ratio of the original source. Further to this the performance of both classifiers exhibits no dependency upon the gender of the speaker.

# Appendix D: The CD ROM

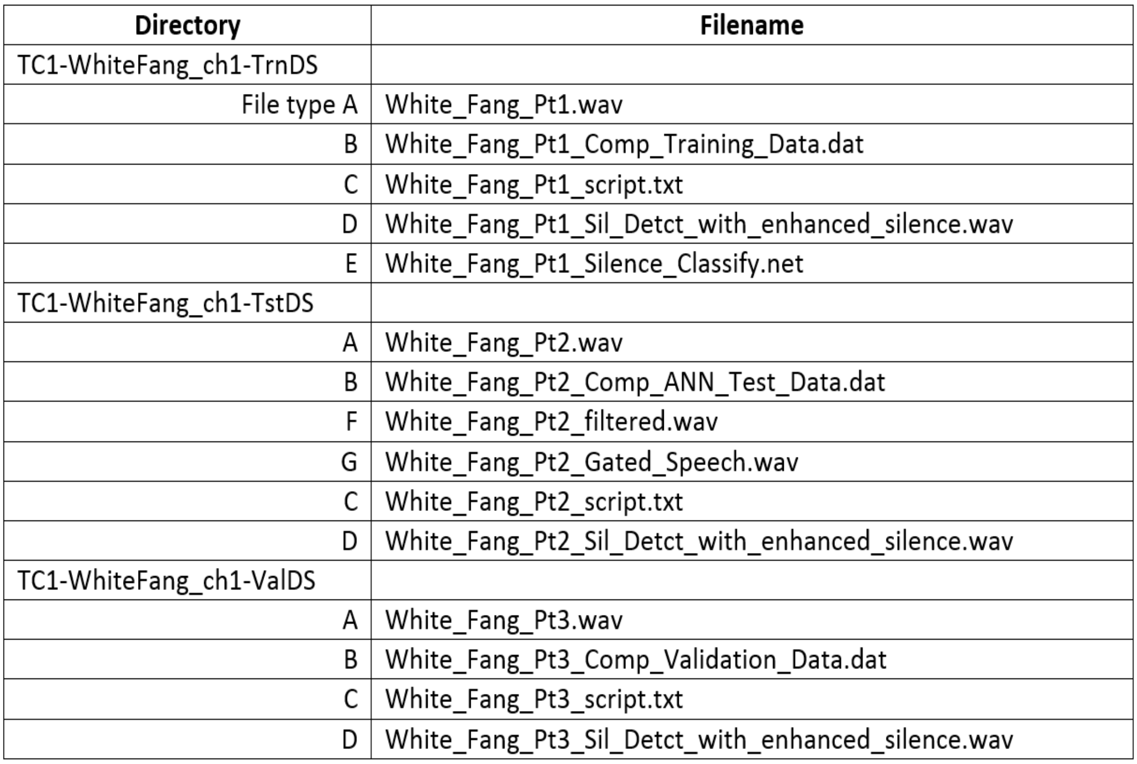
The test-case directory and file structure is as shown in Table D.1; and for each of the test-cases the following information is included.

**File Type A:** Original Audio Source
**File Type B:** The Audio Model Data
**File Type C:** The text of the Original Audio Source
**File Type D:** Filtered Audio with the locations of the silence pauses as detected by the D$_{eterm}$Classifier overwritten with zero.
**File Type E:** The trained DNN.
**File Type F:** The Original Audio Source after it has been decomposed into harmonics, and reconstructed less the first few harmonics. (As defined in Table 4.4.)
**File Type G:** Filtered Audio with the locations of the silence pauses as detected by the DNN_Classifier overwritten with zero.

Note that *.dat and *.net files are readable text files.

The result of the D$_{eterm}$Classifier is demonstrated in File Type D.

The outcome of the process is File Type E, the trained DNN, and the efficacy of the trained DNN as a classifier is demonstrated in File Type G.



| Directory | Filename |
|---|---|
| TC1-WhiteFang_ch1-TrnDS | |
| File type A | White_Fang_Pt1.wav |
| B | White_Fang_Pt1_Comp_Training_Data.dat |
| C | White_Fang_Pt1_script.txt |
| D | White_Fang_Pt1_Sil_Detct_with_enhanced_silence.wav |
| E | White_Fang_Pt1_Silence_Classify.net |
| TC1-WhiteFang_ch1-TstDS | |
| A | White_Fang_Pt2.wav |
| B | White_Fang_Pt2_Comp_ANN_Test_Data.dat |
| F | White_Fang_Pt2_filtered.wav |
| G | White_Fang_Pt2_Gated_Speech.wav |
| C | White_Fang_Pt2_script.txt |
| D | White_Fang_Pt2_Sil_Detct_with_enhanced_silence.wav |
| TC1-WhiteFang_ch1-ValDS | |
| A | White_Fang_Pt3.wav |
| B | White_Fang_Pt3_Comp_Validation_Data.dat |
| C | White_Fang_Pt3_script.txt |
| D | White_Fang_Pt3_Sil_Detct_with_enhanced_silence.wav |

Table D.1: CD File Structure

# Appendix E: The Development Environment

**Primary Development Environment:**

ASUS™ N55S Laptop (ASUSTek Computer Inc) with an Intel® Core™ i7-2670QM CPU @ 2.20GHz (Intel Corporation), running the Windows™ 7 Home Premium/Service Pack 1, 64 bit Operating System (Microsoft Corporation).

**Alternative Development Environment:**

Viglen™ Desktop (XMA Ltd.) with an Intel® Core™ i7-3770S CPU @ 3.40GHz (Intel Corporation), running the Windows™ 7 Enterprise, 64 bit Operating System (Microsoft Corporation).

**Support/Backup Development Environment:**

ASUS™ Desktop (ASUSTek Computer Inc) with an Intel® Core™ i7-3770S CPU @ 3.10GHz x 8 (Intel Corporation), running the CentOS™ 7 (Red Hat Inc.) 64 bit operating System.

**Archive:**

Standalone Seagate 3.63 TB HDD (Seagate Technology LLC) using the NTFS File System (Microsoft Corporation).

**Thesis Preparation:**

This document was prepared using LaTeX release MikTex Version 2.9 (https://miktex.org © 2018 Christian Schenk) with TexStudio Version 2.12.4 (© van der Zander et al. http://www.texstudio.org/), with many tables and graphs imported from Microsoft Office 2013 (Microsoft Corporation), and Figure 3.5 imported from GNUPlot Version 5.0 patchlevel 3 (© 1986-1993, 1998, 2004, 2007-2016, Thomas Williams, Colin Kelley and many others. http://www.gnuplot.info)

# Glossary

**Abbreviations and Acronyms** :

| | |
|---|---|
| ANN | Artificial Neural Network |
| ASR | Automatic Speech Recognition |
| CD | Compact Disk |
| CSSD | Cumulative Sample Standard Deviation |
| dB | Decibel |
| DCT | Discrete Cosine Transfer |
| DFT | Discrete Fourier Transfer |
| DNN | Deep Neural Network |
| FFT | Fast Fourier Transform |
| ERB | Equivalent Rectangular Bandwidth |
| $F_0$ | Fundamental Frequency (of Voiced speech) |
| HMM | Hidden Markov Model |
| LPC(C) | Linear Predictive Coding (Coefficients) |
| MFC(C) | Mel Frequency Cepstrum (Coefficients) |
| MLP | Multi Layer Perceptron |
| MSE | Mean Squared Error |
| PDF | Probability Density Function |
| RBFN | Radial Basis Function Network |
| ROM | Read Only Memory |
| RCF | Robust Categorical Features |
| RMS | Root (of the) Mean (of the) Squared |
| RPROP | Residual Propagation |
| S/N | Signal to Noise (Ratio) |
| SPS/sps | Samples Per Second |
| SSD | Sample Standard Deviation |
| SVM | Support Vector Machine |
| TC | Test-Case |
| VAD | Voice Activity Detection |

**Acoustic Frequency Scale** (Johnson [2012]) : A linear scale for measuring measured frequency in cycles per second in SI units of Hertz (Hz)

**Aperiodic** (Chambers [1999]) : No periodicity; decaying to rest without oscillation.

**Auditory Frequency Scale** (Johnson [2012]) : Any non linear scale for representing frequency that is based upon the empirically derived frequency response of the human auditory physiology. e.g. The Bark and Mel frequency scales.

**Autocorrelation coefficient** : In the Oxford Dictionary of Computing [2008] Pg. 114, correlation is defined to be the degree to which two random variables are associated. Autocorrelation refers to the specific case where the correlation is between one variable and a later version of itself. According to Huang et al. [2001] Pg. 324, autocorrelation is commonly used in the estimation of pitch.

**Baseline** : "...a standard of comparison" — Chambers [1999].

**Bayesian Classifier** (Mlodinow [2009])(Virtanen et al. [2013]) : Based upon the presumption that with full cognizance of the circumstances that surround a problem, the most probable solution identified will be the correct solution. In probability theory, when new information about a problem becomes available, then that information can be used to reduce the sample space (by limiting the possibilities), and update probabilities. This is used in speech processing where as each entity is identified, then the probabilities of what might follow are adjusted, in accord with the rules of the juxtaposition of elements in speech - the language model (also known as a finite-state or context free grammar, or a statistical N-gram model).

**Cepstrum** : "The spectrum of the log of the spectrum." — Jurafsky and Martin [2009].

**Corpus (corpora)** : "... a body of literature, writings, etc; the main part of anything..." —Chambers [1999].

**Deterministic** (Pg 142, Oxford Dictionary of Computing [2008]) : An algorithm, the output of which is determined by the initial state and the inputs.

**Diphthong** (Chambers [1999]) : Two vowels pronounced as a single syllable (e.g. as in the word 'out').

**Dynamic** : “Capable of changing or being changed...” — Oxford Dictionary of Computing [2008].

**Excitation Function** (pg 25: Gold et al. [2011]) : Speech may be viewed as the sound which results when the sound produced by the vibration of the vocal chords (the Excitation Function) is modified by the configurable resonances and obstructions in the vocal tract.

**Fricative** (Johnson [2012]): Turbulent aperiodic sound which is the result of airflow through constrictions in the vocal tract.

**Fuzzy Logic** (Pg 214 Oxford Dictionary of Computing [2008]) : Rather than constraining logic to the states of TRUE or FALSE, Fuzzy Logic provides a multi-valued logic where the logic can indicate the degree of truth.

**Forced Alignment**: See Viterbi.

**Formant** : Formants are resonances in the vocal tract. According to Johnson [2012] (pg 142) vowels may be distinguished from one another by comparing the frequencies of the first and second of the formants.

**Generalise** : “...to comprehend as a particular case within a wider concept, proposition, definition etc...” — Chambers [1999].

**Glottal Source**: (Pg. 255, Jurafsky and Martin [2009]). The vocal folds.

**Heuristic**: (Chambers [1999]) “consisting of guided trial and error”; “depending on assumptions based on past experience”.

**Monotonic**: “(of a function or sequence) having the property of either never increasing or never decreasing (math)” — Chambers [1999].

**MP3 Lossy Audio Compression Format** (Brandenburg [1999]): MP3 is the short form for the Moving Pictures Experts Group MPEG 1/2 Layer 3 lossy audio compression format (MP3-Standard [1995]). Lossy compression, is where information that is not perceived by the listener because of the psychoacoustics effect of masking, is removed from the audio. The intention with MP3 is to compress the audio to the maximum whilst preserving the sound quality. The psychoacoustics model — referred to by Brandenburg [1999], and in MP3-Standard [1995] as the perceptual model — uses a filter-bank and generates masking thresholds for each of the MP3 encoder partitions. These partitions are approximately the same as the critical bands which span the audible frequency range. The MP3-Standard [1995] describes two perceptual models, but these are not prescriptive and the designer of an MP3 encoder is free to implement any perceptual model, though Brandenburg [1999] writes, **"A lot of experience and knowledge is necessary to implement good quality MPEG audio encoders."**.

**Neurological** (Oxford Dictionary of Computing [2008]) : Of the function and structure of the brain.

**Normalisation** ( Oxford Dictionary of Computing [2008]) : The reorganisation of data so that it conforms to a higher normal form.

**NTIMIT** : "...to provide a telephone bandwidth adjunct to TIMIT." — Fisher et al. [1993]

**Paradigm** : "A model or example of the environment" — Oxford Dictionary of Computing [2008]; "a conceptual framework within which scientific theories are constructed" — Chambers [1999].

**Periodic** : "...recurring regularly in the same order..." — Chambers [1999].

**Pitch** : "...the degree of acuteness of a sound that makes it a high or a low, etc, note..." — Chambers [1999]

**Real-Time System** (Oxford Dictionary of Computing [2008]) : "Any system where the time at when the output is produced is significant."

**Sibilants** (pg 155, Johnson [2012]) : The unvoiced fricative sounds, /s/ and /sh/.

**Spectral Distortion** (G.729 [2012]) : is a measure of the difference between the Line Spectral Frequencies (as derived from the LPC coefficients) for the current frame and the running average for the background noise.

**Temporal** (Chambers [1999]) : "relating to time"

**TIMIT** : "The TIMIT corpus of read speech is designed to provide speech data for acoustic-phonetic studies and for the development and evaluation of automatic speech recognition systems." — Garofolo et al. [1993]

**Tonotopic** (Chambers [1999]) : According to the dictionary, one of the definitions for tone is , "Rise or fall in pitch", and the medical definition for topical is, "Affecting only part of the body". Hence tonotopical.
Gold et al. [2011] attribute the use of 'tonotopic' — for having bundles of nerves that are sensitive to different frequencies — to H. von Helmholtz [1862]. That is the concept of frequencies within separate bands being processed together gives rise to the idea of the cochlea as a filter-bank.
Of the two definitions, the first perhaps better reflects the current usage of the word, because the spatial arrangement of sound processing in the brain — where different frequency bands stimulate different parts of the brain — may be labeled as tonotopic (Moerel et al. [2012], Langers [2014]).

**Transducer** : "A device that transfers power from one system to another in the same or in a different form." — Chambers [1999].

**Transient** (Chambers [1999]) : A short term burst or spike of energy in a waveform.

**Viterbi** : Whilst progressing through a sequence of Hidden Markov Models, the judgment on whether the ASR activity is correct is made by summing the probabilities of the individual states on the various paths, and the path with the highest probability is accepted to be the correct solution. It can be seen that there will be many paths, and to test these paths will involve many identical repeated calculations. This is an optimal path problem and can be solved by dynamic programming — using the Viterbi algorithm (Section 8.2.2 Huang et al. [2001]).

For Forced Alignment, the correct observable state/process templates are known in advance and they are concatenated so that the Viterbi algorithm has only the correct model from the outset and thus only one path to evaluate. Transitions at phoneme boundaries in the model are timestamped and these can be used to synchronise the speech with the phonetic transcript.

**White Noise** (pg 556, Oxford Dictionary of Computing [2008]) : Continuous in time, and magnitude, with uniform energy levels over equal frequency intervals.

**Zero Crossing Rate** (pg 81, Lea [1986]): Low frequency voiced sounds such as vowels have a much lower rate of zero crossings than high frequency unvoiced sounds, and so in low noise conditions it is possible to obtain an indication of the instantaneous frequency of voiced speech, and to discriminate between voiced speech and fricatives on the basis of number of zero crossings as a function of time.

**Wavelet Transform**(Pg 7, Addison [2002]): The Mexican Hat (Figure 3.4) is one example of a wavelet, and is defined as the second derivative of a Gaussian. The result of the convolving of a wavelet with a Continuous Time Varying function (CTV function) is a measure of the degree to which the frequencies in the wavelet exist in the CTV function. That is, each wavelet operates as a passband filter where the bandwidth of the filter is a function of the temporal duration of the wavelet. So if the process of convolving the wavelet with the CTV function is repeated for a set of wavelets each with a progressively shorter time span, then this will result in a data-set which indicates to what extent each particular band of frequencies exist in the CTV function. That is, the Wavelet Transform provides an alternative to the Discrete Fourier Transform for obtaining the frequency spectrum of a CTV function.

# Bibliography

Paul S. Addison. ***The Illustrated Wavelet Transform Handbook: Introductory Theory and Applications in Science, Engineering, Medicine and Finance***. Taylor and Francis, Abingdon, OX14 4RN, 2002. ISBN 13: 978-0-7503-0692-8.

T. V. Ananthapadmanabha, A. P. Prathosh, and A.G.Ramakrishnan. **Detection of the closure-burst transitions of stops and affricates in continuous speech using the plosion index**. *The Journal of the Acoustical Society of America*, 135(1), January 2014.

Poul-Erik Andreasen. **FANN Ada Bindings; ©Poul-Erik Andreasen**, 2015. URL `http://leenissen.dk/fann/wp/language-bindings`. Last accessed 2016-06-20.

J. P. Anu and V. Karjigi. **Sentence Segmentation for Speech Processing**. *National Conference on Communication, Signal Processing and Networking (NCCSN), IEEE*, 2014.

Jonathan F. Ashmore and Paul J. Kolston. **Hair cell based amplification in the cochlea**. *Current Opinion in Neurobiology*, 4:503–508, 1994.

Bishnu S. Atal and Lawrence R. Rabiner. **A Pattern Recognition Approach to Voiced-Unvoiced-Silence Classification with Applications to Speech Recognition**. *IEEE Transactions On Acoustics, Speech, and Signal Processing*, ASSP-2(3), June 1976.

Audacity®. **Audacity® free, open source, cross-platform software for recording and editing sounds; Version 2.0.6**, 2014. URL `http://audacityteam.org/`. Last accessed 2017-03-06; Audacity® is a registered trademark of Dominic Mazzoni.

M. Ben-Ari. ***Ada for Software Engineers***. John Wiley and Sons Ltd, Chichester, West Sussex. PO19 1UD, 1998. ISBN 0 471 97912 0.

F. Beritelli, S. Casale, G. Ruggeri, and S. Serrano. **Performance Evaluation and Comparison of G.729/AMR/Fuzzy Voice Activity Detectors**. *IEEE Signal Processing Letters*, 9(3):pp 85–88, 2002.

Steven F. Boll. **Suppression of Acoustic Noise in Speech Using Spectral Subtraction**. *IEEE Transactions On Acoustics, Speech, and Signal Processing*, ASSP-27(2), April 1979.

Karlheinz Brandenburg. **MP3 and AAC Explained**. *AES 17th International Conference on High Quality Audio Coding*, 1999.

G. Bristow. *Electronic Speech Recognition; Techniques, Technologies and Applications; Edited by G.Bristow.* William Collins Sons & Co. Ltd., 1986. ISBN 0-00-383238-4.

Sandrine Brognaux and Thomas Drugman. **HMM-Based Speech Segmentation: Improvements of Fully Automatic Approaches**. *IEEE Transactions On Acoustics, Speech, and Signal Processing ASSP*, 24(1):678–684, 2016.

D. Burileanu, L. Pascalin, C. Burileanu, and M. Puchia. **An Adaptive and Fast Speech Detection Algorithm**. *Speech and Dialogue, Volume 1902 of the series Lecture Notes in Artificial Intelligence (Sub-series of Lecture Notes in Computer Science)*, 1902:177–182, 2000.

Rich Caruana and Alexandru Niculescu-Mizil. **An Empirical Comparison of Supervised Learning Algorithms**. *ACM International Conference Proceeding Series*, 148:161–168, 2006.

Chambers. ***The Chambers Dictionary***. Chambers Harrap Publishers Ltd, Edinburgh, 1999. ISBN 0-550-10008-3.

C. H. Chen. **On Information and Distance Measures, Error Bounds, and Feature Selection**. *Information Sciences*, 10:159–173, 1976.

Corinna Cortes and Vladimir Vapnik. **Support Vector Networks**. *Machine Learning*, 20: 273–297, 1995.

James E. Crouch. ***Essential Human Anatomy: A Text-Atlas***. Lea & Febiger, Philadelphia, 1981. ISBN 0-8121-0755-1.

Ivan Nunes da Silva, Danilo Hernane Spatti, Rogerio Andrade Flauzino, Luisa Helena Bartocci Liboni, and Silas Franco dos Reis Alves. ***Artificial Neural Networks: A Practical Course***. Springer International Publishing, AG, Switzerland, 2017. ISBN 978-3-319-43161-1. doi: 10.1007/978-3-319-43162-8.

Peng Dai and Ing Yann Soon. **An improved model of masking effects for robust speech recognition system**. *Speech Communication*, 55:387–396, 2013.

Hallowell Davis. **An active process in cochlear mechanics**. *Hearing Research*, 9:79–90, 1983.

Steven B. Davis and Paul Mermelstein. **Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences**. *IEEE Transactions On Acoustics, Speech, and Signal Processing*, ASSP-28(4):357–366, August 1980.

G. Deekshitha, J.J. Thennatde, and M. Leena. **Segmentation of Continuous Speech for Broad Phonetic Engine**. *IEEE International Conference on Electrical, Computer and Communication Technologies*, 2015.

Huiqun Deng and Douglas O'Shaughnessy. **Voiced-Unvoiced-Silence Speech Sound Classification Based On Unsupervised Learning**. *Multimedia and Expo, 2007 IEEE International Conference on*, 2007. doi: 10.1109/ICME.2007.4284615.

Wei Dong and Elizabeth S. Olson. **Detection of Cochlear Amplification and Its Activation**. *Biophysical Journal*, 105:1067–1078, 2013.

James P. Egan and Harold W. Hake. **On the Masking Pattern of a Simple Auditory Stimulus**. *2013 The Journal of the Acoustical Society of America*, 22(5):622–630, 1950.

Benjamin Elizalde and Gerald Friedland. **Lost in Segmentation: Three Approaches for Speech/Non Speech Detection in Consumer-Produced Videos**. *2013 IEEE International Conference on Multimedia and Expo*, pages 9–42, 2013.

ETSI-ES-202-050. **Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithms; Compression algorithms**. Standard ETSI ES 202 050 V1.1.5 (2007-01), European Telecommunications Standards Institute, 2007.

Hugo Fastl and Eberhard Zwicker. ***Psycho-Acoustics Facts and Models; Third Edition***. Springer Series in Information Sciences, 2007. ISBN 978-3-642-51765-5.

William Fisher, George Doddington, Kathleen Goudie-Marshall, Charles Jankowski, Ashok Kalyanswamy, Sara Basson, and Judith Spitz. **NTIMIT Speech Corpus**. speech corpus, Philadelphia: Linguistic Data Consortium, 1993.

Harvey Fletcher. **Auditory Patterns**. *Reviews of Modern Physics*, 12:47–66, 1940.

Harvey Fletcher and W. A. Munson. **Loudness, Its Definition, Measurement and Calculation**. *The Bell System Technical Journal*, October 1933.

Harvey Fletcher and W. A. Munson. **Relation Between Loudness and Masking**. *The Journal of the Acoustical Society of America*, 9(1), 1937. doi: 10.1121/1.1915904.

Anders Fridberger, Igor Tomo, Mats Ulfendahl, and Jacques Boutet de Monvel. **Imaging hair cell transduction at the speed of sound: Dynamic behavior of mammalian stereocilia**. *PNAS*, 103(6):1918–1923, 2006.

ITU-T G.729. **Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP) - Annex B**. Standard ITU-T G.729, International Telecommunications Union, 2012. URL `http://www.itu.int/rec/T-REC-G.729/e`. Last accessed on July 6th 2017.

John Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, Nancy L. Dahlgren, and Victor Zue. **TIMIT**. **Acoustic-Phonetic Continuous Speech Corpus LDC93S1**, Linguistic Data Consortium, Philadelphia, 1993. URL `https://catalog.ldc.upenn.edu/ldc93s1`. Last accessed on July 6th 2017.

Zoubin Ghahramani. **An Introduction to Hidden Markov Models and Bayesian Networks**. *Section 8.2.3. International Journal of Pattern Recognition and Artificial Intelligence*, 15(1):9–42, 2001.

Thea Ghiselli-Crippa and Amro El-Jaroudi. **Voiced-unvoiced-silence classification of speech using neural nets**. *IJCNN-91-Seattle International Joint Conference on Neural Networks*, 2:851–856, 1991.

B. Gold, N. Morgan, and D. Ellis. ***"Speech and Audio Signal Processing - Processing and Perception of Speech and Music. Second Edition***. John Wiley & Sons Inc., Hoboken, New Jersey, 2011. ISBN 978-0-470-19536-9.

F. Goldman-Eisler. **The Distribution Of Pause Durations in Speech**. *Journal of Language and Speech*, 4(4):232–237, 1961.

John O. Green. ***Cognitive Processes: Methods for Probing the Black Box—published in A Handbook for the Study of Human Communication: Methods and Instruments for Observing, Measuring, and Assessing Communication Processes (Edited by Charles H. Tardy)***. Ablex Publishing Corporation, New Jersey, 1988. ISBN 0-89391-424-X.

Donald D. Greenwood. **The Mel Scale's disqualifying bias and a consistency of pitch-difference equisections in 1956 with equal cochlear distances and equal frequency ratios**. *Hearing Research*, 103:199–224, 1997.

Jan Gullberg. ***Mathematics from the Birth of Numbers***. W.W. Norton & Company Inc., 10 Coptic Street, London WC1A 1PU, 1997. ISBN 0-393-04002-X.

Mary Hardy. **The length of the Organ of Corti in Man**. *American Journal of Anatomy*, 1938. URL `https://doi.org/10.1002/aja.1000620204`.

Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury. **Deep Neural Networks for Acoustic Modeling in Speech Recognition: The shared views of four research groups**. *IEEE Signal Processing Magazine NOVEMBER 2012*, pages 82–97, 2012. doi: 10.1109/MSP.2012.2205597.

Tin Kam Ho. **Random Decision Forests**. *: Proceedings of 3rd International Conference on Document Analysis and Recognition*, 1:278–282, 1995.

Xuedong Huang, Alex Acero, and Hsiao-Wuen Hon. ***Spoken Language Processing - A guide to Theory, Algorithm, and System Development***. Prentice-Hall, 2001. ISBN 0-13-0226616-5.

George W Hughes and Morris Halle. **Spectral Properties of Fricative Consonants**. *The Journal of the Acoustical Society of America*, 28(2):303–310, 1956.

Aini Hussain, Salina Abdul Samud, and Liew Ban Fah. **Endpoint Detection of Speech Signal using Neural Network**. *2000 TENCON Proceedings*, 1:271–274, 2000.

J. M. Ivison. ***Electric Circuit Theory***. Van Nostrand Reinhold Company Ltd., 1978. ISBN 0-442-30201-9.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. ***An Introduction to Statistical Learning (with Applications in R)***. Springer, 2013. ISBN 978-1-4614-7137-7.

Lloyd A. Jeffress. ***Foundations of Modern Auditory Theory: Volume 1; Editor:J. V. Tobias***. Academic Press, New York, 1970.

K. Johnson. ***Acoustic & Auditory Phonetics, 2nd Edition***. Blackwell Publishing Ltd, 2012. ISBN 978-1-4051-0123-3.

Daniel Jurafsky and James H. Martin. ***Speech and Language Processing - An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition Second Edition***. Pearson Education Inc., New Jersey, 2009. ISBN 13 978-0-13-504196-3.

Robert W. Koch, Hanif M. Ladak, Mai Elfarnawany, and Sumit K. Agrawal. **Measuring Cochlear Duct Length - a historical analysis of methods and results**. *Journal of Otolaryngology - Head and Neck Surgery*, pages 1–11, 2017. doi: DOI10.1186/s40463-017-0194-2.

Michael A. Krasner. **The Critical Band Coder—Digital Encoding of Speech Signals Based On the Perceptual Requirements of the Auditory System**. *ICASSP '80*, 5: 327–331, 1980.

T. Sunil Kumar, Md. Azahar Hussain, and Vivek Kanhangad. **Classification of Voiced and Non-voiced Speech Signals using Empirical Wavelet Transform and Multi-level Local Patterns**. *IEEE International Conference on Digital Signal Processing*, pages 390–394, 2015. doi: 10,1109/ICDSP.2015.7251851.

Dave R M. Langers. **Assessment of Tonotopically Organised Subdivisions in Human Auditory Cortex Using Volumetric and Surface-Based Cortical Alignments**. *Human Brain Mapping*, pages 1544–1561, 2014.

Dave R. M. Langers, Katrin Krumbholz, Richard W. Bowtel, and Deborah A. Hall. **Neuroimaging paradigms for tonotopic mapping (I): The influence of sound stimulus type**. *NeuroImage*, 100:650–662, 2014.

Wayne A. Lea. ***Chapter 2: The Elements of Speech Recognition in "Electronic Speech Recognition; Techniques, Technologies and Applications"; Edited by G.Bristow.*** William Collins Sons & Co. Ltd., 1986. ISBN 0-00-383238-4.

Y. LeCun and Y. Bengio. ***Convolutional networks for images,speech, and time-series. (From The Handbook of Brain Theory and Neural Networks; Editor M. A. Arbib)***. MIT Press, 1995.

Sanjeev Manchanda, Mayank Dave, and S. B. Singh. **An Empirical Comparison Of Supervised Learning Processes**. *International Journal of Engineering*, 1(1):21–38, 2007.

Warren S McCulloch and Walter Pitts. **A Logical Calculus of the Ideas Immanent in Nervous Activity**. *Bulletin of Mathematical Biology, Vol. 52, No. 1/2, pp. 99-115, 1990 (Reprinted from the Bulletin of Mathematical Biophysics, Vol. 5, pp. 115–133, 1943.*, 52(1/2): 99–115, 1990.

Harry McGurk and John MacDonald. **Hearing Lips and Seeing Voices**. *Nature*, 264: 746–748, December 1976.

Harry McGurk and John MacDonald. **Visual Influences on Speech Perception Processes**. *Perception & Psychophysics*, 24(3):253–257, 1978.

James D. Miller. **Sex differences in the length of the organ of corti in humans**. *Acoustical Society of America*, 2007. doi: DOI:10.1121/1.2710746.

Leonard Mlodinow. ***The Drunkards Walk: How Randomness Rules Our Lives***. Penguin Books, London WC2R ORL, 2009. ISBN 13 978-0-141-02647-3.

Michelle Moerel, Federico De Martino, and Elia Formisano. **Processing of Natural Sounds in Human Auditory Cortex: Tonotopy, Spectral Tuning, and Relation to Voice Sensitivity**. *The Journal of Neuroscience*, 32(41):14205–14216, Oct 2012.

Md. Khademul Islam Molla, Keikichi Hirose, and Md. Kamrul Hasan. **Voiced/non-voiced speech classification using adaptive thresholding with bivariate EMD**. *Cell Tissue Res Pattern Analysis Applic.*, 19:390–394, 2015. doi: 10.1007/s10044-015-0449-3.

Sujoy Mondal and Abhirup Das Barman. **Clustering based Voiced-Unvoiced-Silence Detection in Speech using Temporal and Spectral Parameters**. *IEEE International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)*, pages 390–394, 2015.

Jugurta Montalvão and Marcos Renato Rodrigues Araujo. **Is masking a relevant aspect lacking in MFCC? A speaker verification perspective**. *Pattern Recognition Letters*, 33:2156–2165, 2012.

Brian C. J. Moore and Brian R. Glasberg. **Suggested formulae for Calculating Auditory-Filter Bandwidths and Excitation Patterns**. *The Journal of the Acoustical Society of America*, 74(3):750–753, 1983.

MP3-Standard. **Information technology—Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/S—Part 3: Audio**. Standard BS EN ISO/IEC11172-3:1995, British Standards Institution, London, W4 4AL, 1995.

Steffen Nissen. **Fast Artificial Neural Network Library (FANN) ©Steffen Nissen**, 2003. URL `http://tenet.dl.sourceforge.net/project/fann/fann_doc/1.0/fann_doc_complete_1.0.pdf`. Last accessed 2016-06-20.

GNAT Programming Studio©. **GPS GPL Edition of the GNAT Programming Studio: GPS 6.0.1 (20140113) hosted on i686-pc-mingw32; GNAT GPL 2014 (20140331) ©AdaCore**, 2014. URL `http://libre.adacore.com/tools/gps/`. Last accessed 2017-03-07.

Mihaela Oprea and Daniela Şchiopu. **An Artificial Neural Network-Based Isolated Word Speech Recognition System for the Romanian Language**. *IEEE 16th International Conference on System Theory, Control and Computing*, pages 1–6, April 2012.

Douglas O'Shaughnessy. ***Speech Communications: Human and Machine, 2nd Edition***. SPIE—John Wiley and Sons; Inc., 1999. ISBN 978-0-7803-3449-6.

Oxford Dictionary of Computing. ***Oxford Dictionary of Computing; Editors: John Daintith and Edmund Wright***. Oxford University Press, Oxford OX2 6DP, 2008. ISBN 0-8194-4987-9.

Dimitri Palaz, Mathew Magimai Doss, and Ronan Collobert. **Learning linearly separable features for speech recognition using convolutional neural networks**. *arXiv:1412.7110 [cs.LG]*, 2015.

Ki-Young Park and Soo-Young Lee. **An engineering model of the masking for the noise-robust speech recognition**. *Neurocomputing*, 52(54):615–620, 2003.

Gordon E. Petersen and Harold L. Barney. **Control Methods Used in a Study of the Vowels**. *The Journal of the Acoustical Society of America*, 24(2):175–184, 1952.

Kevin L Priddy and Paul E Keller. ***Artificial Neural Networks: An Introduction***. SPIE—The International Society for Optical Engineering, Washington, 2005. ISBN 978-0-19-923400-4.

Fengyan Qi, Changehun Bao, and Yan Liu. **A Novel Two-Step SVM Classifier For Voiced/Unvoiced/Silence Classification of Speech**. *IEEE ISCSLP*, pages 77–80, 2004.

Yingyong Qi and Bobby R Hunt. **Voiced-Unvoiced-Silence Classifications of Speech Using Hybrid Features and a Network Classifier**. *IEEE Transactions On Speech And Audio Processing*, 1(2):250–255, April 1993.

L.R. Rabiner and M.R. Sambur. **An Algorithm for Determining the Endpoints of Isolated Utterances**. *The Bell System Technical Journal*, pages 297–315, 1975.

Helge Rask-Andersen, Wei Lui, Elsa Erixon, Anders Kinnefors, Kristian Pfaller, Annelies Schrott-Fischere, and Rudolf Glueckert. **Human Cochlea: Anatomical Characteristics and Their Relevance for Cochlear Implantation**. *The Anatomical Record*, 295: 1791–1811, 2012.

M. Reidmiller and H. Braun. **A direct adaptive method for faster backpropagation learning: the RPROP algorithm**. *IEEE International Conference on Neural Networks*, pages 586–591, 1993. doi: 10.1109/ICNN.1993.298623.

Patrick F. Reidy. **A Comparison of Spectral Estimation Methods for the Analysis of Sibilant Fricatives**. *The Journal of the Acoustical Society of America*, March 2015. URL `http://dx.doi.org/10.1121/1.4915064`. Last accessed 09-Feb-2016.

F. Rosenblatt. **The Perceptron: A probabilistic Model For Information Storage And Organisation In the Brain**. *Psychological Review*, 65(6):386–408, 1958.

Marek Rudnicki, Oliver Schoppe, Michael Isik, Florian Völlk, and Werner Hemmert. **Modeling auditory coding: from sound to spikes**. *Cell Tissue Res*, 361:159–175, 2015. doi: 10.1007/s00441-015-2202-z.

D.E. Rumelhart, G.E. Hinton, and R.J. Williams. ***Learning internal representations by error Propagation in Parallel Distributed Processing (Volume 1, Chapter 8)***. MIT Press, Cambridge, Massachusetts, USA, 1986.

Tushar Ranjan Sahoo and Sabyasachi Patra. **Silence Removal and Endpoint Detection of Speech Signal for Text Independent Speaker Identification**. *I.J. Image, Graphics and Signal Processing*, 6:27–35, 2014.

Tara N. Sainath, Ron J. Weiss, Andrew Senior, Kevin W. Wilson, and Oriol Vinyals. **Learning the Speech Front-end With Raw Waveform Convolutional, Long Short-Term Memory, Fully Connected Deep Neural Networks (CLDNNs)**. *Google, Inc. New York, U.S.A*, 2015.

V.V.S. Sarma and D. Venugopal. **Studies on pattern recognition approach to voiced-unvoiced-silence classification**. *ICASSP '78*, 3:1–4, 1978.

B. Scharf. ***Foundations of Modern Auditory Theory: Volume 1; Editor:J. V. Tobias***. Academic Press, New York, 1970.

Eckart Scheerer. **The unknown Fechner**. *Psychological Research*, 49:197–201, 1987.

M. R. Schroeder, B. S. Atal, and J. L. Hall. **Optimizing digital speech coders by exploiting masking properties of the human ear**. *J. Acoustical Society of America*, 66(6): 1647–1652, 1979. doi: 10.1121/1.383662.

Claude E. Shannon. **Communication in the Presence of Noise**. *Proceedings of the IRE, Vol 37, Number 1 pp. 10–21, 1949*, 37(1):10–21, 1949.

Paris Smaragdis. ***Techniques for Noise Reduction in Automatic Speech Recognition; Edited by Tuomas Virtanen, Rita Singh and Bhiksha Raj.*** John Wiley & Sons Ltd, 2013. ISBN 978-0-470-97409-4.

Jongseo Sohn, Soo Kim, and Wonyong Sung. **A Statistical Model-Based Voice Activity Detection**. *IEEE Signal Processing Letters*, 6(1):1–3, January 1999.

Richard M. Stern and Nelson Morgan. ***Features Based on Auditory Physiology and Perception. Published in: Techniques for Noise Reduction in Automatic Speech Recognition; Edited by Virtanen, Singh, and Raj.*** John Wiley & Sons Ltd, 2013. ISBN 978-0-470-97409-4.

S. S. Stevens and J. Volkmann. **The Relation of Pitch To Frequency: A Revised Scale**. *The American Journal of Psychology*, 53(2):329–353, 1940.

S. S. Stevens, J. Volkmann, and E. B. Newman. **A Scale for the Measurement of the Psychological Magnitude Pitch**. *J. Acoustical Society of America*, 8:185–190, Jan 1937.

S.S. Stevens. **On the Psychophysical Law**. *The Psychological Review*, 64(3):153–181, 1957.

Tascam™. **Tascam™ DP004 Digital Pocket Studio**, 2017. URL `http://tascam.com./product/dp-004/`. Last accessed 2017-02-09; Tascam™ is a trademark of TEAC Corporation, registered in the U.S. and other countries.

Peng Teng and Yunde Jia. **Voice Activity Detection Using Non-Negative Sparse Coding**. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 289–292, 2012.

William Forde Thompson, Varghese Peter, Kirk N. Olsen, and Catherine J. Stevens. **The effect of intensity on relative pitch**. *The Quarterly Journal of Experimental Psychology*, 65 (10):2054–2072, 2012. doi: 10.1080/17470218.2012.678369.

Doroteo Torre Toledano. **Neural Network Boundary Refining For Automatic Speech Segmentation**. *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 6:3438–3441, 2000.

Hartmut Traunmüller. **Analytical Expressions for the Tonotopic Sensory Scale**. *The Journal of the Acoustical Society of America*, 88(1):87—100, July 1990.

Abhay Upadhyay and Ram Bilas Pachori. **Instantaneous voiced/non-voiced detection in speech signals based on variational mode decomposition**. *The Journal of the Franklin Institute*, 352(7):2679—2707, Jul 2015.

R. Vergin and D. O'Shaughnessy. **Pre-Emphasis and Speech Recognition**. *Canadian Conference on Electrical and Computer Engineering*, 2:1062–1065, 1995.

Tuomas Virtanen, Rita Singh, and Bhiksha Raj. ***Techniques for Noise Reduction in Automatic Speech Recognition***. John Wiley & Sons Ltd, 2013. ISBN 978-0-470-97409-4.

Andrew J. Viterbi. **A Personal History of the Viterbi Algorithm**. *IEEE Signal Processing Magazine*, pages 120–142, 2006.

Georg von Békésy. ***Enlarged Mechanical Model of the Cochlea with Nerve Supply; Figure 19; Pg 325 in Foundations of Modern Auditory Theory: Volume 1; Editor:J. V. Tobias***. Academic Press, New York, 1970.

Wei Wei and Chen Yanpu. **Speech Enhancement by Spectral Component Selection**. *IEEE Proceedings of ICSP2000*, pages 674–678, 2005.

Maria-Barbara Wesenick and Andreas Kipp. **Estimating The Quality Of Phonetic Transcriptions and Segmentations of Speech Signals**. *ICSLP 96. Proceedings: Fourth International Conference on Spoken Language*, pages 129–132, 1996.

George M. White and Richard B. Neely. **Speech Recognition Experiments with Linear Prediction, Bandpass Filtering, and Dynamic Programming**. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-24(2):617–620, 1976.

C. Ray Wylie and Louis C. Barrett. ***Advanced Engineering Mathematics***. McGraw-Hill International Book Company, 1982. ISBN 0-07-066643-1.

W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig. **Achieving Human Parity in Conversational Speech Recognition**. *Microsoft Research; Technical Report MSR-TR-2016-71*, 2016.

Datao You, Jiqing Han, Guibin Zheng, and Tieran Zheng. **Sparse Power Spectrum Based Robust Voice Activity Detector**. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 289—292, 2012.

Weizhong Zhu and Douglas 0'Shaughnessy. **Incorporating Frequency Masking Filtering in a Standard MFCC Feature Extraction Algorithm**. *IEEE Proceedings 7th International Conference on Signal Processing*, 1:617–620, 2004.

E. Zwicker. **Subdivision of the Audible Frequency Range into Critical Bands (Frequenzgruppen)**. *The Journal of the Acoustical Society of America*, 3(2):pg. 248, 1961.

E. Zwicker, G. Flottorp, and S. Stevens. **Critical Band Width in Loudness Summation**. *The Journal of the Acoustical Society of America*, 29(5):548–557, 1957.