Master report, IDE 1271, November 2012
Subject: Master Thesis in Embedded and Intelligent Systems

Speech Intelligibility Measurement on the basis of ITU-T Recommendation P.863

Master thesis

School of Information Science, Computer and Electrical Engineering

Swatantra Ghimire

Speech Intelligibility Measurement on the basis of ITU-T Recommendation P.863

Master Thesis in Embedded and Intelligent Systems

November 2012

Author:

Swatantra Ghimire

Supervisors:

John G. Beerends (TNO, the Netherlands)
Josef Bigun (Halmstad University)

Examiner:

Antanas Verikas

School of Information Science, Computer and Electrical Engineering

Halmstad University

PO Box 823, SE-301 18 HALMSTAD, Sweden

Master Thesis
Report, IDE1271
School of Information Science, Computer and Electrical Engineering
Halmstad University

Preface

First, I would like to thank my supervisor John G. Beerends at TNO for all the support he provided during the course of this thesis work. It was not an easy choice to come all the way from Sweden to do my thesis work at TNO, Netherlands. All the documentations required, travel arrangement and visa issues were quite challenging, but he helped me a lot throughout all these and I sincerely thank him for that. He has been a huge source of inspiration and motivation. Without his guidance, it would not have been possible to finish my thesis on time. I have learned a lot from him, and that has made me see myself as someone working in the fascinating field of acoustics in the future.

I would also like to thank my supervisor Josef Bigun at Halmstad University for his supervision. His comments during the mid-term presentation, motivating words in the beginning of my thesis, support and understanding throughout this work are highly appreciated, I thank him a lot.

Indeed, I have to thank TNO for giving me the opportunity to do my thesis work in their company. All of my colleagues whom I used as subjects were vital part of this work.

This work would not have been possible without the support and love I received from my family and friends. I especially thank my mother for always being there for me.

Finally, I would like to thank the almighty God for all the courage he provided, which kept me going during tough times.

Abstract

Objective speech intelligibility measurement techniques like AI (Articulation Index) and AI based STI (Speech Transmission Index) fail to assess speech intelligibility in modern telecommunication networks that use several non-linear processing for enhancing speech. Moreover, these techniques do not allow prediction of single individual CVC (Consonant Vowel Consonant) word intelligibility scores. ITU-T P.863 standard [1], which was developed for assessing speech quality, is used as a starting point to develop a simple new model for predicting subjective speech intelligibility of individual CVC words. Subjective intelligibility measurements were carried out for a large set of speech degradations. The subjective test uses single CVC word presentations in an eight alternative closed response set experiment. Subjects assess individual degraded CVC words and an average of correct recognition is used as the intelligibility score for a particular CVC word. The first subjective database uses CVC words that have variations in the first consonant i.e. /C/ous (represented as "**kæʊs**" using International Phonetic Association phonetic alphabets). This database is used for developing the objective model, while a new database based on VC words (Vowel Consonant) that uses variations in the second consonant (a/C/ e.g. aH, aL) is used for validating the model.

ITU-T P.863 shows very poor results with a correlation of 0.30 for the first subjective database. A first extension to make P.863 suited for intelligibility prediction is done by restructuring speech material to meet the temporal structure requirements (speech+silence+speech) set for standard P.863 measurements. The restructuring is done by concatenating every original and degraded CVC word with itself. There is no significant improvement in correlation (0.34) when using P.863 on the restructured first subjective database (speech material meets temporal requirements). In this thesis a simple model based on P.863 is developed for assessing intelligibility of individual CVC words. The model uses a linear combination of a simple time clipping indicator (missing speech parts) and a “Good frame count” indicator which is based on the local perceptual (frame by frame) signal to noise ratio. Using this model on the restructured first database, a reasonably good correlation of 0.81 is seen between subjective scores and the model output values. For the validation database, a correlation of around 0.76 is obtained. Further validation on an existing database at TNO, which uses time clipping degradation only, shows an excellent correlation of 0.98.

Although a reasonably good correlation is seen on the first database and the validation database, it is too low for reliable measurements. Further validation and development is required, nevertheless the results show that a perception-based technique that uses internal representations of signals can be used for predicting subjective intelligibility scores of individual CVC words.

Table of contents

List of Tables

List of Figures

1. Introduction

This chapter introduces the thesis. It describes the motivation behind the research work carried out in this thesis. The solution approach and the outline of the thesis are given.

With the introduction of advanced voice processing techniques in voice communication systems such as noise suppression, echo cancellation, automatic level and gain control etc., in combination with the use of mobile handsets in hostile acoustic environments, the voice quality in many situations is becoming unacceptably low and speech intelligibility is becoming a major issue. Interaction between different voice processing units can introduce several types of severe degradations leading to unacceptable low quality. Tools for testing voice quality/intelligibility are therefore of the utmost importance.

Speech quality and intelligibility, although closely tied, are different attributes of a speech signal and they are not equivalent (refer Chapter 2.1 Speech quality and Intelligibility). Optimizing a system based on speech quality may not improve the intelligibility and vice versa. Thus, different assessment methods are required to assess speech quality and intelligibility of a system.

Several subjective and objective measurement methods exist for assessing speech quality and intelligibility. Subjective measurements, which use a panel of listeners, are accurate and reliable, but very expensive and time-consuming. An objective measurement method uses a computerized model, which measures the quality/intelligibility of voice services at a very cheap cost. Objective tools for measuring speech quality are more reliable compared to those for measuring speech intelligibility (Series P ITUT-T standards for measuring speech quality have been extensively used for optimizing speech quality) and they perform well in terms of correlation with subjective measurements (high correlation with several subjective experiments).

ITU-T recommendations P.863 (POLQA, Perceptual Objective Listening Quality Assessment) [1]-[8], P.862 (PESQ, Perceptual Evaluation of Speech Quality) [9], [10] were designed to perform objective assessment of listening quality. The most comprehensive and extensively used objective speech quality assessment method among these is ITU-T P.863, POLQA. This method is well suited for assessing the speech quality in HD voice, VoIP (Voice over Internet Protocol), 3G and NGN (Next Generation Networks). It is important to remember that POLQA does not provide comprehensive evaluation of transmission quality, as it only measures effect of one way distortion and noise on speech quality and it is possible to have high POLQA scores, yet poor overall conversational quality [4].

AI (Articulation Index) [11] and AI based STI (Speech Transmission Index) method [12]-[14] are the most widely used standardized objective methods to predict speech intelligibility in most of the electro-acoustic situations. The STI method can be used to measure and calculate the STI value, which quantifies how well the speech is transmitted through a channel with respect to intelligibility. This value lies within the range of zero (completely unintelligible) and one (perfect intelligibility). This method uses a test signal, which is a noise signal containing frequency characteristics similar to those found in natural speech. The STI method assumes that prediction of intelligibility is based on the weighted contribution from a number of frequency bands. The contribution is based on effective SNR (determined by several factors like noise, reverberation etc.) [12]. Modern digital speech transmission channels, which use advanced coding (minimize bit rate, but maintain quality) and speech enhancement algorithms, will behave differently for speech and modulated-noise STI and the use of STI test signal will possibly result in an incorrect assessment of speech intelligibility. This can be seen in [14], where STI provides an erroneous prediction of intelligibility for speech processed using noise reduction method (spectral subtraction). The extensions of STI method that operate on natural speech signals (use speech as probe) exist, but they cannot cope with the distortions seen in modern telecommunication networks.

Contrary to the standard intelligibility measurement approaches, this thesis focuses on the intelligibility of individual CVC words (Consonant Vowel Consonant). A simple new objective model based on ITU-T P.863, POLQA is developed for assessing speech intelligibility of the individual CVC words. Single CVC identification analysis will allow exact determination of speech intelligibility problems of a transmission channel.

1.1 Problem formulation

With increased usage of mobile devices and network services in hostile acoustic environment i.e. presence of high level of background noises (noise in moving car, train, airplane etc.), speech intelligibility is severely affected. This issue cannot be solved by allocating more bandwidth to the channel/source coding (although users may be willing to pay for the larger bandwidth), but by pre-processing the input speech signal (e.g. noise suppression). Thus, assessing speech intelligibility is very important. In addition, with the emergence of low-rate coding techniques, noise suppression and speech enhancement algorithms, the use of objective speech intelligibility measurement methods like the STI method will provide incorrect assessment of speech intelligibility.

The STI method has shown a robust relation with subjective tests based on CVC words [12]. However, subjective intelligibility score in the STI tests does not represent an individual CVC score, but an average score over several CVC words (CVC words are embedded in a carrier phrase).

Moreover, the STI method has requirements on the length of a test signal (at least 10-20 seconds for reliable measurement). It cannot be used for predicting the intelligibility of a single CVC word because of the short duration of a CVC word (short envelope spectrum). This means that a fundamental perception based approach that uses the psychoacoustic (internal) representation of speech is necessary for measuring intelligibility of the individual CVC words (measuring single CVC intelligibility will allow exact determination of speech intelligibility problems of a transmission channel). This would allow significant improvement in the perceptual optimization of telecommunication systems.

1.2 Solution approach

POLQA is standardized by ITU-T as the recommendation P.863 for the objective assessment of overall listening speech quality in both narrowband (300 to 3400 Hz) and super-wideband (50 to 14000 Hz) mode in telecommunication networks [1]. Subjective speech quality tests for POLQA use an absolute category rating (ACR) 5-point opinion scale (see

Table 1) to obtain the user's perception of speech quality. This scale uses a single number ranging from one (lowest perceived speech quality) to five (the best-perceived speech quality). Results from the listening tests are expressed in terms of MOS-LQS (Mean Opinion Score Listening Quality Subjective), which is an average quality score over a large set of votes by subjects using the ACR 5- point scale. POLQA predicts perceived listening quality in terms of MOS-LQO (Mean Opinion Score Listening Quality Objective) which has a very high correlation with MOS-LQS [1].

Quality of the speech	Score
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

Table 1. ACR 5-point opinion scale used in POLQA. In ACR listening tests, subjects only listen to degraded speech signals and express their opinion using the 5-point opinion scale.

Intelligibility refers to the degree with which humans can understand a speech signal. It cannot be measured using an opinion score (see Table 1), as a speech fragment is either understandable or not. A subjective listening test that uses nonsense CVC (Consonant Vowel Consonant) words is designed to measure intelligibility of the individual CVC words. The CVC words are degraded by a set of common degradations seen in today's telecommunication networks. A well-constructed subjective listening test is crucial in developing an objective model. The listening test is designed to ensure that no extensive training is required for a listener to take the test (refer 3.2 Setup and procedure for further details).

The subjective database, which contains original, degraded CVC words and the intelligibility score related to individual degraded CVC words, is obtained after conducting several listening experiments. ITU-T P.863, POLQA is used 'as is' in the super-wide band mode to predict intelligibility on the subjective database. The results obtained from the measurements are used to develop a simple model based on ITU-T P.863 for assessing the intelligibility of individual CVC words. The model is then validated on a new subjective database based on nonsense VC words (Vowel Consonant). A diagrammatic representation of the solution approach is given in Figure 1.

1.3 Boundary conditions

Intelligibility has three domains that contribute to the correct identification of CVC words:

- How well the CVC word is pronounced (Speaker)
- How well is the CVC word transported through transmission channel
- How well the listener is trained in the identification task (Listener)

The simple objective model developed in this thesis for assessing intelligibility of individual CVC words focuses only on the second aspect and quantifies how well a transmission channel transports speech in terms of intelligibility (measures perceived intelligibility). So, this model cannot be used to predict impact on intelligibility because of other two aspects. Moreover, the model cannot be used to replace subjective testing.

1.4 Contribution

The contributions of the work done in this thesis are listed below:

- The simple objective model developed in this thesis can be used to determine speech intelligibility problems of a voice link, enabling service providers to optimize their services to provide better intelligibility.
- The subjective databases developed in this thesis can be used for developing and validating new objective models. The speech signal along with the model developed in this thesis can be used as a tool for measuring intelligibility of a transmission channel in real time (speech signals at one end of transmission channel and the objective model at the other end to measure intelligibility of the channel)
- The simple model is based on ITU-T P.863, POLQA. This shows that a single model that can assess speech quality as well as intelligibility could be developed in future, allowing telecommunication operators to use a single tool for optimizing both the quality and intelligibility.
- Measuring and optimizing voice services in terms of intelligibility is important, e.g. a telephone call used as an evidence loses its value if unintelligible.

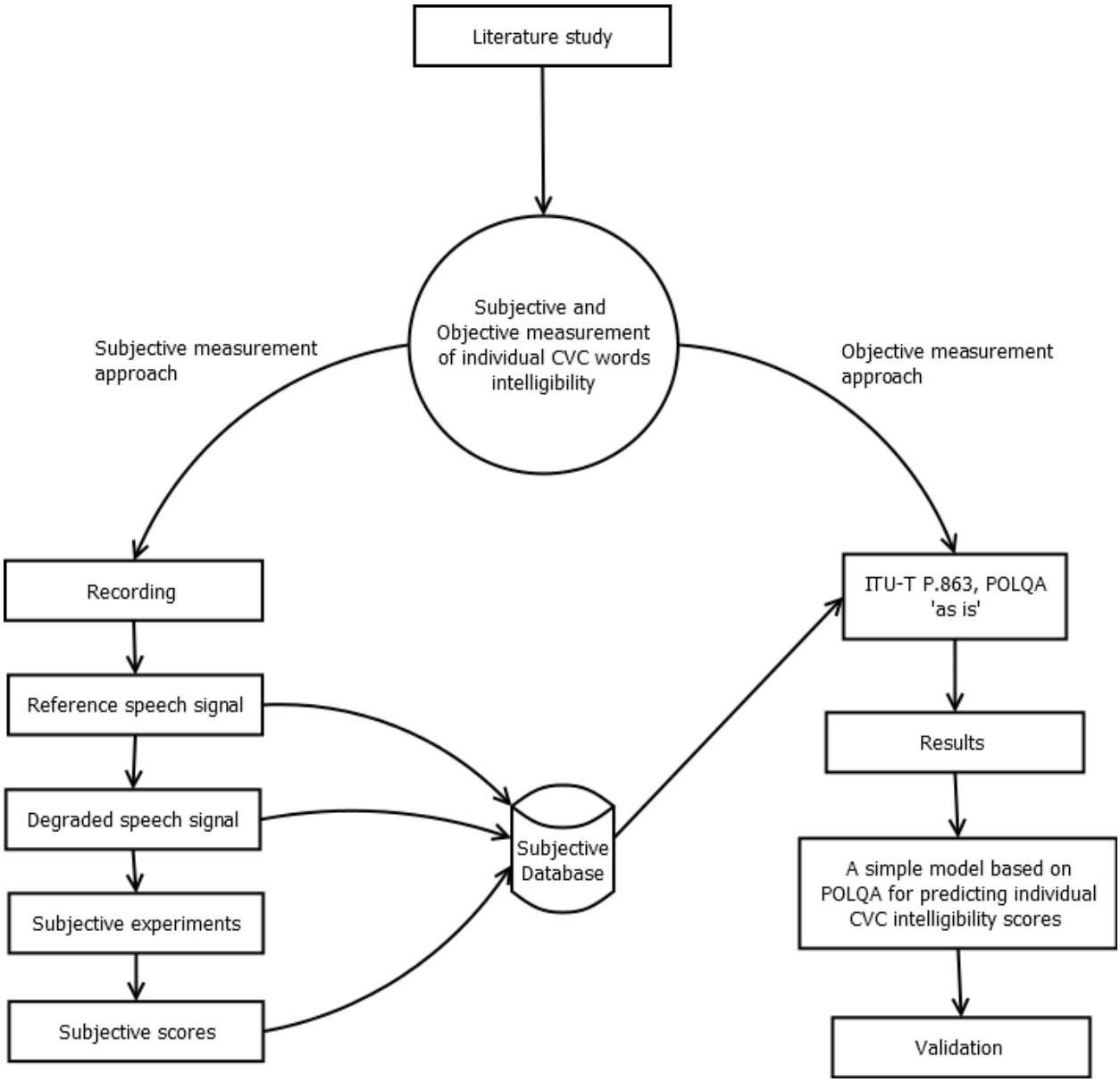

Figure 1. A flowchart showing the solution approach used in this thesis.

1.5 Outline

The structure of the thesis work is given below:

Chapter 2.

This chapter introduces speech quality and intelligibility, presents differences and similarities between them and their impact on the voice services in today's telecommunication networks. The psychoacoustic model used in ITU-T P.863 is described. A short description of AI and AI based STI method is also presented.

Chapter 3.

This chapter describes the subjective speech intelligibility measurement approach used in this thesis. Setup and procedure used for conducting listening tests is elaborated.

Chapter 4.

This chapter describes the objective speech intelligibility measurement approach that was developed in this thesis. The results and validation of the simple model developed in this thesis are discussed.

Chapter 5.

This chapter presents the conclusions derived from the research done in this thesis. Some suggestions for further improvement are given.

2. Background

Quality and intelligibility are highly correlated, but different from each other. In this thesis work, a simple perception-based objective model is developed for measuring speech intelligibility of individual CVC words. ITU-T P.863, POLQA is used as a starting point, which is an objective speech quality assessment technique. This chapter describes the POLQA psychoacoustic model in detail. A short description of AI and AI based STI technique is also presented.

2.1 Speech quality and Intelligibility

Quality is a measure of the difference between idealization and realization, while intelligibility refers to the degree with which a normal person can understand the speech.

Quality is a highly subjective measure and the same speech material can have a varying quality measure among different listeners. Speech quality and intelligibility have a strong correlation, but they are different. The best example to illustrate this is the missing word phenomenon where a complete word is omitted from the voice transmission, which in some cases is not noticed by subjects leading to a non-intelligible high-quality speech fragment.

As new voice services are rolled out by the telecommunication industry, the quality of speech is very important, but these systems need to ensure that they get the message across i.e., transmitting intelligible information to the intended recipient is crucial [15]. The noise suppression and the speech enhancement algorithms used in these systems focus on optimizing speech quality at the expense of speech intelligibility, especially in low SNR conditions. However, when SNR is low or when packet loss is high, the focus should shift to intelligibility, as conveying what is being said is very important. In addition, one should not confuse audibility with intelligibility, i.e., a speech might be audible (loud enough), but still maybe unintelligible (not understandable).

The first measurement results from the initial work under the title P.OSI (series P recommendations Objective Speech Intelligibility) started by ITU-T show that the traditional narrowband PSTN (Public Switched Telephone Network) has better intelligibility than a wideband Skype service, but Skype has better quality than PSTN [2]. This means that we are enjoying high-quality voice services at the expense of reduced intelligibility.

There are several primary and secondary factors, which influence intelligibility. Some of the primary factors are listed below [15]:

- Bandwidth and Frequency Response (DUT)
- Loudness and Signal-to-noise ratio
- Reverberation and Echo
- Loss of information (time clipping)
- Time varying lag
- Various distortions etc.

Apart from these, there are several secondary factors; some of them are listed below:

- Gender
- Age
- Context and content of the speech
- Microphone and speaker quality
- Talker to listener distance
- Ability of the listener to discern etc.

2.2 Objective speech quality measurement

Several objective speech quality assessment techniques exist, which include techniques based on SNR, spectral distance, linear predictive coding parameters to name a few. These techniques quantify speech quality in terms of some system parameters (SNR, frequency response, harmonic distortion, log-area ratio). Although these techniques are simple to model and easy to evaluate, they do not take human auditory processing into the account (non-linear frequency analysis, masking, loudness etc.). Thus, these techniques are limited in terms of prediction of the speech quality.

Perception-based speech quality measurement methods measure speech quality of a system in a way that closely resembles the human perception, which is necessary to characterize modern speech processing and enhancement systems since these systems are highly time-varying and non-linear. A perception-based technique uses natural speech signal to measure the speech quality of a system under test. It has good correlation with different subjective measures and can handle several distortions commonly encountered in modern telecommunication networks. ITU-T recommendation P.863, POLQA is the most comprehensive and extensively used perception-based objective method for measuring overall listening speech quality.

2.3 ITU-T P.863, POLQA

POLQA, ITU-T recommendation P.863 is an objective speech quality assessment technique that can be used to predict overall listening speech quality (as perceived by the user) in both narrowband and super wideband listening conditions. ITU-T P.863 emulates the subjects that rate speech quality using the ACR 5 point opinion scale (refer Table 1). The basic concept behind ITU-T P.863 is the same as used in its predecessor ITU-T P.862. Both of them use the same full reference method, where both the reference and degraded signals are used to predict overall listening quality. No further information about the system under test is available and the system is considered as a black box [1]. POLQA only measures the impact of one way distortion and noise on speech quality and it is possible to have high output score from POLQA, yet poor overall conversational quality [4]. Figure 2 gives the basic overview of ITU-T P.863. There are three major functional blocks in ITU-T P.863 namely:

- Temporal alignment
- Perceptual Model
- Cognitive Model

The perceptual model maps both the reference and degraded signal onto their corresponding internal representations i.e. psychoacoustic representations (pitch-loudness-time). The cognitive model uses the difference between the two internal representations to predict speech quality, which is expressed in terms of objective MOS-LQO (Mean Opinion Score Listening Quality Objective). The MOS-LQO obtained from ITU-T P.863 has a very high correlation with the MOS-LQS (Mean Opinion Score Listening Quality Subjective), which is the average quality score from a number of subjective tests using the five point ACR (Absolute Category Rating) opinion scale (see Table 1) [1].

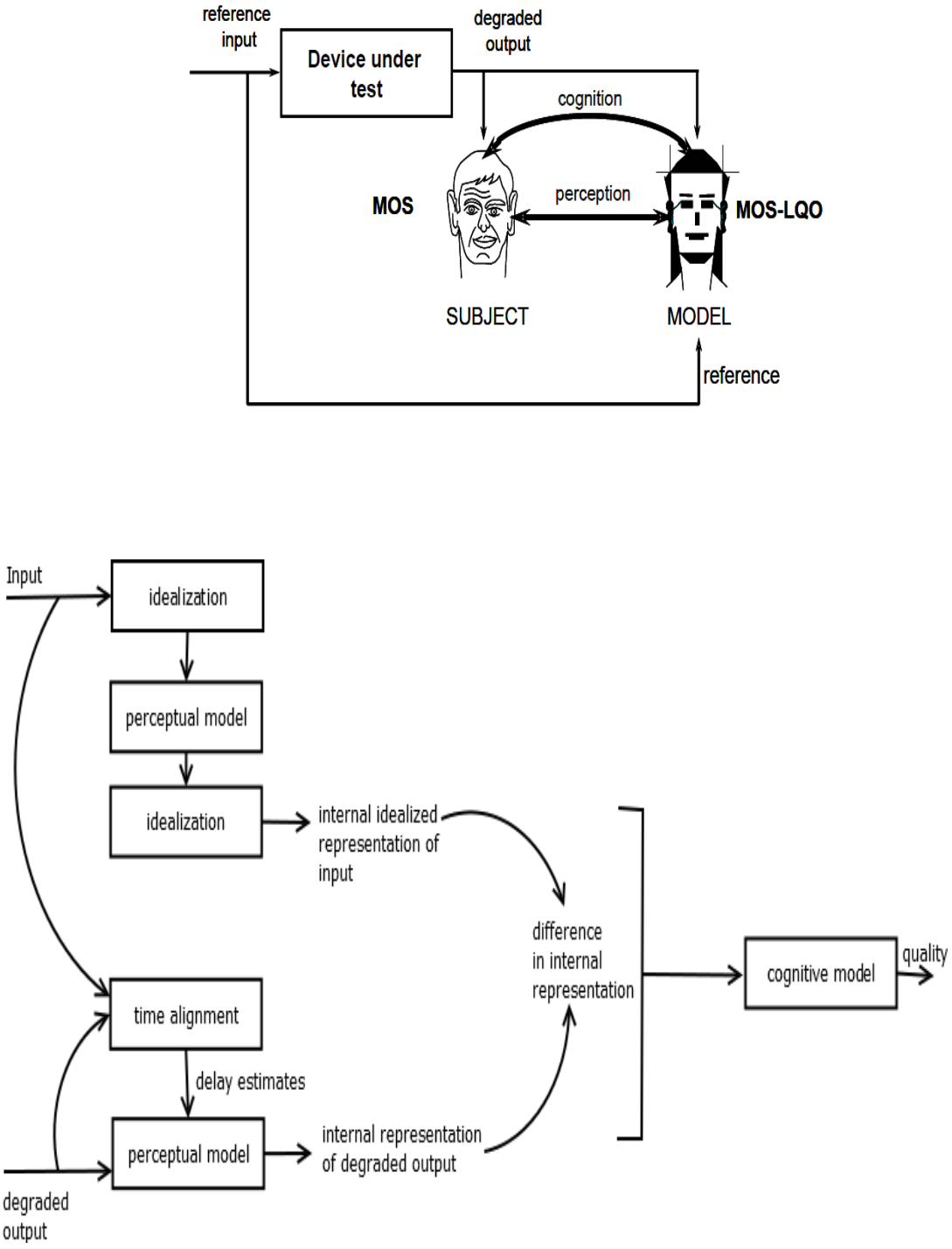

Figure 2. Basic overview of ITU-T P.863, POLQA. The computer model comprises a perceptual and a cognitive model (taken from [1]).

One of the most important improvements in ITU-T P.863 compared to ITU-T P.862 is the compensation of the quality difference in recording of original speech signals by a process called "idealization". This process suppresses the low level of noise inherently present in recording and tunes timbre of the original recording to a global acceptable timbre. Due to this "idealization" approach, a transparent system can be judged as a perceptually non-transparent system when a poor quality recording is used as an input signal. This is because of the ACR based listening tests used in the subjective experiments. Subjects do not hear the reference speech, but it is assumed that subjects have an idea of what the reference speech sounds like and a reference signal, which is not perfect, will be scored worse than the perfect. One should always remember that any perceptual assessment method does not measure the quality of the system under test, but it measures the quality of the output of the system. The quality of the system can be measured only by averaging the quality assessments over a large set of relevant speech signals [1].

2.3.1 Temporal alignment

The temporal alignment algorithm aligns reference and degraded signal, ensuring that the processing in the core model is based upon the comparison of the same speech segments in two signals. Both the reference and degraded signals are transformed to the time-frequency domain using overlapping FFT frames, the length of which depends on the sampling rate (1024 for 48 kHz, 512 for 16kHz and 256 for 8 k Hz) and is not necessarily the same as the ones used in the perceptual model (refer 2.3.2 Perceptual model). For each frame, a delay value is calculated. The delay value is always defined as the delay of reference signal compared to the degraded signal [1].

Temporal alignment algorithm consists of five major blocks namely:

- Filtering
- Pre-alignment
- Coarse-alignment
- Fine-alignment
- Section combination

Initially, both the reference and degraded speech signal are bandpass filtered. Most of the spectral energy of a speech signal lies in the range of 300 to 3500 Hz, so this range of

signal will provide the most reliable delay estimation. ''Pre-alignment' finds the active speech part of the signals and computes an initial delay estimate per frame and the theoretical maximum and minimum of the initial delay estimate. After pre-alignment, the signals go through 'Coarse alignment', which refines the delay value in each frame. It is a backtracking search algorithm and the resolution of coarse alignment is varied stepwise. 'Fine alignment' determines the exact sample delay of each frame. Now, all the frames with approximately the same delay values are combined (see Figure 3) [1]. Global time expansion and compression (time scaling) distortion is often seen in modern voice and audio systems e.g. as seen in many Voice over IP systems (VoIP). POLQA detects this distortion and if found compensates it. Finally, after determining the correct delay value and compensating the effect of time scaling distortion, the delay value and signals (reference and degraded) are passed onto the perceptual model [1].

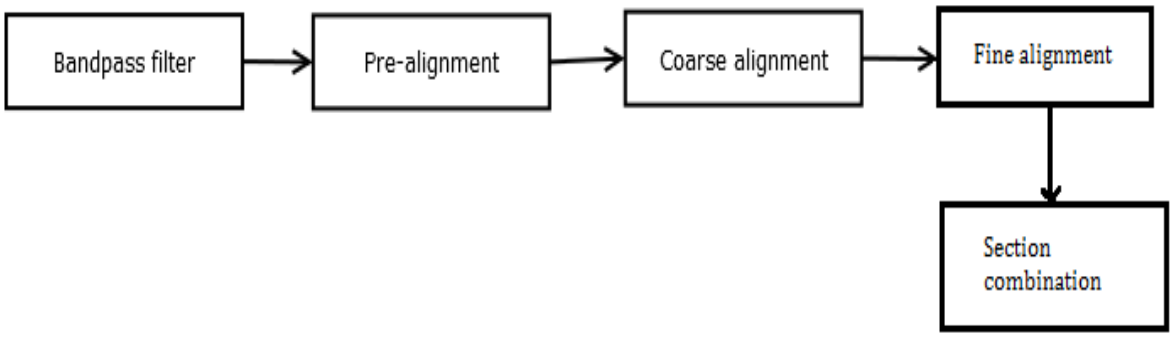

Figure 3. Five major blocks of the Temporal Alignment algorithm used in ITU-T P.863, POLQA (Redrawn from [8]).

2.3.2 Perceptual model

The perceptual model converts the incoming original and degraded signal into their corresponding internal representations (pitch-loudness-time).

Both the reference and degraded speech signals are scaled. The degraded signal is scaled towards playback level, while the reference signal is scaled towards a predefined fixed optimal level of about 73 dB SPL equivalent. After scaling both the original and reference signals are transformed to time-frequency domain with 50% overlapping FFT frames using a Hann window. The size of the window is varied according to the sample rate (2048 for 48 kHz, 1024 for 16 kHz and 512 for 8 kHz) to match the time analysis

window of the human auditory system [1]. Small pitch shifts resulting in a distorted frequency axis if found in the degraded speech signals, are compensated. Now, the frequency axis in 'Hz' is mapped onto the pitch scale in 'Bark' (the psychoacoustic scale) [3]. The resulting reference and degraded speech signals are termed as *'Pitch Power Density'* representations (power as the function of time and frequency). The three quality indicators in POLQA for frequency response distortions (FREQ), additive noise (NOISE) and room reverberations (REVERB) are calculated at this stage. Now, the masking in both the time and frequency domain is modeled to determine the perceived loudness for each frequency component. This is done for every frame in the reference and degraded speech signal. After modeling the masking effect, pitch density representations are then transformed onto pitch loudness density representations (loudness scale in 'Sone'). Low level of noise in the reference signal is suppressed (idealization) and the constant background noise in the degraded signal is partially removed (steady state noise has a lower impact than non-steady state noise). The reference signal is partially compensated for linear frequency response distortions (less perceptible than non-linear distortions). Now, the final noise suppression is done for both the reference and degraded speech signal (steady state noise in the loudness domain having a big impact) [3]. The resulting signals are perceptually relevant internal representations (pitch-loudness-time) of the reference and degraded speech signal [1].

Four different variants of the internal representations of the reference and degraded signal are calculated. Two of them focus on a normal range of degradations and loud degradations introduced by the system under test. The other two variants focus on normal and loud degradations for the added disturbances. These four variants are used to calculate two disturbance densities, one focusing on overall degradations while the other on the processing of added degradations. This models the asymmetry in impact among the degradations caused by the time-frequency components missing from the reference signal and the degradations caused by the introduction of new time-frequency components. The two final *'Disturbance densities'* (measure of perceptibility of distortions) are then passed onto the cognitive model [3]. An overview of the POLQA perceptual model can be seen in Figure 4 (taken from [1]).

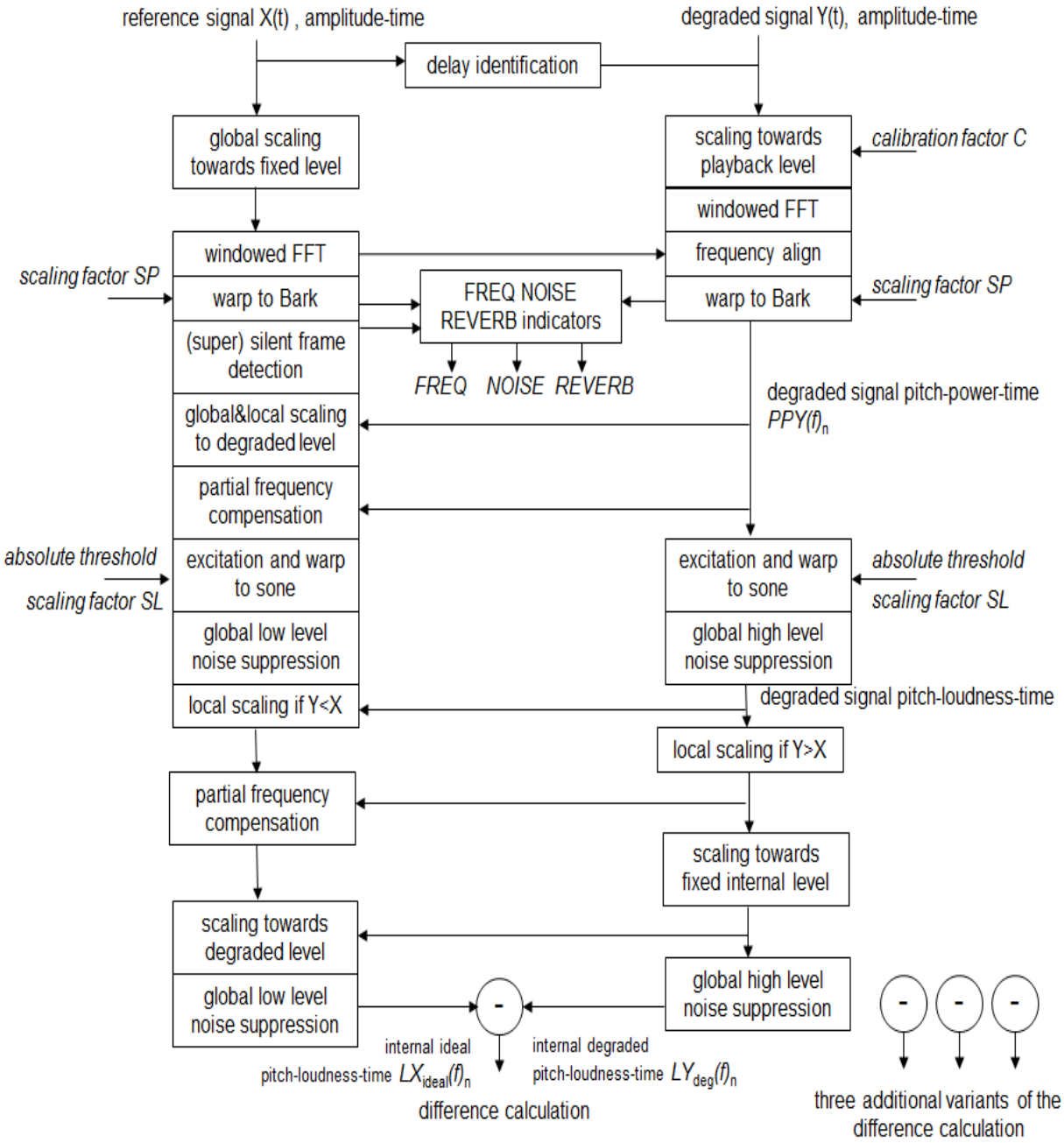

Figure 4. ITU-T P.863, POLQA - Perceptual model (Taken from [1]). Difference function is calculated in four different flavors.

2.3.3 Cognitive model

Disturbance density is an indicator for the perceptibility of distortions; cognitive effects are not taken into consideration yet. The cognitive model considers cognitive aspects which is important for humans to score the quality of what they perceive. It converts the perceptibility measure i.e. disturbance density into the annoyance measure. The conversion is performed by compensating disturbance density for various conditions [3]:

- Level variations
- Frame repetitions
- Timbre
- Spectral flatness
- Noise switching during speech pauses
- Delay variations
- Variation of disturbance density
- Loudness jumps

Two more indicators namely LEVEL (severe deviations from the optimal listening level) and FLATNESS (spectral flatness) are calculated at this stage [9]. The disturbance densities are aggregated over pitch, spurts and time. The added disturbance density is compensated for loud reverberations and loud noises using REVERB and NOISE indicators. The two disturbance densities are combined with the FREQ indicator to derive a MOS like intermediate indicator. The raw POLQA score is derived from the MOS like intermediate indicator using LEVEL, FLATNESS indicators and 2 different aggregations of disturbance densities over pitch, spurts and time. The raw POLQA score is then mapped on to MOS-LQO using third order polynomial mapping, which is based on a huge set of databases [1]. More details on the temporal alignment and perceptual model used in POLQA can be found in [1].

2.4 Objective speech intelligibility measurement

There are several objective assessment methods for measuring speech intelligibility. Signal-to-noise ratio based techniques like AI (Articulation Index) and AI based STI (Speech Transmission Index) are widely used standardized methods worldwide.

The AI technique, which was developed in late 40's, is based on the assumption that the intelligibility of a speech signal is the sum of contribution from individual frequency bands. This technique defines 20 conterminous frequency bands, which contribute equally to a defined index, *Articulation Index (AI)*. The AI value ranges from 0 to 1. The closer the value is to 1, the better is the intelligibility [13].

The AI based STI method uses test speech signals, which have spectro-temporal characteristics similar to those found in natural speech [16]. It evaluates the test signal as human ears would evaluate it. This method assumes that prediction of intelligibility is based on the weighted contribution from a number of frequency bands, which is similar to the AI concept. However, the AI method uses 20 contiguous frequency bands (not equal bandwidth) with similar band importance value, while the STI uses a weighted contribution (not similar). This contribution is based on effective SNR, which is determined by several factors like noise, reverberation, echoes etc. [12]. In the STI method, the amount of reduction in the intelligibility due to distortions is modeled as the reduction in temporal envelop modulations [17].

This method gives a single value between zero (completely unintelligible) and one (perfect intelligibility). This value quantifies how well a transmission channel transmits information with respect to intelligibility [16]. A diagram depicting the metric used by the STI method is shown in Figure 5.

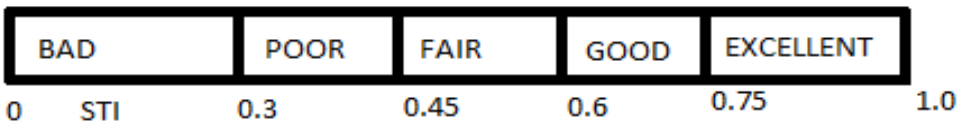

Figure 5. Scale used by STI (taken from [13])

The AI based STI method fails when the test speech signal is introduced to some non-linear processing. This processing may introduce new modulations and the model predicts improvement in the SNR value, giving an incorrect measure of speech intelligibility. Moreover, modern digital speech transmission channels use advanced coding techniques and these codecs behave differently for speech and for STI test signal

(modulated-noise). Thus, using the STI method in modern telecommunication networks will yield an incorrect prediction of intelligibility [17]. The STI method has been verified for several linear distortions (especially when the speech signal passes through a noise suppression algorithm) and although a lot has been done to model STI for non-linear distortions, these modifications have not been verified yet.

3. Subjective intelligibility assessment of individual CVC words

This chapter describes the subjective method used for measuring speech intelligibility in this thesis. The reason for choosing the type of listening test used in this thesis is presented and the procedure used for conducting listening tests is described in detail with necessary diagrams.

Subjective experiments for measuring intelligibility use recorded or live speech signals to be assessed by a panel of listeners. The speech material presented to a listener represents the output of the transmission of a system under test. Subjective listening tests are used to measure the impact of distortions on speech intelligibility. Careful design and planning of listening experiments ensure that the uncontrolled factors (context effect, individual variation, bias, cultural and gender variation etc.) do not affect the listening tests. Test materials can be presented randomly to prevent bias in the response of subjects (one of the examples of this effect can be a subjective quality test where the quality of previous speech can bias a subject's opinion score for the next speech). Expertise (in terms of listeners) and number of listeners can vary depending on the type of the test conducted.

Subjective assessment of intelligibility can vary [13]:

- Based on speech signals used for assessment i.e. phonemes, words (meaningful or nonsense) or sentences. Nonsense CVC words (Consonant Vowel Consonant) are widely used.

- Based on the way the test material is presented to the listeners (e.g. embedding words into a carrier phrase).

- Based on the response from the listeners: Closed or open response. In open response, there are no alternatives to choose from and subjects respond based on what they think they heard, while in closed response subjects choose from a list of alternatives.

Closed response does not require extensive training in terms of listeners. However, open response requires extensive training, especially when nonsense CVC words are used [13]. Subjective listening experiments provide an accurate and reliable intelligibility measure provided the experiments be conducted in stringent conditions, but setting it up is quite expensive and time consuming.

3.1 Used approach

In this thesis work, a simple computer-based subjective listening experiment based on nonsense CVC words and closed response is used for measuring the speech intelligibility of individual CVC words. A paper-based test can be used, but analyzing results becomes easier with a computer-based approach. Use of nonsense CVC words with closed response has many advantages (see Figure 6):

- The use of nonsense CVC words minimizes the context effect (no interference regarding the meaning of the word, spelling).

- CVC words have better qualification range compared to other speech signals (short sentences, numbers etc.). They discriminate over a wider range. This can be seen in Figure 7, which shows the qualification and relation between several intelligibility scores and the STI value. As one can see, other tests suffer from the ceiling effect (saturation) [13].

- Closed response does not require listeners to go through extensive training to be a part of a listening experiment.
-

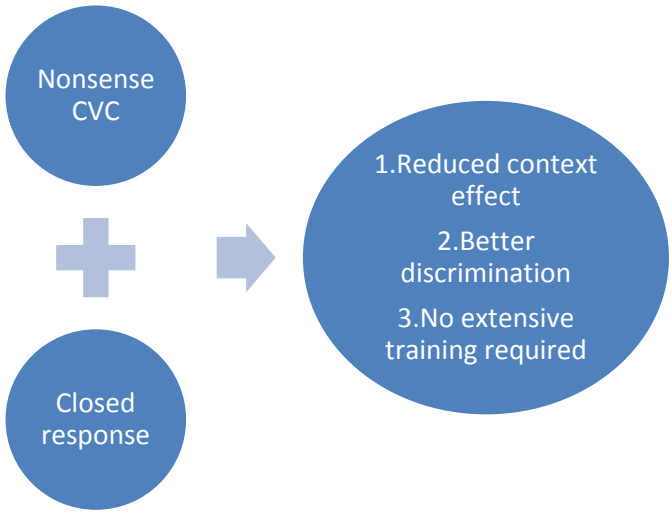

Figure 6. Advantages of nonsense CVC words and closed response subjective listening tests.

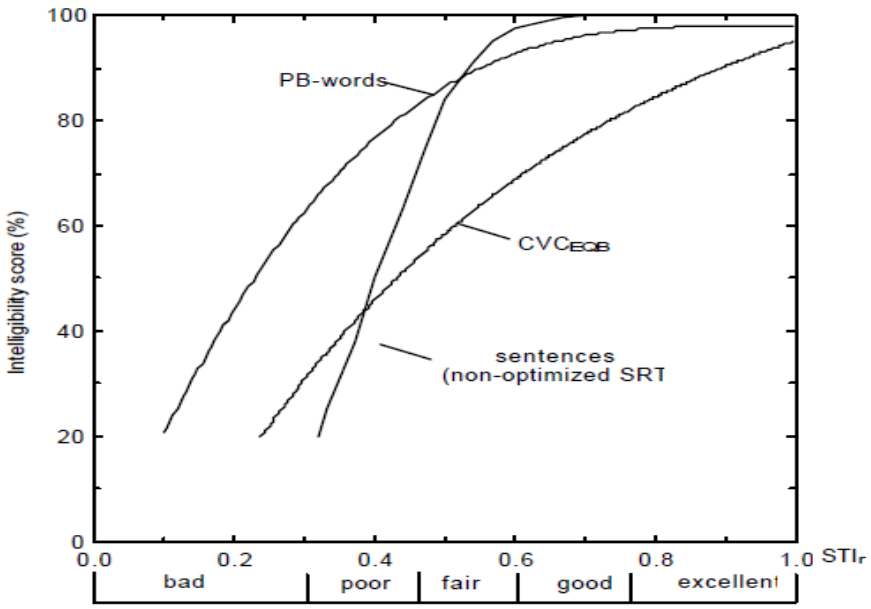

Figure 7. Qualification and relation between various intelligibility scores and the STI (taken from [13]).

3.2 Setup and procedure

Developing a good objective speech intelligibility assessment method requires a large amount of reliable subjective data (speech signals and subjective scores) [1]. High fidelity recording of reference speech signals, degraded speech signals containing wide variety of distortions and a good listening test interface are thus essential.

Nonsense CVC words used as the speech material are recorded from two different sources (two male voices), using a high quality omni-directional microphone and a digital audio interface. The recordings are done with a sampling frequency of 48 kHz, 16 bit and mono channel settings in a low noise and reverberation chamber. The recordings meet the requirements set for ITU-T P.863 super wideband measurement (refer section 1 in [1] for further details). Adobe Audition version 1.5 is used for audio processing.

An example list of CVC words recorded for the listening tests is given in Table 2. The list comprises seven different CVC words. The generic representation of the CVC words on the list below is called '/C/ous', where initial consonant 'C' varies in each word whereas 'ous' remains the same.

Dous
Fous
Hous
Kous
Lous
Mous
Rous

Table 2. A list of CVC words. Seven different '*Cous*' words constitute the list.

The recorded files (Figure 8 shows one of the recorded files) are pre-filtered using 50 to 14,000 Hz band-pass filter and leveled to -26 dB overload as per ITU-T recommendation (refer section 1 in [1] for further details). Individual'*/C/ous*' words (for e.g. 'Dous') are extracted and these extracted '*/C/ous*' words represent the reference speech signals, i.e. original speech signals without any degradation (see Figure 9).

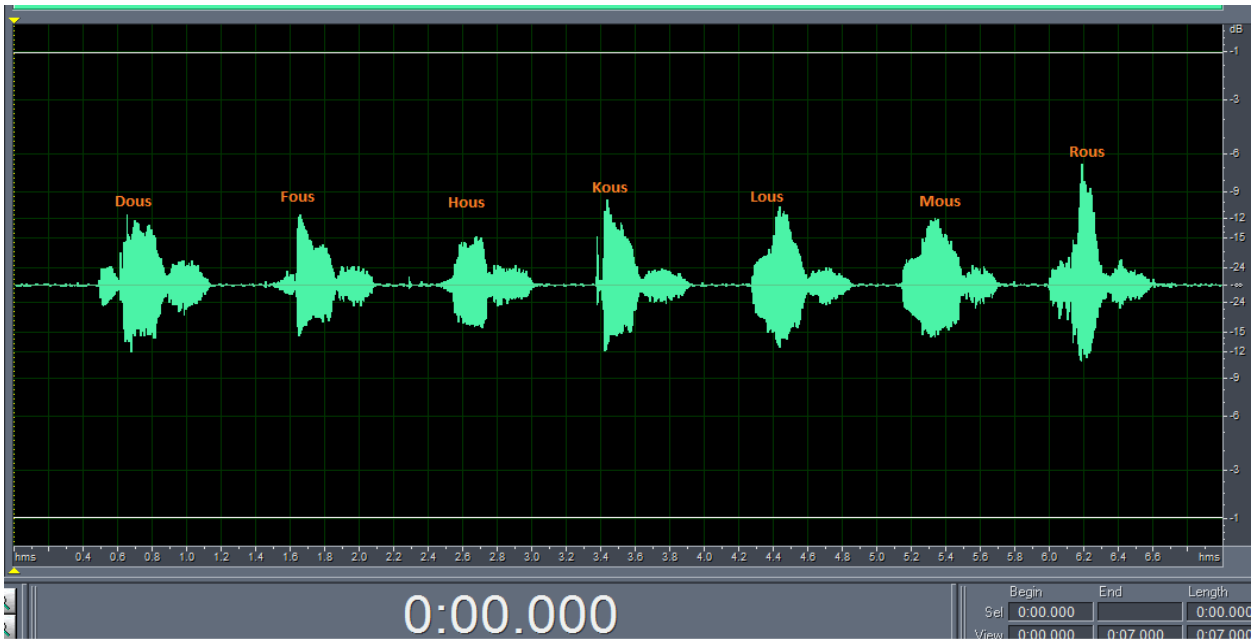

Figure 8. A recorded 'C/ous/' file from a source. It contains seven different CVC words. This file is pre-filtered and leveled before extracting individual CVC words.

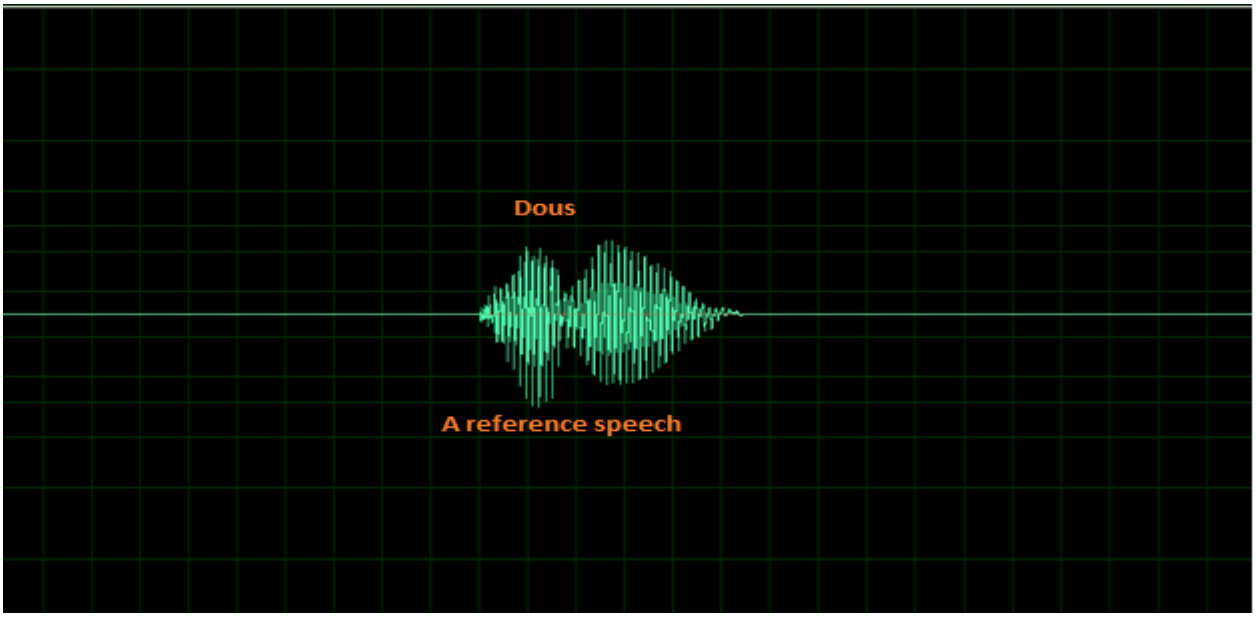

Figure 9. An extracted CVC word ('Dous'). This is an individual '*/C/ous*' word and represents a reference speech signal.

A degraded speech signal represents the output signal from the device under test (DUT). In this thesis work, degraded speech signals are generated by degrading original speech signals (individual '*/C/ous*' words) using several degradations commonly seen in the modern telecommunication networks. The degradations include background noise, time clipping (loss of information), bandwidth limitation, and pulse distortion to name a few. A detailed description of the degradations used in the listening tests can be found in Appendix A: List of degradations.

The degraded speech signals are randomized and concatenated. Each concatenated audio file contains 42 degraded '*/C/ous*' words (each '*/C/ous*' word is degraded with six different degradations). Silence is added after each '*/C/ous*' word to ensure that the listeners get sufficient time to choose an alternative. A concatenated file used for the listening test can be seen in Figure 10.

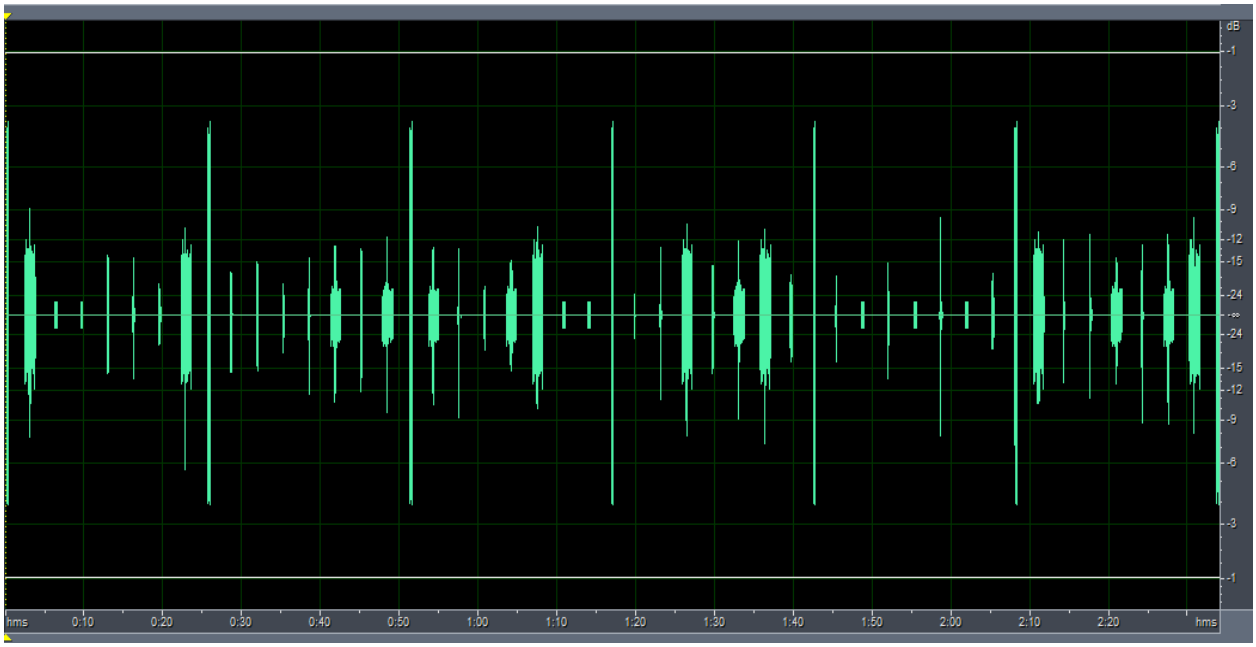

Figure 10. Concatenated audio file used for a listening test. A silence of 2 seconds is added after every degraded '*/C/ous*' word.

There are several ways to present the results from the subjective listening experiments. ITU-T P.863 uses MOS scale, which allows subjects' to rate the quality of speech fragments on a five-point opinion scale. However, speech intelligibility cannot be assessed based on some opinion scale, as a speech fragment (individual '*/C/ous*' words) is either understandable or not.

Subjects choose an alternative (a '*/C/ous*' word) from a list of alternatives based on what they think they heard. There are eight alternatives in the interface, and alternatives do not contain any extra word other than the '*/C/ous*' words. 'Not recognized' as the name suggests can be chosen when a listener cannot recognize the heard '*/C/ous*' word. For each correct assessment, a score of one is given and for an incorrect or skipped/not recognized assessment, a score of zero is given. The listening test setup is designed in Microsoft Excel. A snapshot of the interface (a portion of it) before and after a listening test is given in Figure 11 and Figure 12.

Tick the appropriate box as you hear the recording e.g. If you hear Dous tick Dous Column									
BEEP	Dous	Fous	Hous	Kous	Lous	Mous	Rous	Not Recognized	Sco
Stimulus 1									-
Stimulus 2									-
Stimulus 3									-
Stimulus 4									-
Stimulus 5									-
Stimulus 6									-
Stimulus 7									-
BEEP	Dous	Fous	Hous	Kous	Lous	Mous	Rous	Not Recognized	Sco
Stimulus 8									-
Stimulus 9									-
Stimulus 10									-
Stimulus 11									-

Figure 11. A snapshot of the interface (Microsoft Excel) used for listening tests. Each stimulus represents a degraded '*/C/ous*' word.

Tick the appropriate box as you hear the recording e.g. If you hear Dous tick Dous Column									
BEEP	Dous	Fous	Hous	Kous	Lous	Mous	Rous	Not Recognized	Sco
Stimulus 1							X		-
Stimulus 2								X	-
Stimulus 3			X						-
Stimulus 4	X								-
Stimulus 5			X						-
Stimulus 6				X					-
Stimulus 7					X				-
BEEP	Dous	Fous	Hous	Kous	Lous	Mous	Rous	Not Recognized	Sco
Stimulus 8								X	-
Stimulus 9								X	-

Figure 12. Snapshot of the test interface (Microsoft Excel) at the end of a listening test. Automatic highlighting of the chosen cell ensures that users keep track of the stimuli heard.

Every listening test uses five concatenated audio files. These concatenated files are obtained by concatenating degraded speech signals based on five different random orders of degraded speech signals. This means that a listener hears the same 42 degraded '*/C/ous*' words five times, but played in five different random orders and thus assessing each degradation five times. Other listeners in the panel go through the similar procedure. Therefore, every degraded '*/C/ous*' word (42 in total) is scored 5 times (five random orders) the number of listeners in the panel (four listeners in a panel), i.e. twenty times. The average of these scores represents the subjective intelligibility score for a particular degraded '*/C/ous*' word. The value of the subjective score lies between 0 (not intelligible) and 1 (perfect intelligibility). A general flowchart describing the subjective test setup and procedure is given in Figure 13.

If a subject misses assessing any '*/C/ous*' word, he/she can restart the test from that particular point. No restrictions are made regarding the number of times a listener can take the same test (in terms of restarting the test if subjects miss assessing a '*/C/ous*' word).

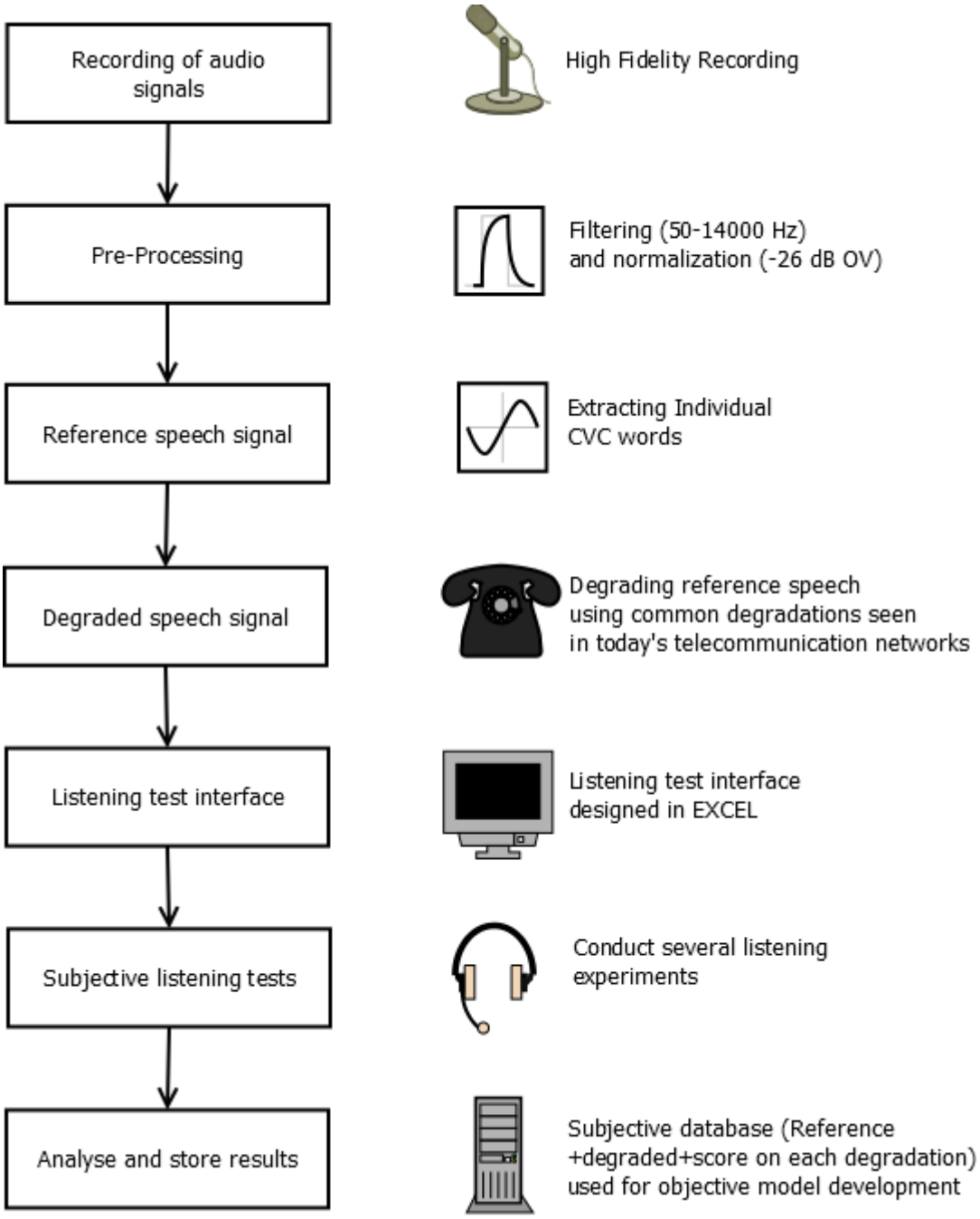

Figure 13. Subjective Experiment Setup. A flowchart describing the sequence of steps followed for the subjective intelligibility assessment of individual CVC words.

4. Objective intelligibility assessment of individual CVC words

This chapter describes the simple new objective model developed in this thesis for measuring intelligibility of individual CVC words. ITU-T P.863 is used 'as is' in the beginning for predicting single CVC word intelligibility. Because of its poor performance (low correlation, a first extension based on the restructuring of speech files is developed. This produces a slight improvement in correlation. A simple new model is developed which performs reasonably good in the first subjective database. The validation of this model in a new subjective database is also presented.

Objective assessment of speech intelligibility refers to the use of a computer model, which simulates the perceived intelligibility based on the subjective measurements. Subjective experiments depend on listeners (context effect, individual variation, bias, cultural and gender variation etc.) and balance of the conditions tested, whereas an objective assessment is context independent and represents the average score by a panel of listeners. It is less laborious and highly reliable in terms of repeatability. A good objective measurement technique has very high correlation with several subjective experiments. Validating the model on various subjective databases gives us the validity range of that model.

Before developing an objective model, the assessment of limitations is essential. One of the limitations is the 'bias effect' seen in the subjective tests, which cannot be predicted using an objective model. Time-clipped '*/C/ous*' words can be a source of such effect. Clipping the initial consonants '*/C/* ' from '*/C/ous*' words might bias subjects towards one particular '*/C/ous*' word (a subject might hear all the time-clipped '*/C/ous*' words as 'Dous'), resulting in a perfect intelligibility score for that particular word (cannot be predicted). Such points (which show 'bias effect') if encountered are omitted during the modeling in this thesis work. This effect is not seen in the */C/ous* database, but was clearly seen in the validation database i.e. *a/C/* database (refer 4.4 Validation) and the points showing this effect were removed.

This clearly means that objective assessment based on a certain model might have many restrictions and an objective assessment should not be used as a substitute for a subjective experiment.

4.1 ITU-T P.863 'as is'

ITU-T P.863, POLQA is used as a starting point because:

- It can handle a wide range of degradations encountered in modern telecommunication networks.
- Although speech quality and intelligibility are different attributes of speech, they have a strong correlation.

ITU-T P.863 in super-wideband mode is used 'as is' in the beginning for predicting intelligibility scores. After conducting several listening experiments, a large amount of subjective data (speech signals and subjective scores) is obtained. Individual reference (e.g. 'Dous') and degraded '*/C/ous*' words (e.g. time-clipped 'Dous') are used as speech material. The result shows a very disappointing correlation of 0.30 when using standard ITU-T P.863 output values. A plot between the subjective intelligibility scores and standard ITU-T P.863 output values is given in Figure 14. For many degraded '*/C/ous*' words, subjective scores are high while P.863 scores are low i.e. subjects can understand these words (intelligible), while the model predicts it otherwise. Although, the distribution (data points in Figure 14) is not even across the intelligibility scale, one should understand that the subjective experiments are used to measure the impact of distortions and noise on intelligibility and behavior of subjects cannot be predicted beforehand (to have an even distribution).

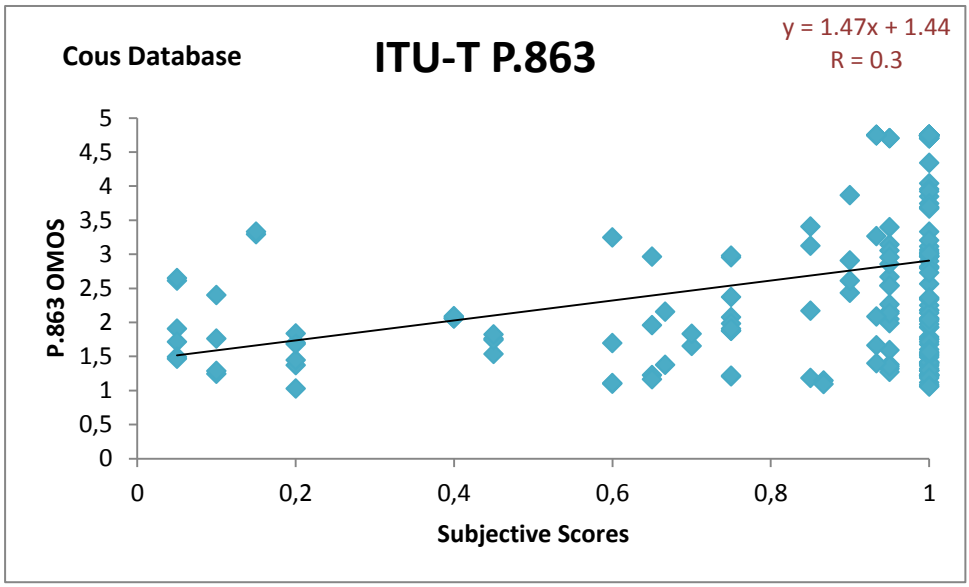

Figure 14. Correlation between subjective intelligibility scores and standard ITU-T P.863 output in super wideband mode for the '*/C/ous*' database. There were 196 degraded '*/C/ous*' words, the degradations used were: reduced playback level, bandwidth limitation, background noise, time-clipping and pulse distortion.

4.2 Temporal restructuring of speech material

The initial problem identified for such a disappointing correlation is the temporal structure of the speech material. The structure does not comply with the temporal structure followed in the standard ITU-T P.863 measurements. The reference speech files used in the standard ITU-T P.863 measurements contain two sentences separated by a gap of silent interval. The length of silent parts i.e. leading, trailing and the silent interval between sentences/speech fragments in the reference speech file should be neither too long nor too short. A value in the range of 500-2000 ms is desirable. Thus, restructuring of speech material is necessary to meet the above requirements.

The easiest way to restructure speech material is by concatenating every reference and degraded '*/C/ous*' word with itself (a pair of same '*/C/ous*' word e.g. 'Dous+Dous'). A generic name '*Cous*_double' is used to represent restructured (concatenated) speech material. Digital silence is added to ensure that the length of the silent parts in every '*Cous*_double' file is at least 500 ms. A restructured reference speech file, 'Dous_double' is shown in Figure 15.

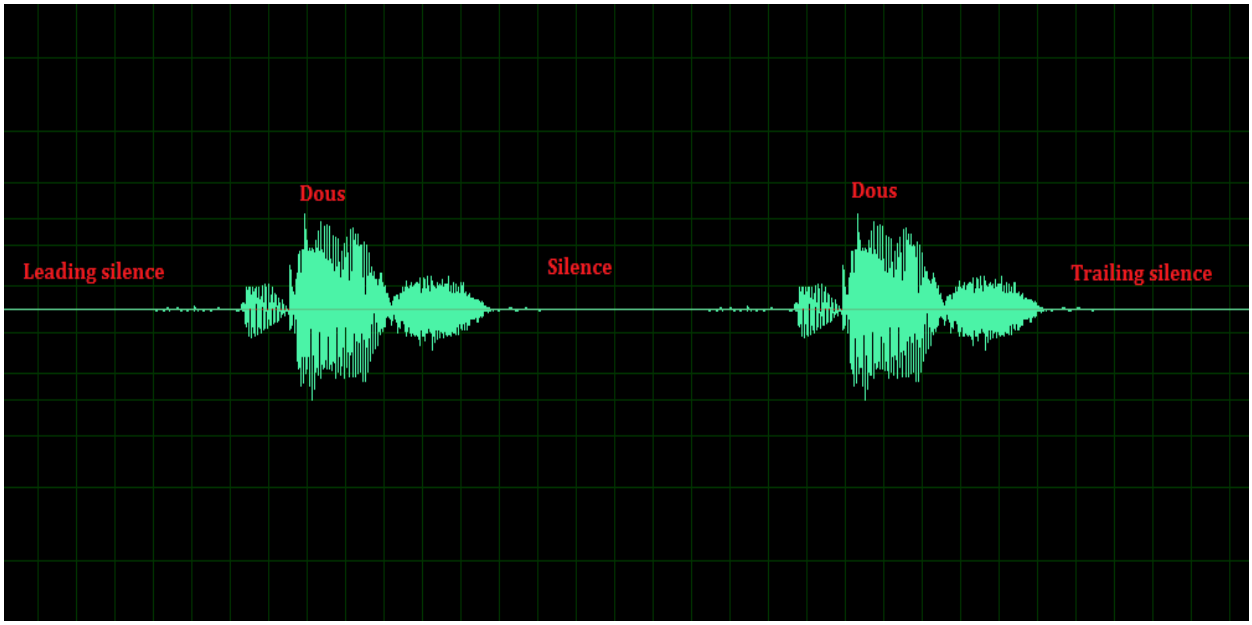

Figure 15. 'Dous_ double' reference speech signal. Leading silence, trailing silence and the silent interval between two 'Dous' words are shown.

Using the '*Cous*_double' database does not improve the correlation significantly. A slight increase in correlation from 0.30 to 0.34 is observed (see Figure 16). Nevertheless, '*Cous*_double' is used for further development, as it meets the temporal requirements set for the standard ITU-T P.863 measurements.

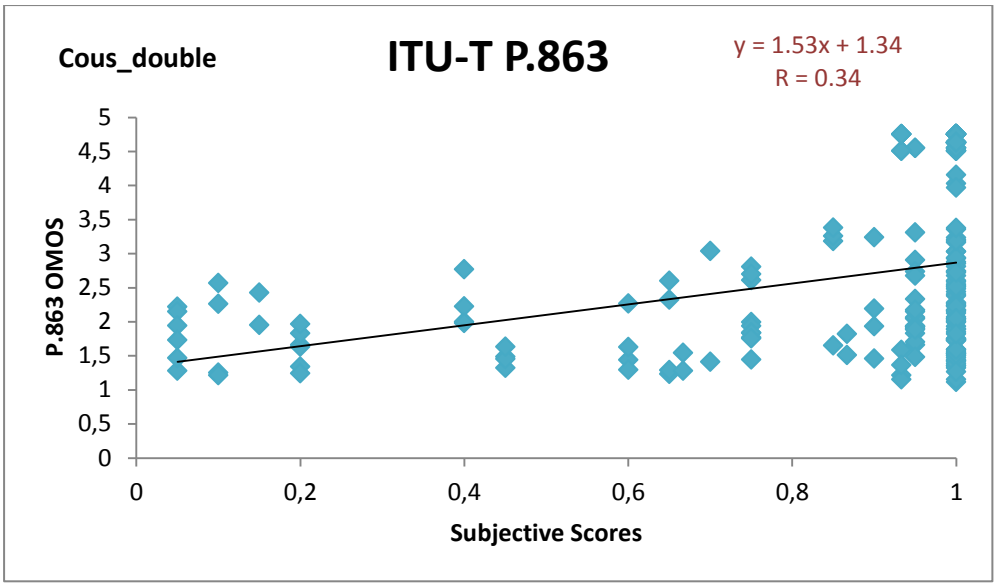

Figure 16. Correlation between subjective intelligibility scores and standard ITU-T P.863 output in super wideband mode for the '*Cous_double*' database. There were 196 degraded '*/C/ous*' words, the degradations used were: reduced playback level, bandwidth limitation, background noise, time-clipping and pulse distortion.

4.3 A simple new model for predicting individual CVC intelligibility

The ITU-T P.863 model was developed for assessing speech quality and a disappointingly low correlation for '*Cous_double*' database when using standard P.863 output values reflects the need for an improved model for predicting subjective intelligibility scores.

A simple new model that uses a simple time clipping indicator (missing speech parts) and a "Good frame count" based on the local perceptual SNR (frame by frame) is developed in this thesis for predicting individual CVC scores. It is integrated to the P.863 perceptual model (see Figure 17). The cognitive model of P.863 is totally discarded.

- ✓ *A simple time clipping indicator* (Indicator1)
- ✓ A "Good frame count" indicator based on local (frame by frame) *Perceptual SNR* (Indicator2)

Perceptual SNR, although not being a good indicator for measuring speech quality can be a good indicator for measuring speech intelligibility. A linear combination of these two indicators (Indicator1 and Indicator2) is a simple yet effective model for predicting speech intelligibility of individual CVC words. A short description and pseudo codes for calculating these two indicators are given below:

Indicator1:

This indicator is calculated separately and is the ratio of the number of speech active frames in the original signal to the number of speech active frames in the degraded signal. It gives a rough indication of the impact of time clipping (missing speech parts) distortion. When the value of **Indicator1** is less than 1, this indicator does not represent time clipping, but the introduction of loud noise and pulses. In order to only model time clipping, all the values less than 1.0 are set to 1.0.

All the values of **Indicator1** that are more than 2.0 are set to 2.0. This models the saturation effect seen in the subjective tests (Clipping more than a certain extent will not vary the subjective score). Δ ensures that **Indicator1** does not attain the value of infinity when there are no speech active frames in the degraded signal. The value of Δ is optimized to get maximum correlation (optimized value of Δ is 1.0). Implementation is given in the form of pseudo code below:

```
Indicator1 = (num_active_original + Δ)/( num_active_degraded + Δ);
IF (indicator1 < 1.0) indicator1 = 1.0; // Noise and pulses not modeled
IF (indicator1 > 2.0) indicator1 = 2.0;    // Saturation effect
```

Indicator2:

This indicator is calculated from the internal representations (pitch-loudness-time representation) of the reference and degraded signal. Original and degraded speech signal are scaled towards each other globally and locally using a long time constant. This scaling compensates for the slow variations in gain (inaudible) and also for the global loudness gain factor. The difference between the loudness levels of original and degraded speech signal is calculated for every speech active frame. This difference is termed as the **level_difference.** Two variants of **level_difference** are calculated. One focuses on time clipping allowing better modeling of missing speech parts (**Indicator1** gives a rough estimate of time clipping distortion), while the other focuses on loud noises and pulses. This models the asymmetry in impact caused by the missing time-

frequency components from the reference signal and the introduction of new time-frequency components.

The ratio of **level_difference** to the original loudness level is computed. This ratio is termed as **perceptual_SNR**, and tells us how loud the difference is compared to the original signal. A condition based on **perceptual_SNR** value is used to count the number of acceptable frames in terms of intelligibility. The total count is called **good_frames**. A ratio of **good_frames** to the total number of speech active frames (**num_active_frame**) gives **Indicator2**. Implementation is given in the form of pseudo code below:

```
global_scaling (original_signal, degraded_signal);
local_scaling (original_signal, degraded_signal);

FOR (frame = start_frame,frame <= stop_frame, frame++)

        IF (speech_active_frame) num_active_frame++;

                IF (original_frame_level > degraded_frame_level) //time clip
                                level_diff = (original_frame_level -

degraded_frame_level);
                        perceptual_SNR = level_diff/original_frame_level;
                                        IF (perceptual_SNR < 0.2) good_frames++;
                                        // increase the count of good frames

                ELSE                                                    //noise and pulse
                        level_diff = (degraded_frame_level -
                                                                        original_frame_level);
                        perceptual_SNR = level_diff/original_frame_level;
                                        IF (perceptual_SNR < 0.2) good_frames++;
                                        // increase the count of good frames
                END
        END
END

Indicator2 = good_frames/ num_active_frame;
```

A flowchart describing the simple model can be seen in Figure 18.

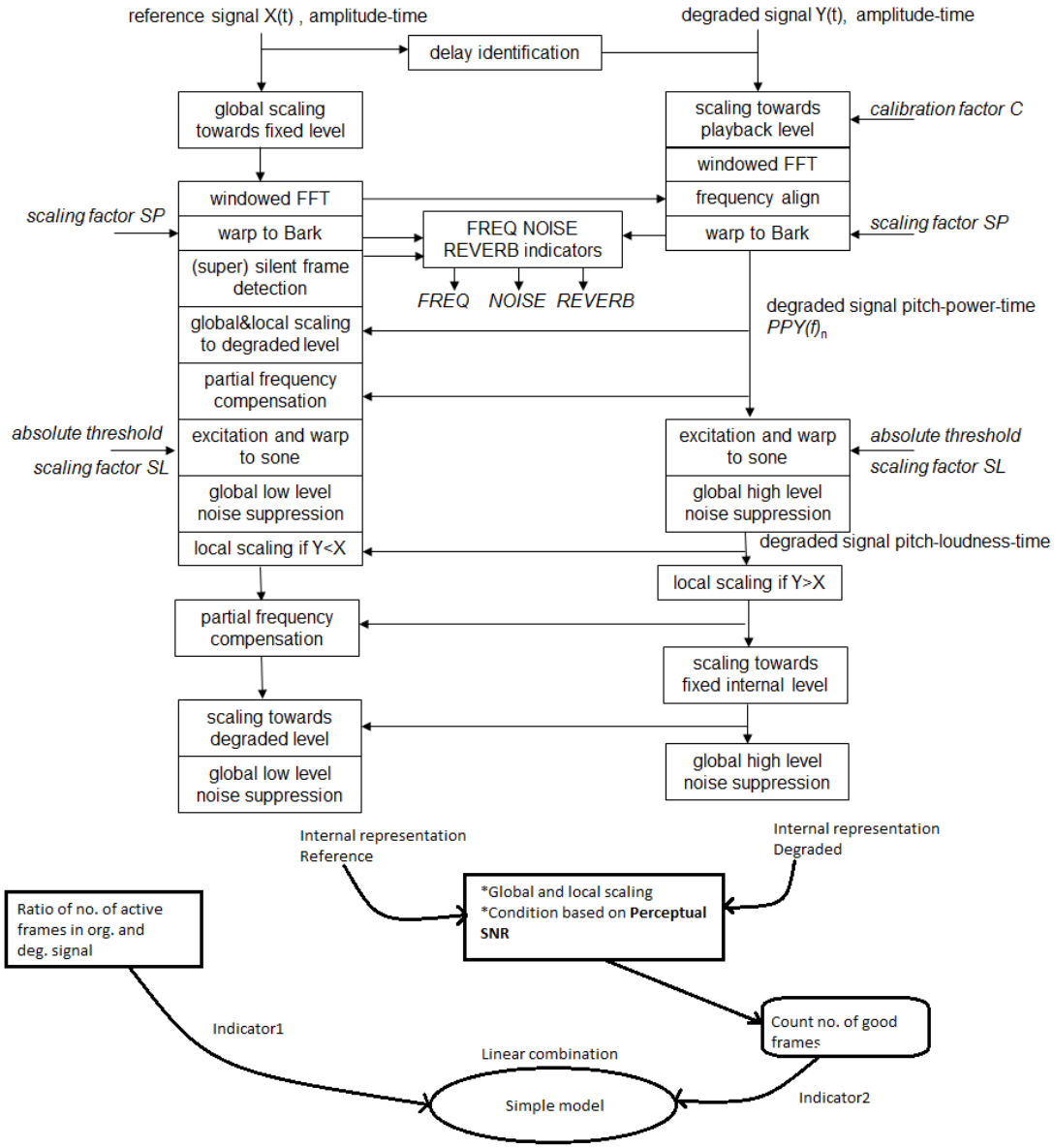

Figure 17. Integration of the simple model in ITU-T P.863 perceptual model. A linear combination of Indicator1 and Indicator2 is used as the model to predict individual CVC intelligibility scores. Cognitive model of ITU-T P.863 is totally discarded.

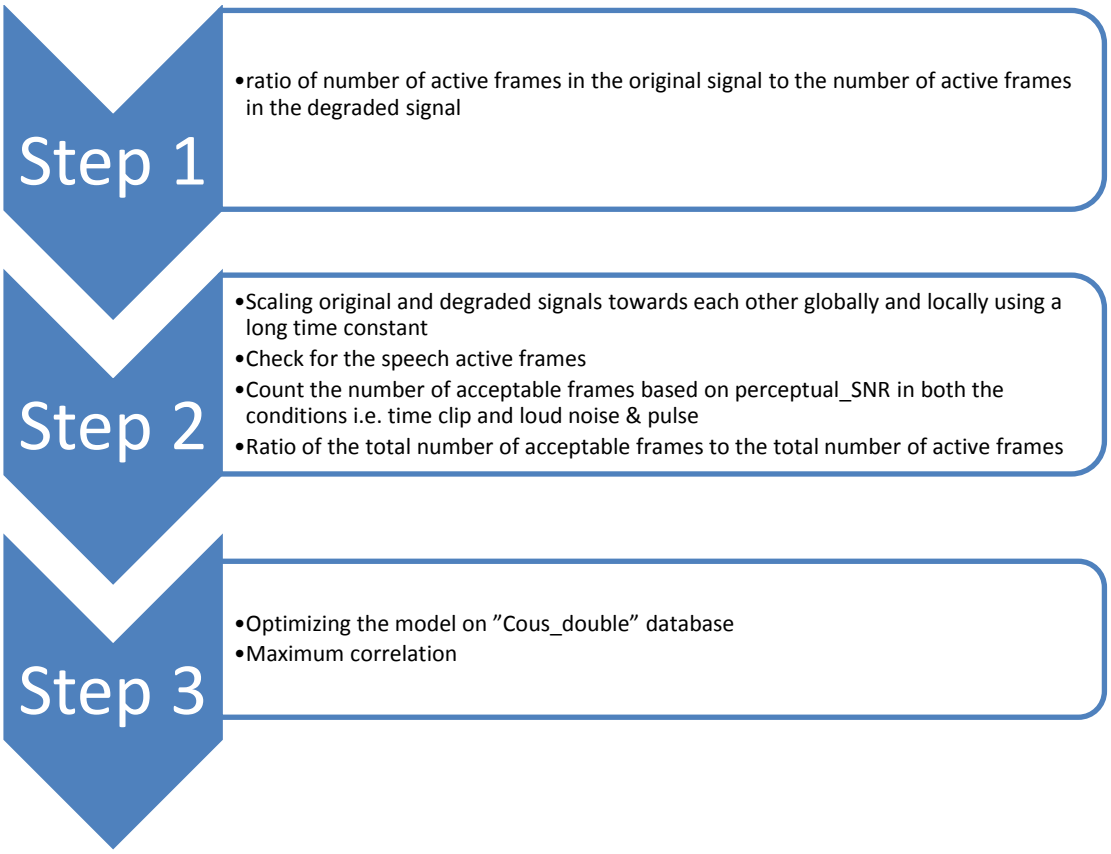

Figure 18. A flowchart illustrating the simple model.

The simple new model is used to predict subjective intelligibility scores for '*Cous*_double' database. The correlation between the model output values and the subjective intelligibility scores is 0.81 (see Figure 19), which is a significant improvement compared to a correlation of 0.34 obtained when using standard ITU-T P.863 output values. The prominent outliers are some time-clipped degraded speech signals, for which the predicted scores are low whereas the subjective scores are high indicating perfect intelligibility. The next step is validation of the simple model on a new subjective database based on nonsense VC words (Vowel Consonant).

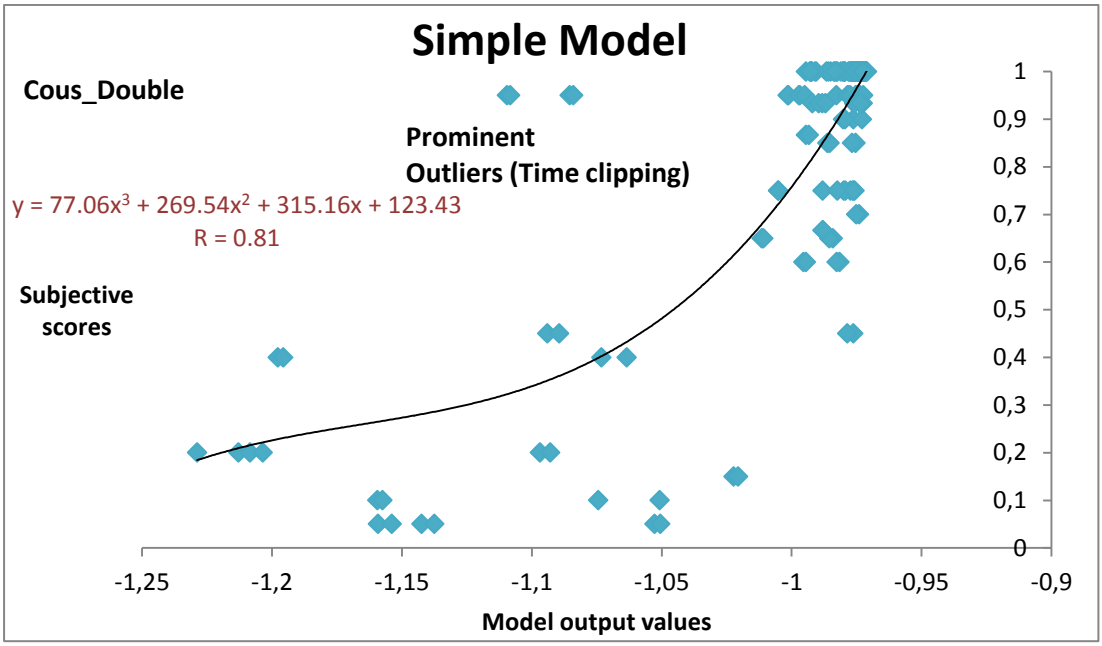

Figure 19. Correlation between the subjective scores and the model output values for '*Cous*_double' database. A reasonably good correlation of 0.81 is obtained. There were 196 degraded '*/C/ous*' words, the degradations used were: reduced playback level, bandwidth limitation, background noise, time-clipping and pulse distortion.

4.4 Validation

A new subjective database based on nonsense VC words (Vowel Consonant) is used for validating the simple model. The subjective experiments for VC words are conducted the same way it was done for the '*/C/*ous' words (refer 3.1 Used approach). The test interface and pre-processing of signals remain the same. The only difference is the inclusion of some new distortions (refer Appendix A: List of degradations for further details) that were not used for the first subjective database (*/C/*ous database).

An example list of VC words recorded is given in Table 3. The list comprises seven different VC words. The generic representation of the VC words in the list is called '*a/C/*', where 'C' represents the varying consonant at the end, whereas 'a' remains the same in the beginning of each word. Concatenating every reference and degraded '*a/C/*' word with itself gives the '*aC*_double' database.

The 'bias effect' is seen in the subjective tests conducted for '*a/C/*' words. All the time clipped '*a/C/*' words were perceived as 'aH' by the subjects, resulting in the perfect intelligibility score for 'aH'. The model cannot predict this, so 'aH' words are removed from the '*aC*_double' database (the model is used to predict intelligibility for '*aC*_double' database excluding 'aH' words).

The result for the validation database ('*aC*_double' database) shows a reasonably good correlation of 0.76 when using the model output values (Figure 20). It is important to remember that a different mapping is used for '*aC*_double' database compared to */C/*ous database ('*a/C/*' words contain a single consonant and a vowel contributing equally towards intelligibility, while '*/C/*ous' words have two consonants and a vowel and consonants contribute more towards intelligibility).

A clear outlier pattern is seen. These outliers are degraded speech material suffering from severe bandwidth limitation and amplitude clipping distortion, which are new set of degradations added to '*a/C/*' database (not present in '*/C/*ous' database).

aF
aG
aH
aJ
aK
aL
aM

Table 3. A list of VC words. Seven different '*aC*' words constitute the list.

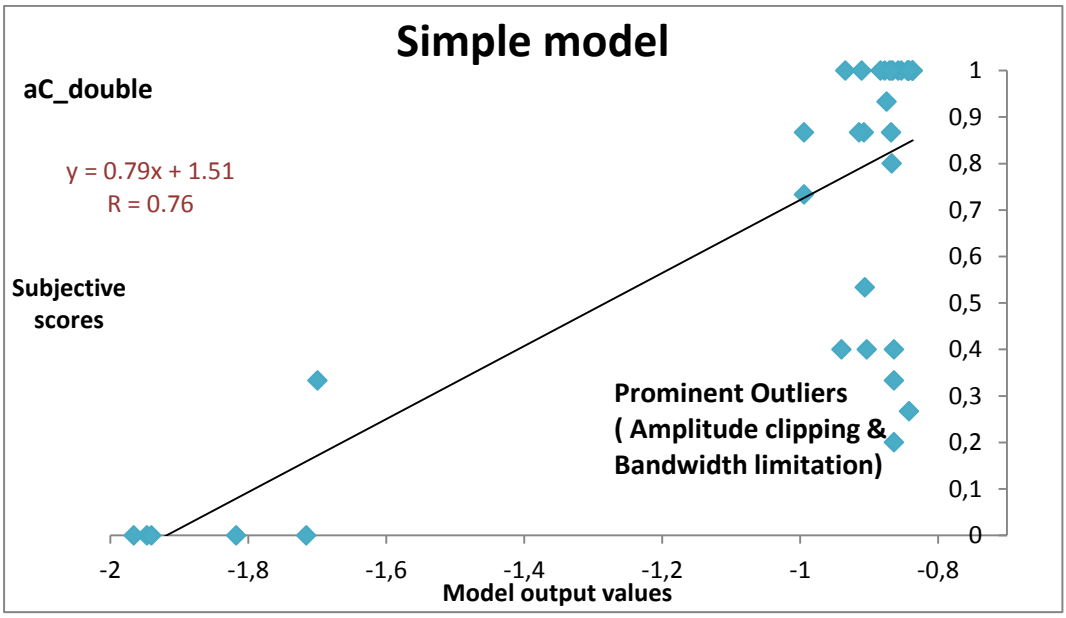

Figure 20. Correlation between the subjective scores and the model output values for '*aC*_double' database. This database does not contain 'aH' words. A total of 36 degraded '*a/C/*' words were present in the database. The degradations used were: background noise, bandwidth limitation, time-clipping, amplitude-clipping and codec distortions. A reasonably good correlation of 0.76 is obtained.

A further validation is performed on an already existing intelligibility database at TNO (Existing_database). This database contains degraded speech signals suffering from severe time clipping distortion. An excellent correlation of 0.98 is obtained between subjective scores and the model output values (see Figure 21).

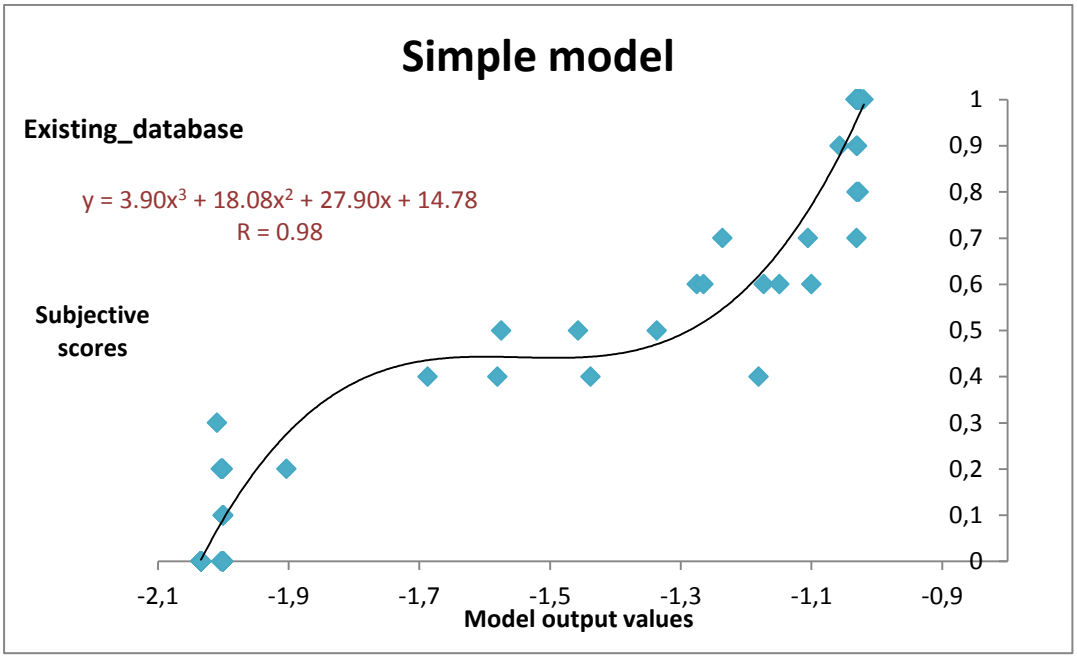

Figure 21. Correlation between the subjective scores and the model output values for an existing database at TNO. An excellent correlation of 0.98 is obtained. A total of 81 degraded speech signals suffering from severe time-clipping distortion were present in the database.

5. Conclusion and future work

This chapter provides the conclusions derived from the research work done in this thesis and presents some suggestions for future improvement.

Today's telecommunication networks are optimized to provide better voice services in terms of quality, while ignoring speech intelligibility. However, systems delivering high quality but unintelligible voice services are useless, as these systems cannot get the message across to the recipients. A cheap, fast and reliable objective speech intelligibility assessment technique is thus essential to allow operators and telecom equipment manufacturers to measure and then their systems to provide better speech intelligibility, ensuring an enhanced end user experience.

The research on this thesis was carried out to develop a simple perceptual objective model for measuring speech intelligibility of individual CVC words. ITU-T P.863 is used as a starting point to build a simple objective model for predicting subjective intelligibility scores of individual CVC words.

Some of the important conclusions and limitations of the research work carried out in this thesis are listed below:

- ITU-T P.863 cannot be used for predicting subjective intelligibility scores of individual CVC words. Although temporal structure of speech material was restructured to meet P.863 requirements, the correlation between subjective and standard P.863 output values was far too low (0.34).
- A simple model based on ratio of number of active frames and perceptual SNR provides good results on the first subjective database (correlation of 0.81), the second validation database (correlation of 0.76) and on an existing database at TNO which contains severely time-clipped degraded speech (correlation of 0.98). This indicates that a simple perceptual model can be developed for predicting individual CVC intelligibility.
- The prominent outliers in the first subjective database are some degraded speech signals suffering from time clipping distortion. The model predicts low scores for these outliers, while subjects find it intelligible resulting in high subjective scores. This behavior suggests that some higher level of modeling apart from the simple model used in this thesis is required. Nevertheless, the simple new model in general performs well under time-clipping distortion, which is clearly reflected by the results obtained from the existing database at

TNO that contains severely time-clipped degraded speech signals (correlation of 0.98).

- The prominent outliers in the validation database are some degraded speech signals suffering from severe bandwidth limitation and amplitude clipping. These degradations were not present in the first subjective database. This implies that other indicators besides Indicator1 and Indicator2 are required, for the model to be able to cope with these distortions.
- Removing the points reflecting the 'Bias effect' improved the correlation in the second validation database. A confusion matrix can be constructed for the first subjective database and the points reflecting the 'Bias effect' if present should be removed to improve correlation.

Further development and validation of the simple model should be done in the near future. Only male voices were used for recording speech material. A new subjective database containing both the male and female voices and new sets of degradations should be used for validating and developing the model further.

The simple new model developed in this thesis can be extended in future for assessing speech intelligibility for video calls. Video calls are being increasingly used in VoIP services today. It has been shown in the past that adding moving visual image of speaker's face can improve the intelligibility of the degraded speech signal [19]. The audio-visual speech intelligibility test done in [20] shows that visual cues like lip-motion, eye brow gestures, head nods etc., can increase the recognition rate. Improved results for speaker and speech recognition have been achieved by fusing audio and video signals (lip motion), confirming the importance of visual cues as a complimentary information to speech [21]. This means that in future these non-acoustic cues (visual cues) in addition to the acoustic cues should be taken into account when developing an objective model for assessing intelligibility for video calls.

Appendix

Appendix A: List of degradations used

'*/C/ous*' database:

Length of a reference file ~ 3 seconds

Length of a listening test file ~ 12 minutes

Degradation 1: Pink Noise

0 dB SNR

Degradation 2: Time clipping (varying)

Clip: 0.05 sec, 0.1 sec 0.15 sec and 0.2 sec

Degradation 3: Pulse (varying)

Pulse duration: 0.05 sec, 0.1 sec 0.15 sec and 0.2 sec

Degradation 4: Bandwidth limitation (varying)

Filters used: 200-4000 Hz, 200-2000 Hz (Windowing function: Blackman, FFT size: 8192)

Degradation 5: Reduced presentation level

-10 dB

'a*/C/*' database:

Length of a reference file ~ 2.6 seconds

Length of a listening test file ~ 12 minutes

Degradation 1: Pink Noise

0 dB SNR, -6 dB SNR

Degradation 2: Time clipping

Clip: 0.2 sec

Degradation 3: Amplitude clipping

Clip: -30 dB

Degradation 4: Bandwidth limitation

Filters used: 100-1000 Hz (Windowing function: Blackman, FFT size: 8192)

Degradation 5: Codec distortion

Codec used: G.729a (8 kHz), 10% loss

References

[1] John G. Beerends, Christian Schmidmer, Jens Berger, Matthias Obermann, Raphael Ullmann, Joachim Pomy and Michael Keyhl, "Perceptual Objective Listening Quality Assessment (POLQA), The third generation ITU-T standard for end-to-end speech quality measurement (Part I and II),", *J. Audio Eng. Soc. , submitted Dec. 2011*.

[2] John G. Beerends, "Extending P.863 'POLQA' towards intelligibility testing, first results," ITU-T study group 12 - contribution 300, May. 2012, Geneva, Switzerland.

[3] "POLQA application guide," HEAD acoustics GmbH, Germany, Mar. 2012.

[4] Recommendation ITU-T P.863, Perceptual objective listening quality assessment, SERIES P: TERMINALS AND SUBJECTIVE AND OBJECTIVE ASSESSMENT METHODS, Methods for objective and subjective assessment of speech quality, Jan. 2011.

[5] Dr.Irina Cotanis, "Voice Service Quality Evaluation Techniques and the New Technology, POLQA," ASCOM, 2010.

[6] Ottmar Gerlach, "Next-Generation (3G/4G) Voice Quality Testing with POLQA," Rhode & Schwarz, Mar. 2012.

[7] Dr.Irina Cotanis, "Moving from PESQ to POLQA, the Next-Generation Mobile Voice Quality Testing Standard," ASCOM.

[8] "An Introduction to ITU-T Rec. P.863 POLQA," Malden Electronics Ltd, UK, Sep. 2011.

[9] Beerends, J. G., Hekstra, A. P., Rix, A. W. and Hollier, M. P., "Perceptual Evaluation of Speech Quality (PESQ), the new ITU standard for end-to-end speech quality assessment, part II - psychoacoustic model," *Journal of the Audio Engineering Society,* Vol. 50, No.10, pp. 765-778, Oct. 2002.

[10] Antony W. Rix, John G. Beerends, Michael P. Hollier and Andries P. Hekstra, "Perceptual evaluation of speech quality (PESQ) – A new method for speech quality assessment of telephone networks and codecs," *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01),* vol.2, pp.749-752, 2001.

[11] N. R. French and J. C. Steinberg, "Factors Governing the Intelligibility of Speech Sounds, " J. Acoust. Soc. Am., Vol. 19, Issue 1, pp. 90-119, 1947.

[12] Herman J.M., and Tammo Houtgast., "Basics of the STI measuring method," *Past, present and future of the Speech Transmission Index. Soesterberg, the Netherlands: TNO Human Factors*, pp. 13-43, 2002.

[13] Herman J.M. Steeneken, "THE MEASUREMENT OF SPEECH INTELLIGIBILITY," TNO Human Factors, Soesterberg, the Netherlands, Retrieved from: http://www.gold-line.com/pdf/articles/p_measure_TNO.pdf.

[14]. C. Ludvigsen, C. Elberling, and G. Keidser, “Evaluation of noise reduction method: Comparison between observed scores and scores predicted from STI,” Scan. Audiology, vol. 39, pp. 50–55, 1993.

[15] "Speech Intelligibility, technical notes volume 1," A JBL professional technical note, number 26.

[16] John G. Beerends , Ronald Van Buuren, Jeroen Van Vugt and Jan Verhave, “Objective Speech Intelligibility Measurement on the Basis of Natural Speech in Combination with Perceptual Modeling,” *J. Audio Eng. Soc. , Vol.57, No. 5, May. 2009* .

[17] Jianfen Ma, Yi Hu and Philipos C. Loizou, “Objective Measures for Predicting Speech Intelligibility in Noisy Conditions Based on New Band-Importance Functions,” J. Acoust. Soc. Am. Vol. 125, No.5, May. 2009.

[18] Yi Hu and Philipos C. Loizou, "Evaluation of Objective Quality Measures for Speech Enhancement," IEEE transactions on audio, speech, and language processing, vol. 16, no. 1, pp. 229-238, Jan. 2008.

[19] Olov Ostberg, Bjorn Lindstrom and Per-Olof Renhall, “Contribution of display size to speech intelligibility in videophone systems,” *Int. J. Human-comput. Interact. , Vol.1, No.2, pp. 149-159, 1989*.

[20] Samer Al Moubayed and Jonas Beskov, “Effects of Visual Prominence Cues on Speech Intelligibility”, Auditory-Visual Speech Processing Proceedings AVSP’09, 2009, Vol.9, Norwich, England.

[21] Maycel-Issac Faraj and Josef Bigun "Synergy of Lip-Motion and Acoustic Features in Biometric Speech and Speaker Recognition," IEEE transactions on computers, vol. 56, no. 9, pp. 1169-1175, Sep. 2007.