Objective assessment of speech intelligibility.

Liu, Wei Ming

Objective Assessment of Speech Intelligibility

by

Wei Ming LIU

A thesis submitted to the
UNIVERSITY OF WALES
in fulfilment of the requirements
for the Degree of
DOCTOR OF PHILOSOPHY

SWANSEA UNIVERSITY

Year 2007

Declaration/Statements

Declaration

This work has not previously been accepted in substance for any degree and is not being concurrently submitted in canditature for any degree.

Signed ..... (candidate)

Date ..... 16/05/08 .....

Statement 1

This thesis is the result of my own investigations, except where otherwise stated. Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signed ..... (candidate)

Date ..... 16/05/08 .....

Statement 2

I hearby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed ..... (candidate)

Date ..... 16/05/08 .....

Dedication

This thesis is dedicated to my grandpa, Liu Nge, whose loving eyes never fail to bring out a new surge of energy in me to strive on, no matter how huge the challenges and how painful the setbacks were during this difficult research journey. Also, this thesis is dedicated to Swansea Chinese Christian Church (SCCC) for their unconditional love and support throughout the course of this research.

Acknowledgements

I would first like to express my deepest gratitude to my supervisor, Dr. J.S.D Mason, for giving me the opportunity to undertake this research at Swansea University. His guidance and time invested throughout the course of this research are deeply appreciated and are undeniably keys to my success. Also I owe a sincere debt to Dr. Nick Evans, who has been selflessly giving his time whenever I needed help, even after leaving Swansea to pursue his career in France. His valuable advice, as well as the effort put in proof reading this thesis, are deeply remembered.

Additionally I would like to thank all members in the Speech and Image Research Group, namely Benoit, Keith, Alicia, Neil, Richard and Ruben, for the assistance and friendship and fun throughout the years. My research life would not be the same without them.

A sincere love and gratitude to the brothers and sisters in Swansea Chinese Christian Church (SCCC), particularly May Jie, Brenda, Michelle, Ken, Alvin, Chong, Jade, Samson, Amy, Stephanie and dearest darling David Loke. Their patience and encouragement during the time of disappointment, despair and frustration set my feet on the ground and enabled me to strive on. I realise how extremely blessed I am to have such a supportive family in Christ. Also, a big thank you to my fellow friends See Ju, Molly, JiaJen, SoonAo, Chris, Jann, Mable and Maylynn for their encouragement and help in many areas.

Next, I would like to thank Overseas Research Student (ORS) Scholarship, as well as Her Majesty Government Communication Centre (HMGCC) for their financial support. In particular I would also like to extend gratitude to Mr K. Worrall, the leading acoustician in HMGCC whom also is the initiator of this research. Thanks must also go to Mr R. Fellows and Miss L. Craigie for the many insightful thoughts we shared and for coming all the way from London for the meetings.

Needless to say, none of this would have been possible without the constant support of my family back in Malaysia whom I miss dearly especially my 85-year old grandparents. I would not have got this far without their unlimited confidence in me. They trust that I would persevere and succeed without a shadow of doubt. They are the motivation for me to work hard and never give up.

Most importantly, I want to thank my Heavenly Father for His everlasting love and comfort throughout the challenging years during research, for I know that I never walk alone, for 'Love never gives up, never loses faith, is always hopeful, and endures through every circumstance *[1 Corinthian 13:7]*'.

Abstract

This thesis addresses the topic of objective speech intelligibility assessment. Speech intelligibility is becoming an important issue due most possibly to the rapid growth in digital communication systems in recent decades; as well as the increasing demand for security-based applications where intelligibility, rather than the overall quality, is the priority. Afterall, the loss of intelligibility means that communication does not exist.

This research sets out to investigate the potential of automatic speech recognition (ASR) in intelligibility assessment, the motivation being the obvious link between word recognition and intelligibility. As a pre-cursor, quality measures are first considered since intelligibility is an attribute encompassed in overall quality. Here, 9 prominent quality measures including the state-of-the-art Perceptual Evaluation of Speech Quality (PESQ) are assessed. A large range of degradations are considered including additive noise and those introduced by coding and enhancement schemes. Experimental results show that apart from Weighted Spectral Slope (WSS), generally the quality scores from all other quality measures considered here correlate poorly with intelligibility. Poor correlations are observed especially when dealing with speech-like noises and degradations introduced by enhancement processes.

ASR is then considered where various word recognition statistics, namely word accuracy, percentage correct, deletion, substitution and insertion are assessed as potential intelligibility measure. One critical contribution is the observation that there are links between different ASR statistics and different forms of degradation. Such links enable suitable statistics to be chosen for intelligibility assessment in different applications. In overall word accuracy from an ASR system trained on clean signals has the highest correlation with intelligibility. However, as is the case with quality measures, none of the ASR scores correlate well in the context of enhancement schemes since such processes are known to improve machine-based scores without necessarily improving intelligibility. This demonstrates the limitation of ASR in intelligibility assessment.

As an extension to word modelling in ASR, one major contribution of this work relates to the novel use of a data-driven (DD) classifier in this context. The classifier is trained on intelligibility information and its output scores relate directly to intelligibility rather than indirectly through quality or ASR scores as in earlier attempts. A critical obstacle with the development of such a DD classifier is establishing the large amount of ground truth necessary for training. This leads to the next significant contribution, namely the proposal of a convenient strategy to generate potentially unlimited amounts of synthetic ground truth based on a well-supported hypothesis that speech processings rarely improve intelligibility.

Subsequent contributions include the search for good features that could enhance classification accuracy. Scores given by quality measures and ASR are indicative of intelligibility hence could serve as potential features for the data-driven intelligibility classifier. Both are in investigated in this research and results show ASR-based features to be superior. A final contribution is a novel feature set based on the concept of anchor models where each anchor represents a chosen degradation. Signal intelligibility is characterised by the similarity between the degradation under test and a cohort of degradation anchors. The anchoring feature set leads to an average classification accuracy of 88% with synthetic ground truth and 82% with human ground truth evaluation sets. The latter compares favourably with 69% achieved by WSS (the best quality measure) and 68% by word accuracy from a clean-trained ASR (the best ASR-based measure) which are assessed on identical test sets.

Abbreviations

AI	Articulation Index
ASR	Automatic Speech Recognition
CSNR	Classical Signal-to-Noise Ratio
DRT	Diagnostic Rhyme Test
ETSI	European Telecommunications Standard Institute
GMM	Gaussian Mixture Models
HMM	Hidden Markov Models
HSR	Human Speech Recognition
IS	Itakura Saito
ITU	International Telecommunication Standard Institute
LAR	Log Area Ratio
LLR	Log likelihood ratio
MBSD	Modified Bark Spectral Distortion
MNB	Measuring Normalizing blocks
MOS	Mean of Opinion Score
MRT	Modified Rhyme Test
NLSS	Non-linear Spectral Subtraction
PESQ	Perceptual Evaluation of Speech Quality
SegSNR	Segmental Signal-to-Noise Ratio
SNR	Signal-to-Noise Ratio
STI	Speech Transmission Index
WSS	Weighted Spectral Slope

Contents

List of Figures

List of Tables

Chapter 1

Introduction

S. F. Boll, one of the pioneers of spectral subtraction who has contributed significantly to speech enhancement since the late 1970s, made a rather intriguing statement in 1991 [1]. To quote directly, he stated: "Why has no one found a way to remove noise from speech in order to improve intelligibility?". The statement not only highlights the difficulty of improving intelligibility, but also the difference between quality and intelligibility. More importantly as far as this research is concerned, the statement exposes the paradox that under many circumstances it is relatively easy to process speech to obtain improved machine-based scores such as word recognition for an automatic speech recognition (ASR) system without actually improving intelligibility [2,3]. This hints one of the big challenges in machine-based intelligibility assessment.

Reliable assessment of intelligibility is critical due to its increasing importance. Intelligibility is the essence without which communication does not exist. In many applications where conveyance of information is the main concern intelligibility is no doubt given priority over other criteria such as ease-of-listening, naturalness and so on. This is especially true for security-based applications such as military which may operate under adverse noise conditions and bandwidth constraints. The threat of terrorism and increasing social disturbances might also play a part in the demand for systems (e.g., monitoring systems) which maintains high levels of intelligibility. Apart from that, the escalating boom in the mobile communications industry in recent years is also believed to have contributed to increasing awareness of the importance of intelligibility. This new trend of communication brings to the users an 'anywhere and anytime' expectation, as coined by Martin Rainer [4]. However, along with convenience of mobility brings exposure to uncontrollable, inevitable and often hostile degradations, for example, street noise, subway noise, crowd, rain, etc. In summary, speech with degraded intelligibility seems to be an ever-growing problem.

Ultimately intelligibility is defined by humans. Various standards are available as guidelines for conducting intelligibility listening tests including the Diagnostic Rhyme Test (DRT), Modified Rhyme Test (MRT) and Phonetically Balanced (PB) sentences. These tests have specific requirements on details such as vocabulary, database, number and condition of listeners, procedures, equipments, room condition, etc in order to improve consistency and reliability of the ground truth obtained. However, such listening tests are often too costly both in terms of time and money. There is hence a need for alternative, machine-based assessment approaches to replace or at least supplement the human listening tests. Machine-based approaches are commonly referred to as the objective measures, while human listening tests as the subjective measures.

Despite the importance of intelligibility, ironically objective assessment of intelligibility has not received great attention in the open literature relative to the more general overall quality. This is evident from the consistent development in objective quality measures from the simple time-domain measures to the more advanced spectral domain measures (1970s and 1980s), to the more recent psychoacoustic domain measures (1990s to present) Quackenbush et al [5] reported an investigation of over 2000 variations of waveform-based and spectral-based quality measures in 1988. Several prominent quality measures of that era include signal-to-noise measure (SNR), the Itakura-Saito (IS) distance, the log area ratio (LAR), the weighted spectral slope (WSS) and the cepstral distance (CD). As technology has advanced and new forms of degradation are introduced by modern speech processing systems, more recent developments follow a perceptual-based approach. Explicit models for some of the known attributes of human auditory perception are incorporated into the quality assessors with the motivation to create assessors that better mimic the human hearing systems. Well-known perceptual-based measures include bark spectral distortion measure (BSD) [6], measuring normalising blocks (MNB) [7], Perceptual Evaluation of Speech Quality (PESQ) [8] and Single-sided (i.e., non-intrusive) Speech Quality Measure (3SQM) [9]. The last two measures, namely PESQ and 3SQM are ITU-T standardised and are generally deemed as the state-of-the-art for quality measurement.

While significant research efforts have been directed to the area of objective assessment for overall quality, development in the more specific case of intelligibility assessment is relatively small. Early attempts date back to 1947 when Bell Labs developed the articulation index (AI) [10]. The speech transmission index (STI) introduced by Houtgast and Steeneken [11] in 1973 is a variation based on AI and is included in IEC standard 60268-16. Both AI and STI are reported to correlate well with human intelligibility but their applicability is rather limited to linear systems, rendering the measures less suited to modern applications such as testing with vocoders [12, 13]. In fact, the STI standard clearly points out the exclusion of vocoder applications from its scope of intelligibility measurement. Besides, the STI measure is primarily intended for natural room acoustic transmission (cathedrals, auditoriums or classrooms). Subsequent developments evolved mainly around enhancement or simplification of the STI, resulting in measures such as STI for public announcement systems (STIPA), rapid STI (RASTI), and STI for communication systems (STICOM). To quote directly from Hu and Louzou [2] whose works in speech enhancements require an intelligibility measure to evaluate a range of enhancement algorithms in terms of speech intelligibility: "Given the absence of accurate and reliable objective measure to predict the intelligibility of speech processed by enhancement algorithms, we must resort to formal listening tests...". This work of Hu and Louzou was published in May, 2007. It is probably safe to say to date there is no PESQ-equivalent in the area of intelligibility assessment.

The lack of reliable objective intelligibility measures is somewhat evident when out of 10 contributions in the recent speech intelligibility conference organised by Institute of Acoustics in London (Sept, 2006), 5 describe subjective testing performed in their respective sectors, 4 report objective tests involving STI and 1 reviews errors of objective intelligibility measures particularly STI due to its dominance in the field. Mercy and Aitchison [14] review the requirements for speech intelligibility

in communications systems which is set out by the UK MoD in DEFSTAN 00-25 Part 16. According to [14], it was noted in the MoD document that available objective methods are not sufficiently precise for assessment of military communications; and that certain assumptions about the nature of speech implicit in the AI and STI do not apply to vocoded speech and synthetic speech. Mercy and Aitchison [14] also quote the standards set out by the US Department of Defence stating that AI and STI should be used to estimate system performance only during the concept and design stage but not as a substitute for intelligibility testing when a production system is available. The contributors of the 4 articles reporting on subjective tests in the said London conference include military [15] and police force [16]. It seems that in sectors where intelligibility testing is critical, for example security services, human listening tests are sought and research efforts are instead invested into refining human listening tests.

Reasons for the relative inactiveness of development in objective intelligibility assessment is perhaps attributed to the difficulty of the task itself. Conceptually humans are better assessors of speech intelligibility because humans are also better at hearing speech (i.e., better speech recognisers). As much as perceptual-based quality measures such as PESQ are built on rules or knowledge of how humans perceive or assess quality, if the rules of how humans assess intelligibility are known then those rules can be modelled to build an objective intelligibility measure. Unfortunately, both Lippmann [17] and Scharenborg [18] who are involved in research into automatic speech recognition (ASR) versus human speech recognition (HSR) observe that many details of the internal processes of how humans recognise speech are still unknown. This lack of knowledge inevitably impairs the development of objective intelligibility measures.

In conclusion, objective assessment of speech intelligibility has proved to be a challenging task. Existing objective intelligibility measures have either fallen out of use (for e.g., AI [19]) or are incompatible with modern degradations [12, 13] or are designed for room acoustics; meanwhile quality measures are most likely to be limited in their potential usefulness for intelligibility assessment, one reason being that it could be easy to artificially improve quality scores without actually improving intelligibility. In short, there is a lack of development in objective intelligibility assessment but a real and perhaps pressing need for such technology due to the importance of some specific applications (for e.g. security services) that require intelligibility testing.

1.1 Objectives

The primary objective of the research reported in this thesis relates to the development of an objective (machine-based) assessment system that gives scores reflecting human opinion of intelligibility. The context is illustrated in Figure 1.1 where an original signal, S is processed independently by two different processes resulting in output speech signals S_A and S_B. The objective of this research is to develop an intelligibility measure to identify the process giving the more intelligible output.

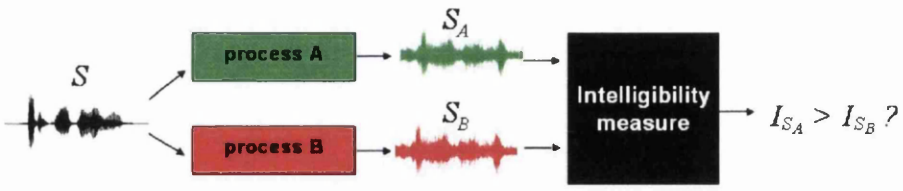


Figure 1.1: The intelligibility assessment system is to predict which of the signals S_A and S_B is more intelligible.

The two different processes could refer to competing systems or the same system with different configurations to be optimised during developmental stages. They could also refer to communication links consisting of different systems combinations or various operating environments which introduce different environmental degradations. In order for the system to be practicable for general use, it should be capable of assessing intelligibility of speech signals influenced by any combination of factors commonly found to affect intelligibility. Here we consider specifically environmental noises and degradations introduced by speech coding and enhancement algorithms.

The difficulty of the task is three-fold. Let I_{S_A} and I_{S_B} be the intelligibility of S_A and S_B in Figure 1.1. In order of ascending difficulty the task can be defined as:

(i) to identify the which signal is the more intelligible, i.e., is I_{S_A} bigger than I_{S_B}?

(ii) to measure the intelligibility *difference* between the signals, i.e., by how much is I_{S_A} larger or smaller than I_{S_B}?

(iii) to grade the absolute level of intelligibility of each signal under comparison, i.e. what are I_{S_A}? what is I_{S_B}? In other words, what is the intelligibility of S_A and S_B?

The ideal solution is to achieve task (iii) since clearly achieving (iii) also gives (i) and (ii). However, given that assessment of intelligibility is extremely complex [19] this research embarks on tackling task (i) with the hope of providing a foundation for tackling tasks (ii) and (iii). In short, the aim of the work described in this thesis is to develop an objective intelligibility assessment system that compares/ranks two or more signals in terms of their intelligibility levels. The output of the system reflects relative intelligibility for a given pair of signals, i.e., which is the more intelligible?. The usefulness of the system is judged based on the correlation between the system's output scores and intelligibility scores derived from human for the same set of signals. Note that correlation is on trend rather than absolute scores. For the works described in the thesis, the 2 signals under comparison originate from 1 signal processed in 2 different ways hence are of the same utterance, speaker, length, etc.

1.2 Claims and Contributions

This research attempts to provide a solution for objective assessment of comparative intelligibility. This section summarises the claims and contributions of this research.

Firstly, considering that intelligibility is an attribute of overall speech quality [20] and given the wealth of accomplishment in the field, this raises the question of how useful quality measures might be in the context of intelligibility assessment. Though excellent correlation between objective quality scores with human opinions of intelligibility is not anticipated but such potential of the quality measures has rarely been investigated especially on large range of degradations. Hence one contribution of the research is to

- investigate the suitability of objective quality measures for intelligibility assessment. Nine prominent measures which encompasses time, spectral and perceptual domain are considered.

Degradations considered consist of environmental noises, those introduced by speech coding and enhancement schemes, and their combinations. Intelligibility ground truth for these test sets are established by conducting human listening tests. Performances of the quality measures are judged based on correlation between the objective quality scores and intelligibility ground truth provided by humans.

Secondly, speech intelligibility is thought to be closely related to recognition of words [21]. Investigations into potential of automatic speech recognition (ASR) systems for intelligibility assessment are not new but neither common. Little work has followed that of Chernick [22], Jiang [23] and Hicks et al [24] who all reported promising results on the potential of ASR in such context. Nonetheless, the degradation conditions considered are rather specific, for example, packet loss by Jiang, bit errors by Chernick and co-channel noise by Hicks. This thesis claims to be the first to

- investigate the potential of ASR in intelligibility assessment using a wider range of degradations allowing side-by-side analysis.
- investigate the potential of various ASR statistics, namely word accuracy, percentage correct, deletion, insertion and substitution.
- configure ASR for intelligibility assessment through different training.
- investigate the potential benefit of missing data techniques in conjunctions with ASR in the context of intelligibility assessment. The motivation being that the missing data techniques mimics humans' ability to perform 'auditory scene analysis' (pick out and pay selective attention to reliable sound source) hence might lead to better correlation.

Experiment findings show that a simple clean-trained ASR system is already able to outperform the quality measures in correlating with human intelligibility especially in the context of environmental noises and speech coding degradations. Comparing quality measures and ASR, the former approach is rule-driven and is indirectly linked to intelligibility through overall quality; the latter approach is data-driven and is linked to intelligibility through to the more relevant word recognition. Instead of modelling words as in ASR, one step closer to achieving the research objective would be direct modelling of intelligibility difference. Other motivations to use a data-driven approach are successful applications of the approach in many equally high-end speech-related classification tasks including speaker recognition, speech synthesis, and even recently speech quality assessment [25]. It is claimed that the work in this thesis is the first to

- apply a data-driven classification approach to intelligibility assessment. The product is a classifier trained on intelligibility differences computed from signals in pairs ; the output scores of the classifier hence relate directly to comparative intelligibility for each signal pair.

The reason that a data-driven approach has not been adopted for objective intelligibility assessment in the past is perhaps due to the lack of labelled data needed to train the classifier. In this case the training data are signals with ground truth regarding relative levels of intelligibility (i.e., which signal in the pair is the more intelligible?). One main contribution of this thesis is overcoming this constraint by

- proposing a method that could provide large, potentially unlimited amounts of diversely degraded signals with known intelligibility relationships, while requiring minimal human effort.

An interesting parameter to study is 'where does the minus sign go?' referring to the stage where signal differencing is performed. Assessment using the quality measures and ASR deduces comparative intelligibility for signal pairs through score-level differencing where intelligibility for each signal is first independently and explicitly estimated. Here a more direct approach is taken where

- feature-level differencing is performed where features for the classifier is a vector of differences between features from the two comparing signals. This is thought to be more targetted towards assessment of comparative intelligibility.

Whilst most other speech-related classification tasks employ low-level features such as short-term spectral estimates, this research investigates features stemming from other levels of the signals, namely utterance, word and frame level. Some contributions in this area are

- a systematic search for potential features for the classifier using scores stemming from the quality measures and word recognition process in ASR as well as low-level features.

- introducing a novel feature based on the concept of anchor models where intelligibility is characterised by the similarity between the degradation it underwent and a cohort of chosen anchoring degradations. The feature set addresses 2 problems associated with objective intelligibility assessment.

For any data-driven system better performance can be expected when the representativeness of the training data improves. Here this nature can be fully exploited given the convenience to generate training data (using the data generation method proposed earlier) by involving only relevant degradations (i.e., degradations likely to be encountered during testing) and by including relevant anchor models in the feature generation process. This classifier proposed in the research provides the foundation and flexibility for the development of such application-specific classifier. This is discussed as future work.

1.3 Thesis Structures

The investigatory chapters of this thesis is divided into 2 parts: Part I (chapters 4 and 5) focuses on the investigation of existing measures including the quality measures and ASR; Part II (chapters 6 to 10) extends the work towards a data-driven approach where direct modelling of intelligibility difference is introduced and the comparative intelligibility classifier is proposed.

Chapter 2 reviews the nature of the problem and identifies some foreseeable practical difficulties of objective intelligibility assessment.

Chapter 3 describes the listening tests conducted in order to collect human opinions of intelligibility as ground truth for evaluation of quality measures in Chapter 4 as well as ASR in chapter 5. The standard Aurora2 digit-string corpus [26] is used. Though not specially designed for intelligibility assessment, the digit database is chosen due to its simplicity and the desire to investigate ASR as an intelligibility assessor in Chapter 5. Implementation of a web-based listening test facility enables collection of responses from large number of listeners. Three categories of degradations are considered which give rise to six test sets. The degradations are realistic, daily-encountered degradations which include additive environmental noises, as well as degradation introduced by standard speech coding and enhancement schemes.

Chapter 4 investigates the potential of nine quality measures in the context of intelligibility assessment. They are CSNR, SegSNR, IS, LAR, LLR, WSS, MNB, MBSD, and PESQ. Objective scores are obtained for the six test sets introduced in Chapter 3. Performances of the measures are evaluated based on correlation between the objective scores and human opinion of intelligibility, the ground truth. Pearson correlation and Kendall tau distance (a form of Kendall tau rank correlation) are employed for the correlation analysis.

Chapter 5 considers the potential of ASRs for intelligibility assessment. Various training schemes are investigated. The same experimental setup as in Chapter 4 is used, namely the same test sets and correlation methods.

In Part II the work is extended towards a data-driven approach. Chapter 6 presents the concept of direct modelling of intelligibility information, followed by the proposal of a semi-automatic method which provides large amounts of training data needed for the modelling. Lastly a differential classifier trained on intelligibility difference is proposed where pairs of signals under comparison are differenced at the feature level.

Chapter 7 presents experimental framework where the training and test sets for both the development and evaluation stage of the classifier are introduced. Procedures involved to generate the data sets are also described. Benchmark experiment is reported.

Chapter 8 presents a systematic search for potential features for the classifier. Features considered include scores stemming from various level of processing in quality measures as well as word level recognition in ASR. Low-level frame-based holistic features are also considered where the GMM supervector approach is employed to reduce dimensionality.

Chapter 9 proposes a novel feature set based on the concept of anchor models used in speaker indexing or verification. A selection of degradations are chosen and differently trained ASR systems function as anchor models. Signal intelligibility is characterised by ASR scores given by the cohort of anchor models.

Chapter 10 evaluates the classifier using the six human-evaluated test sets introduced in Chapter 3. The best performing feature sets as identified in Chapters 8 and 9 are assessed here.

Finally, conclusions detailing the contributions of this thesis and some final thoughts are presented in Chapter 11 together with ideas for future work.

Chapter 2

The Nature of the Problem

Assessing speech intelligibility is an extremely challenging task [19]. This chapter reviews some of the practical difficulties associated with objective assessment of intelligibility.

2.1 Small Dynamic Range

Intelligibility is commonly regarded as an attribute of overall speech quality alongside other attributes such as naturalness, ease of listening and loudness [27]. Considering that quality encompasses many criteria while intelligibility is just a single specific attribute, should measuring quality not be more challenging than measuring intelligibility? Besides, if quality could successfully be measured objectively with state-of-the-art measures such as Perceptual Evaluation of Speech Quality (PESQ) [8], how difficult can it be to measure intelligibility? Answers to these questions are not readily available but Hansen [28] made a good comparison of the two tasks by saying that: "Ordinarily, unintelligible speech would not be judged to be high quality; however, the converse need not be true." This statement does not imply that one is easier than the other, but it definitely implies that the two tasks have very different operational ranges, with the first half of Hansen's statement referring to the region where the two ranges overlap. It is postulated here that with the exception of intelligibility, other aspects of overall quality, namely ease of listening, loudness and naturalness have more of a linear relationship with overall quality. The relationship between quality and intelligibility is thought to exhibit a trend shown in Figure 2.1 where high intelligibility corresponds to not only high quality but also a large range of lower quality. The horizontal axis refers to quality range which can be thought of as corresponding to the Mean Opinion Score (MOS) score of poor to excellent. A large portion of this range (as indicated by the green arrow) corresponds to 100% intelligibility. On the other hand, when intelligibility is at threshold between 0% and 99%, all other aspects of quality are probably swamped and the notion of overall quality is long gone.

Certainly therefore, if quality measures are to be applied for intelligibility assessment, then quality scores would be constrained to a relatively small section of their dynamic range. A large part of the meaningful quality score range would correspond 100% intelligibility, hence in this range, intelligibility is essentially not measured. This is illustrated in Figure 2.2 where the red profile refers to human intelligibility response and blue profile refers to intelligibility scores estimated by a objective (or machine-based) quality measure. Intelligibility as perceived by humans falls from the notional

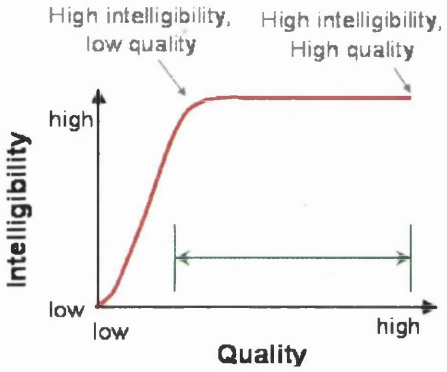


Figure 2.1: An illustration of the hypothesis of the relationship between intelligibility and overall quality, where high intelligibility not only corresponds to high quality but also to a large range of lower quality.

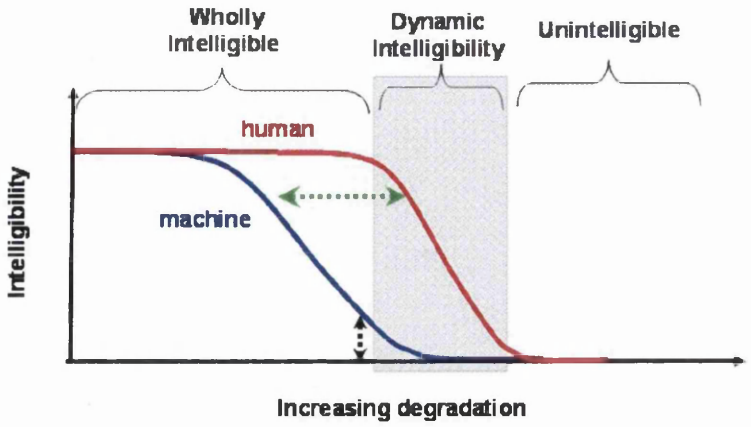


Figure 2.2: An illustration of the huge offset (horizontal arrow) between human scores and machine scores. Grey shaded area refers to dynamic intelligibility region where assessment is an issue. Vertical arrow indicates that machine scores are on the verge of saturation in this region.

100% to 0% within a region referred to here as the 'dynamic intelligibility region' ('dynamic' since intelligibility changes more rapidly in this region compared to the 'wholly intelligible' and 'unintelligible' region where intelligibility is notionally constant, see Figure 2.2). Notice that at the region of dynamic intelligibility where intelligibility assessment is often needed, objective scores are nearly saturated as indicated by the short vertical arrow. Meanwhile, the horizontal arrow emphasises the notable offset between the two profiles. Figure 2.3 shows an example using actual data considered in Chapter 3 and 4 where the blue profile shows scores estimated by PESQ for signals degraded by car noise over the SNR range of 70dB to -20dB. The red profile shows intelligibility as perceived by humans for the same signals. All scores are normalised to 0% and 100% (PESQ score is mapped from its original scale of -0.5:4.5 to 0%:100% here). First of all, notice that span of dynamic intelligibility region of the human profile is relatively narrow compared to the profile of PESQ, as indicated by the black and pink arrows below the x-axis; above and below the black arrow (i.e., before 5dB and after -15dB on the SNR axis) signal intelligibility as perceived by humans is either constantly 100% or 0%. Notice that at 5dB while the human scores is still at 100%, the corresponding PESQ score is only about 45%. Another interesting observation is that PESQ scores saturates not at 0% but at around 33%, corresponding to approximately 1.2 out of the PESQ original scale of -0.5 to 4.5. The reason for this is unknown but most probably because the measure is not intended to measure signals at such low quality. Therefore, while human perceived intelligibility moves from 100% to 0% in the dynamic intelligibility region, PESQ scores obtained are only in the narrow range of approximately 45% to 33%. Given that the fundamental task of the research is to determine the order/ranking of intelligibility of two signals S_A and S_B, this will clearly be more difficult when the dynamic range is small since there

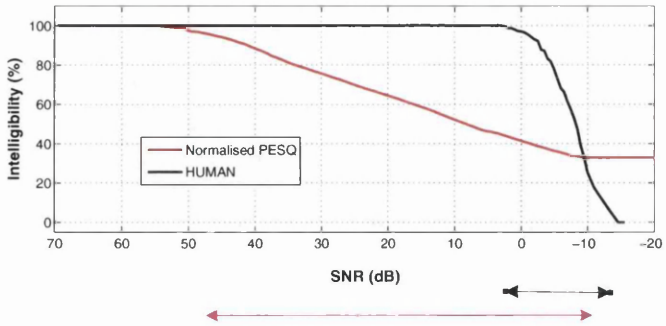
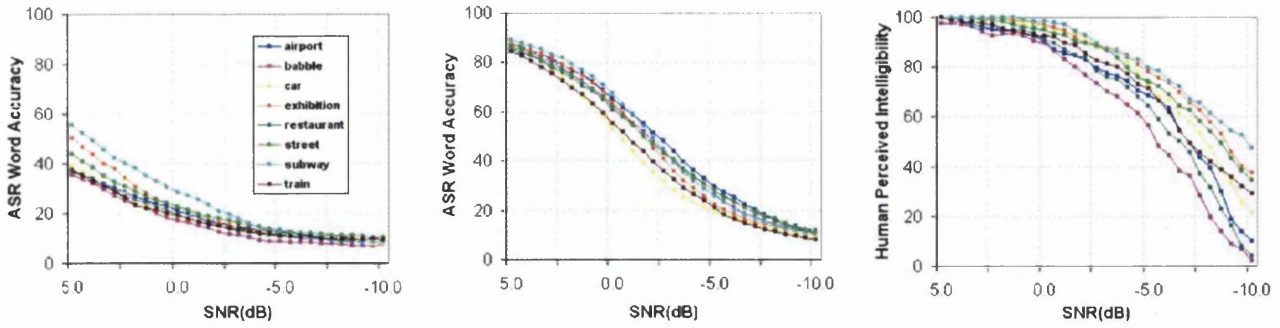


Figure 2.3: An illustration of huge offset in operational ranges of quality and intelligibility. Profiles show PESQ score for signals degraded by car noise versus intelligibility as perceived by human for the same signals. Two distinctive differences: (i) dynamic intelligibility region of human profile begins at lower SNRs (approx. 5dB) than that of PESQ profile (approx. 50dB); (ii) dynamic intelligibility region of human profile is narrower (indicated by arrows).

is greater tendency towards occurrence of unintended crossovers.

Arguably the phenomenons/characteristics illustrated in Figures 2.1 to 2.2 exist not just when quality measures are applied, but also for other objective measures such as ASR that are often used to assess speech in severely degraded conditions. Another example is the so-called gap between ASR and HSR (human speech recognition), with machine performance still very much below that to of humans. Lippmann [17] compared the recognition accuracy for machines and people with six recognition corporas that represent many different potential applications of speech recognition technology. It is important to note that machine recognisers used in the evaluation are the ones that had been tuned to provide the best performances for each corpus. Yet it was found that machine word recognition error rates are often more than an order of magnitude lower than those of humans in both quiet and degraded environments, for example, 4% error for humans and 40% for machines. Superiority of humans becomes even more obvious in noisier conditions. Lippmann also quotes findings from relevant studies by other researchers. Among many are Ebel and Picone [29] who reported that even when noise compensation algorithms aiming to improve machine performance is applied, machine error rates are still much higher than those of humans. For instance, at 10dB, error rates of an HMM recogniser with noise compensation for Wall Street Journal sentences with additive automobile noise is reported to be 12.8%, which is more than 10 times higher than human error rates of 1.1%. A separate study by Varga and Steeneken [30] reports that at 0dB where human listeners give an error rate of less than 1%, machine error rates are around 40% even with noise compensation. Scharenborg [18] reports similar observations that humans are much better able to recognise speech in adverse conditions and non-stationary noises. A 'helpless' fact both Scharenborg and Lippmann [17] pointed out is that the reasons behind humans' superiority in recognising speech is so far not fully comprehended. This evidence regarding ASR seems to suggest that the problem of limited dynamic range when operating



(a) (b) (c)

Figure 2.4: An illustration that the problem of limited dynamic range may apply not only to quality measures but to all machine-based assessors in general. Figures show ASR word accuracy plotted against SNRs for ASR trained on (a) clean signals and (b) mixtures of clean and degraded signals. (c) corresponding human scores.

under high degradation may apply to machine-based assessors in general.

Figure 2.4 show some examples using real data taken from Chapters 3 and 5. Figure 2.4(a) shows recognition performance when signals degraded by different background noises are tested against a clean-trained ASR. Figure 2.4(b) shows the same except that the ASR is trained on the Aurora 2 multi-condition training set which consists of signals degraded by various background noises over the SNR range of 20dB down to 5dB. Figure 2.4(c) shows the corresponding human performance. Notice the relatively small dynamic range and profiles overlap in Figure 2.4(a) as compared to human performance in Figure 2.4(c). Of course, it is the correlation that is important and as long as the profiles are ranked as according to human opinion, then the scores are useful in that they reflect human trends. But the fact that the profiles are 'crammed' together in practice makes such patterns far less likely. Figure 2.4(b) illustrates that with data-driven systems such as ASR it is possible to configure the training to increase its scores and the dynamic range, however, it is unsure whether this increment of the dynamic range also increases the correlation between machine and human scores. Notice that the order of the profiles in Figure 2.4(b) are now very different to those in Figure 2.4(a), for example, the babble profile (pink) is at the bottom in Figure 2.4(a) but becomes one of the top profiles in Figure 2.4(b).

In conclusion, intelligibility assessment is essentially most needed in the dynamic intelligibility region (see Figure 2.2) which is normally a region of high degradation, for example, at SNR below 0dB. This means that intelligibility assessment invariably (and inevitably) operates under high degradation where machine-derived scores are on the verge of saturation, resulting in limited dynamic range and potentially lacking correlation with human scores (in terms of profile ranking). This makes the task more particular than other assessment tasks which do not always have to operate under high degradation.

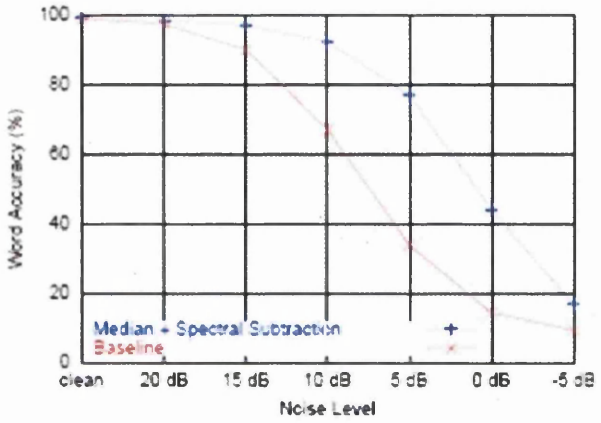


Figure 2.5: An illustration that noise compensation process can readily improve machine scores eventhough humans' perceived intelligibility are likely to remain the same or even degrade. The figure shows ASR performance for car-noise degraded signals with and without non-linear spectral subtraction. The blue profile shows significant improvement over the red profile demonstrating the effectiveness of the process in improving ASR scores. However, human listening tests testify that there is no improvement in intelligibility.

2.2 Intelligibility Enhancement

In 1991, S.F.Boll [1] posed this question: "Why has no one found a way to eliminate noise from speech in order to improve intelligibility?". Ironically Boll is one of the few pioneers in speech enhancement contributing significantly to the growth of spectral subtraction since the late 1970s. For such a statement to be made by someone who has been so actively involved for more than a decade in the research of reducing noise from degraded speech, there must be some truth in the statement. Yet at the same time there is a great wealth of literature reporting systems capable of enhancing speech signals, all demonstrated and verified with improved performance using some form of assessment methods such as ASR, Segmental SNR, Itakura-Saito measure (IS), etc [31–34]. If the reported improvement includes improvement of intelligibility considering the fact that overall quality incorporates intelligibility as one of its components, it is not possible for both Boll's statement and the wider consensus in the mentioned literature to be true. Although speech enhancement systems in general primarily aim and claim to improve overall quality rather than intelligibility specifically, thus to be strictly fair not improving intelligibility does not necessarily deny an improvement in quality; however, the fact that machine scores could be easily manipulated makes intelligibility assessment difficult since most of the time the improved scores do not reflect improvement of intelligibility as perceived by humans.

An example is given in Figure 2.5 which shows ASR performances for car-noise corrupted signals taken from Evans [35]. The red profile refers to ASR word accuracy of original noisy signals while the blue profile refers to word accuracy obtained after spectral subtraction is applied to the noisy signals. As shown the enhanced signals score significantly higher accuracy than the unprocessed signals with nearly 40% improvement at 5dB and 30% improvement at 0dB. However, informal listening tests suggest that though the enhanced signals sound less noisy and more comfortable, intelligibility has not changed.

Obviously, it is not sure whether Boll's statement is still true today. However, there are good reasons for believing so given that if a genuine, prominent and well-established intelligibility enhancer

does exist, such a breakthrough would be well-received in many applications. In fact, Boll's observation is supported by both past and recent publications. In 1986 Lim [36] published a review on a large range of speech enhancement techniques that account specifically for additive noise, and reverberation (plus echoes) respectively. Lim observed that while some techniques appear to improve speech quality, none actually improves speech intelligibility. In fact, some are even reported to decrease intelligibility. For instance, it is reported that even when accurate pitch infomation is given, the adaptive filtering technique developed by Frazier et al [37] tends to decrease the intelligibility at various SNRs despite sounding less noisy. Now, two decades after Lim's review, Hu and Louzou [2] report a comprehensive investigation on eight speech enhancement methods encompassing four classes of algorithms: spectral subtractive, subspace, statistical model-based and Wiener-type algorithms all in the context of their influence on speech intelligibility. The degradations considered are babble, car, street and train noise conditions at 0dB and 5dB. It is reported that while the majority of the algorithms seem to be able to preserve the intelligibility at 5dB, almost all degrade intelligibility at the lower SNR of 0dB. This is shown in Figure 2.6 with the data taken from Hu and Louzou [2]. The left most bars (red) for each group of bars in Figure 2.6 refer to human intelligibility of signals degraded by different additive noises; while the subsequent bars are human intelligibility of the same noisy signals after being further processed by various enhancement algorithms. Of the 64 cases of noise conditions and algorithms considered (8 algorithms $*$ 4 noise types $*$ 2 SNRs $=$ 64 cases), the only improvement reported is obtained by Wiener algorithm for 5dB car noise; with car noise being the most stationary noise among the 4 noises considered. In fact, one of the motivations for the investigation carried out in [2] is to identify algorithms that preserve or maintain intelligibility; it seems that intelligibility improvement was not actually expected. One other interesting observation is that among the algorithms investigated, the Wiener based algorithm appears to be best in perserving speech intelligibility, perhaps because of its relatively small amount of noise attenuation. At the other extreme, the perceptual Karhunen-Loeve transform (pKLT)) attenuates a great deal of noise which results in a more negative impact on speech intelligibility. This simply points out the difficulty in intelligibility enhancement.

In conclusion, it seems that processing speech to improve intelligibility is very difficult if not impossible. Yet it is relatively easy to improve machine scores. This paradox makes intelligibility assessment difficult as it implies that machine-based assessors can easily be fooled.

2.3 Ground Truth of Intelligibility

Opinion and fact are two different issues. For instance, 'Amy is female' is a fact, but 'Amy is pretty' is an opinion. Some speech research-related parallels are: speaker identity (in speaker recognition) is a fact; while speech quality is an opinion. Fact is consistent while opinion is subjective and varies across humans. Empirical opinions obtained through controlled and well-designed methods are often treated as fact or ground truth. In the context of intelligibility assessment, level of intelligibility can also

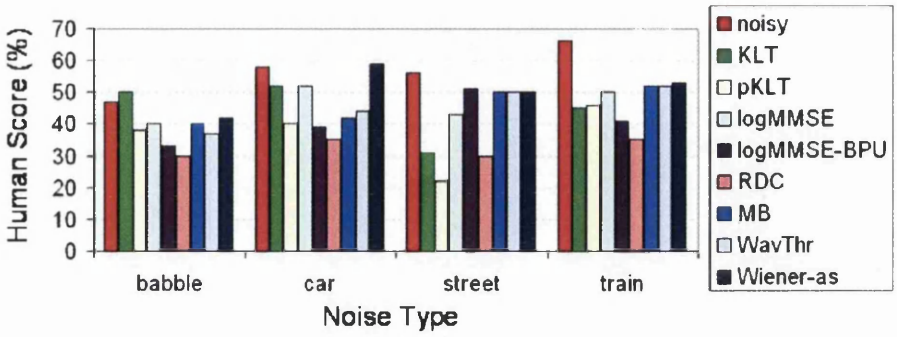


Figure 2.6: An illustration of intelligibility reduction when signals are processed by enhancement algorithms. Signals degraded by babble, car, street and train noise are processed by 8 different enhancement algorithms. Results shown are for signals at 0dB. The first bar (red bar) of every set of results shows intelligibility of original noisy signals without enhancement. The majority of the 'enhanced' signals have lower intelligibility. Data are taken from [2].

be considered as an opinion. Although it is not wholly subjective in the sense that it is immediately quantifiable ($x\%$ words correct); however, opinions of intelligibility do differ across listeners, equipment, databases, test approaches, listeners' attention or even mood. Objective assessors need to be assessed on well-established ground truth. This is made difficult if the ground truth itself is difficult to establish. Human listening tests are no doubt still deemed as the gold standard for intelligibility assessment but extensive resources are needed to conduct proper listening tests which adhere to the standards (i.e., number of listeners, listener training, anachoic chamber, etc), the laboriousness of which is exactly why objective measures are desired. In situations where ground truth is not readily available and carrying out full-scale listening tests is infeasible, any development of objective measures could only be assessed with semi-reliable ground truth obtained from a small group of listeners. The challenge of this research thus includes developing an objective measure that can be assessed by sufficient and more reliable ground truth.

2.4 Epilogue

The difficulty in objective assessment of intelligibility is foreseen in three areas. The first is the relatively small dynamic range of machine scores due to operating under high degradation. The second difficulty is the paradox that while it has been proven difficult to improve speech intelligibility for humans, it is relatively easy to improve machine equivalent scores. The third difficulty is lack of reliable and readily available ground truth for development of objective measure. This is due to the fact that intelligibility measurement is subjected to variation across human hence making the establishment of ground truth difficult. Having foreseen these difficulties, the goal of the work described here is to overcome them in a practical manner leading to a reliable objective measure of speech intelligibility.

Chapter 3

Human Listening Tests

Ground truth needs to be established for evaluation of the objective measures. This chapter describes the listening tests conducted to collect ground truth from human listeners. First of all, it has to be made clear that (i) this research has neither facility nor resource to conduct formal listening tests which may have rigorous requirements such as specially trained listeners and anechoic rooms; as an indication, in fact, despite the popularity of quality assessment, there are only a few laboratories in the world that can run subjective tests fully adhering to P.800 [38, 39]; (ii) secondly, it is not the aim of this research to develop state-of-art listening test methods. On the other hand, the aim relates to the development of the automatic machine-based alternatives. Tests conducted here are purely to give some indication of human opinion. Test methods as well test materials used may not be those that are commonly used in the field. Errors are perhaps larger than if tests are conducted following formal procedures, however, errors and inconsistencies are inevitable and they exist even with standard tests. The errors introduced by the less formal test method and materials used here only add to the existing, unavoidable error bar.

The chapter begins by reviewing the standard, formal listening tests available. Section 3.2 introduces the database used including the vocabulary and degradations considered, followed by a description of the test procedure in Section 3.3 where an online test system is introduced. Lastly, Section 3.4 presents the results obtained which serve as ground truth for evaluation of objective measures for the rest of the thesis. These results are referred to interchangeably as 'human scores' or 'human intelligibility' hereafter.

3.1 Background

Research on the measurement of intelligibility started as early as the beginning of the 19th century [40]. Subjective tests involving humans were the focus during early times and perhaps still are the preferred approach (over objective approaches) for measuring intelligibility [28]. Well-known subjective tests include the Articulation Test introduced by Fletcher and Steinberg at Bell Labs [21] and those specified in the ANSI S3.2-1989 (entitled *Method for Measuring the Intelligibility of Speech Over Communication Systems*) including the Modified Rhyme Test (MRT) [41], the Diagnostic Rhyme Test (DRT) [42] and Phonetically-Balanced (PB) word lists. Each of these measures will be briefly introduced here followed

by an overview of subjective tests in general touching on the difficulty in establishing reliable ground truth from subjective tests results.

Standard subjective intelligibility tests always involve a specifically chosen set of vocabulary. For instance, the MRT [41] uses 50 sets of words where each set consists of 6 rhyming monosyllabic English words. Each word is of the consonant-vowel-consonant (CVC) structure and words in each set differ only in their initial or ending consonant (e.g. 'hold - cold - told - fold - sold - gold'). Similar to the MRT, the DRT [42] also uses monosyllabic CVC English words except that words are in rhyming pairs rather than sets of 6, and words in each pair differ only in their initial consonants (e.g., 'yield - wield'). The DRT word sets faciliated the investigation of six distinctive acoustic-phonetic features of speech, namely voicing, nasality, sustenation, sibilation, graveness and compactness. Variants of the DRT include DMCT which uses bi-syllabic rather than monosyllabic words, and DALT which uses word pairs that differ in the final consonant. In both MRT and DRT the listener is shown all word options and asked to identify which in the set has been heard. Normally a carrier sentence is used both for the speaker to control his vocal effort and for the listener to be attentive when the test word is spoken. Besides rhyming words, phonetically-balanced words are also commonly used in intelligibility tests. One of the ANSI standards is Phonetically Balanced (PB) Word Lists which consists of 20 phonetically-balanced word lists with 50 monosyllabic English words in each list.

Over the years different test methods as well as materials have been developed to measure speech intelligibility but there is no single universal method that is THE method for obtaining ground truth. Different methods (and test materials) give different results and each has its advantages and limitations. Due to the subjectiveness of the measurement it is not suprising that sometimes contradicting comments regarding the same methods are found. One example is DRT which is designed such that the two consonants (of the two words in a pair) differ only in a single distinctive feature (for e.g., voicing, nasality, etc). In other words DRT assumes that speech errors can be adequately predicted with confusions over single feature. Yet Greenspan et al [43] in his investigations contradicts this assumption and suggests that low-bit-rate coded speech yield multi-feature confusions that could not be easily predicted from the single-feature confusions. The DRT is also critisised for using words that only differ in initial consonants. This led to the introduction of the minimal pairs of intelligibility tests (MPIT) by Santen in 1993 [44] where paired words differ in initial, medial as well as ending consonants. However, on another account, Steeneken's investigation in [45] recommends the use of material based on both consonants and vowels. Another criticism for the DRT is that the test material is limited and hence could unintentionally encourage listeners learning.

Besides, both MRT and DRT are closed-response methods in that the answer has to come from the selection of alternatives given. The advantage is that only simple and minimal listener training is required, as opposed to the PB sentences that requires at least 10 hours of training for the listeners to maintain stable performance [46]. Open-response approaches require extensive listening training and give significantly lower intelligibility scores [47]. However, Steeneken is of the opinion that the

open-response approach gives better discrimination [45].

Both MRT and DRT are considered as segmental-level tests, i.e. only a single segment or phoneme intelligibility is tested. A question that arises is whether intelligibility can be evaluated at segmental level since such tests are context-isolated. An alternative is tests at comprehension level where listeners are posed questions and indications of intelligibility are based on whether the questions are correctly answered. An example question could be 'what is 2 + 2?'. De Logu et al [48] introduced a test where listeners hear two passages and are then asked to answer 10 multiple choice questions. New form of comprehension-level tests also use response time as an indication of intelligibility. In comprehension tests it is not crucial to recognise every single word or phoneme as it is the gist of the message that matters. Presumably comprehensive tests are more realistic than the context-isolated segmental-level tests and better reflect daily communication tasks, however, such tests are difficult to construct due to involvement of cognitive factors, i.e., intelligence of the listeners, which is almost impossible to quantify.

Besides the advantages and disadvantages of different test approaches, the test setup is also important and could influence results. Rix [49] discusses the variation between subjective quality tests and give three examples of factors that affect subjective opinions: cultural variation, individual variation and balance of conditions. The last factor points out that within a group of conditions (degradations) under test, the opinions obtained are affected by the average / overall quality of all conditions. This means that if all conditions in the group range only from 'poor' to 'very poor', then the ones with poor conditions are likely to be rated 'good' simply because they are the best examples within that group. In contrast, if all conditions are of 'good' quality and above, then the conditions with 'good' quality are likely to be rated 'bad' simply because they are poorer than the 'very good' examples. This discrepancy further emphasises the difficulty in obtaining reliable human scores. Though the discussion in [49] relates predominantly to quality tests, it is a reasonable assumption that the same discrepancy applies to all tests involving humans.

Furthermore, there are also problems that are commonly expected with subjective listening tests such as learning effects and listener concentration problems. Besides, difference in hearing ability across listeners and performance of the same listeners across different sessions are also inevitable. Hawley [50] constructed a list of 6 categories of factors that affect speech intelligibility. Two factors in the list are in fact listener characteristics and listening conditions. Some relevant points are stated here:

- Listener characteristics: intelligence, motivation, vocabulary size, native language, age and sex.
- Listening conditions: free field or on earphone, reverberation time, listening level, monaural or binaural presentation, high or low air pressure, etc.

All these factors could contribute to variation in listening test results.

The difficulty in obtaining reliable ground truth is supported by Mapp [19] who points out that though word score testing (for e.g., DRT and MRT) has been well standardised, surprisingly there are significant variations in the results between testers. The main problems contributing to these discrepancies range widely from too small a listening sample and too few talkers, to poor listeners training or preparation, and over familiarity of listeners with test materials. To further illustrate this point, Mapp compares the listening test results for the same range of conditions from two leading and highly respected research institutions. The two sets of results are significantly different and a question posed by Mapp is that "if well respected institutions make such error, what chance does the less experienced testers have?" He continues: "Subject-based testing is the traditional method and has been referred to as the 'Gold Standard' - though in practice it is often far from being that."

In conclusion, clearly certain methods have wider error bars than others, but in general, errors with human listening tests are inevitable.

3.2 Database

3.2.1 Digits Vocabulary

Raw signals used throughout this thesis are a subset of the ETSI-Aurora2 digit-string corpus [26] which is derived from the TIDigits database. There are 11 words in the vocabulary, namely the digits one - nine, 'oh' and zero. Signals are taken from Test Set A defined in the Aurora2 framework which consists of 4004 digit strings collected from 104 adult speakers (52 male and 52 female). The test set contains a mixture of clean and degraded signals where 'clean' signals are the original 20kHz signals from the TIDigits downsampled to 8kHz and filtered with the G.712 characteristic [26]. The test set consists of digit strings of varied length: from 1-digit to strings of 7-digits. Here all the 4-digit strings in the Aurora2 Test Set A are extracted, totalling 566 utterances. Examples of signals are '139O' (one-three-nine-oh), '9486' (nine-four-eight-six), etc. Fixed-length signals are used because varying signal length would cause unnecessary inconsistencies across tests conducted for different conditions. For instance, intelligibility scores obtained from a test signal of 1-digit length at 0dB cannot be fairly compared with that of a test signal of 7-digit length at 5dB. A length of 4-digits is used as it is a moderate choice between over-long digit strings which causes much strain for the listeners to memorise and over-short digit strings that lacks sufficient variance for normalisation and causes performance to be utterance-dependent.

Obviously, the choice of vocabulary is important in order for the test result to reliably reflect the actual state of intelligibility. While it is well recognised that this digit database is not phonetically rich and is not optimal for quality nor intelligibility testing, it has the following few advantages: (i) it provides a straightforward scoring process for the listening tests; (ii) results will exhibit minimal influence from listener vocabulary power and experience; (iii) the database is explicitly configured for

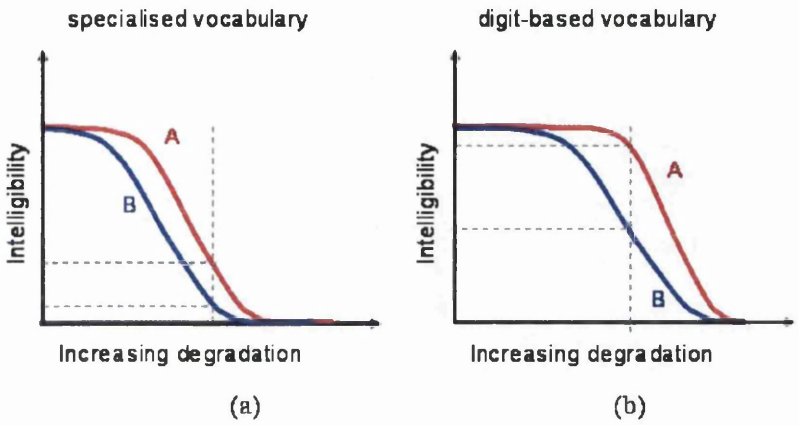


Figure 3.1: An illustration of the postulation that the use of different test materials affect mostly the offset and less so the correlation, i.e. ranking of profiles. if (a) is the intelligibility response obtained when specialised vocabulary is used, it is postulated that the use of digit vocabulary would give higher scores as shown by grey dotted lines in (b) but the order of A and B would remain.

ASR. Obviously advantage (iii) is driven by the wish to investigate ASR as an intelligibility assessor in later chapters. In fact, Hicks et al [24] use the same database for ASR testing of speech intelligibility.

It is hoped that the use of different vocabularies at most affects only the offset, while maintaining the ranking correlation between different test conditions (i.e., which test condition is the more intelligible?). This is illustrated in Figure 3.1 using hypothetical profiles where Figure 3.1(a) uses specialised vocabulary such as DRT rhyming word sets and 3.1(b) uses digits. Profile A and B refer to intelligibility of signals processed in two different ways. Intelligibility scores obtained when a digits-only database is used are most probably higher than those from a specialised vocab database as shown in Figure 3.1(b) due to the relatively low word difficulty (compare horizontal dotted lines in both figures). But it is postulated that the order / ranking of profiles A and B would not be affected. As far as this research is concerned, it is the profile ranking that matters (i.e., objective (i) in Section 1.1).

The set of 566 four-digit clean signals becomes the origin, from which all objective and subjective test sets are generated by processing the clean signals through various degradations.

3.2.2 Degradations

Six sets of degradations are considered leading to the generation of 6 test corpuses which will be used throughout this thesis for evaluation of the objective assessment systems. The degradations cover 3 main types: additive or environmental noises, distortions introduced by speech coding schemes and speech enhancement processes. These three types of degradations are considered as they are the fundamental degradations found in speech processing. The 6 degradation sets (DS) are referred to as are $DS1_{add}$, $DS2_{add}$, $DS3_{cod}$, $DS4_{cod}$, $DS5_{enh}$ and $DS6_{enh}$ respectively where the suffix indicates the degradation type considered in that set (add=additive noise, cod=coding, and enh=enhancement). In detail, the 6 DSs are:

- **DS1**$_{add}$: Degradations considered here are 8 real life background noises used in the production of the Aurora2 digit-string corpus [26]. The noises are airport, babble (crowd of people), car, exhibition hall, restaurant, street, subway (suburban train) and train station noises. They represent some of the most probable application scenarios for telecommunication terminals. Noise characteristics include the relatively 'nice' stationary ones (for e.g., car noise) and the more 'hostile' and speech-like ones (for e.g., babble). All noises are sampled at 8kHz. Noises are added individually to the clean signals using the standard noise addition software from ITU-T Rec., P.56 [51] at SNRs ranging from -10dB to 5dB at 0.5dB intervals[1]. Note that the same noise addition software is used in the creation of the Aurora2 corpus. The noises are referred to as airport, babble, car, exhibition, restaurant, street, subway and train.

- **DS2**$_{add}$: Degradations considered here are 10 additive noises obtained from the Center for Spoken Language Understanding (CSLU). Noise sources include aircraft cockpit, city rain, flat communication channel, automobile highway, helicopter fly-by, large city, large crowd, IBM cooling fan, SUN cooling fan, and white Gaussian noise. The noises are added to the clean signals in the same way as for DS1$_{add}$. The noises are referred to as aircraft, cityrain, flatch, highway, heliflyby, largecity, largecrowd, ibmcoolfan, suncoolfan and gaussian.

- **DS3**$_{cod}$: Degradations considered here are those introduced by 7 speech coding schemes namely GSM (13kbps) [52], G728 low delay CELP(16kbps) [53], Federal Standard MELP (2.4kbps) and LPC-10e (2.4kbps) [54], and three CCITT ADPCM coder [55]: G.721 4-bit (32kbps) , G.723 3-bit(24kbps) and G.723 5-bit (40kbps). Firstly, car noise is added to the clean signals at SNR between of -10dB to 10dB at 1dB intervals. There is no particular reason why car noise is chosen. The SNR range here differs from those of DS1$_{add}$ and DS2$_{add}$ because the degradations here are more adverse hence the higher intelligibility threshold region. The car-noise degraded signals are en-decoded using the coding systems at various numbers of tandemings giving a total of 16 degradations in this test set. Example degradations in this database are 3GSM (degraded by car noise and en-decoded thrice by GSM), 2LPC (degraded by car noise and en-decoded twice by LPC), etc.

- **DS4**$_{cod}$: The same coding algorithms in DS3$_{cod}$ are considered but now with train or street noises and tandemings involve mixed coding schemes (rather tandeming of the same coding scheme). The SNR range is from 10dB down to 0dB at 0.5dB intervals. There are 15 degradations in this test set. One example of degradation in this database is Train.CELP.GSM.GSM (degraded by train station noise, then en-decoded by CELP, then en-decoded by GSM twice).

- **DS5**$_{enh}$: Degradations considered here are 13 speech enhancement processes. First of all the signals are degraded by car noise at SNRs ranging from 0dB to -10dB at 0.5dB intervals. The

[1]This SNR range is not as in the Aurora standard which considers only clean, then 20dB to -5dB.

noisy signals are then processed with the various enhancement processes[2]. The algorithms and their abbreviations are (1) spectral substraction (SS) from Boll (SSBoll) [56], (2) SSBoll79 with quantile filtering (SSBoll-qf) [56], (3) SS from Berouti (SSBerouti) [57] , (4) SS from Kamath (SSKamath) [58], (5) minimum mean-square error short-time spectral amplitude estimator (MMSE) [59], (6) Wiener filter (Wiener) [60], (7) SS with modified minimum statistics (SS-mms) [61], (8) perceptual wavelet filter optimised for car and factory noise (PWF-cf) [62], (9) PWF optimized for hearing aids (PWF-ha) [62], (10) PWF optimized for noise recognition (PWF-nr) [62], (11) PWF-nr with continuous buffer (PWF-nr2) [62], (12) lattice filter optimized for hearing aids (LF-ha) [63], and (13) LF optimized for ASR in car and factory noise (LF-asr) [64].

- **DS6**$_{enh}$: Degradations considered here are those introduced by the non-linear spectral subtraction (NLSS) algorithm contributed by Evans [35]. Firstly different noises from DS1$_{add}$ are added to the clean signals at SNR ranges of 5dB to -10dB at 0.5dB intervals. Noisy signals are then processed by NLSS at different configurations. 13 different configurations are obtained by tuning the two parameters commonly associated with implementations of spectral subtraction [57] namely the noise over-estimate and noise floor. One example of degradations in this database is Street.NLSS_a4.0_b0.01 (degraded by street noise then processed by NLSS with noise over estimate set to 4.0 and noise floor set to 0.01).

Two test sets are created for each degradation type in order to demonstrate different findings in the experiments carried out on objective measures in coming chapters. The reason that additive noises in DS2$_{add}$ are put in a separate group to DS1$_{add}$ is because noises in DS1$_{add}$ are taken from the Aurora2 framework which has been explicitly configured for ASR, whereas DS2$_{add}$ noises have not. It is interesting to investigate any differences that might exist between the two especially when ASR is assessed as an objective intelligibility measure (Chapter 5). Meanwhile the coding test sets DS3$_{cod}$ and DS4$_{cod}$ are in different sets since it is postulated that tandemings of single codec in DS3$_{cod}$ is less challenging than DS4$_{cod}$. Lastly, DS6$_{enh}$ differs from DS5$_{enh}$ in that it represents the scenario of a new system during developmental stage where the configurations giving the best performance needs to be identified (parameter optimisation). Brief description of the 6 test sets are presented in Table 3.1. More detailed description (as well as sources of the processes/degradations) can be found in Appendix A.

3.2.3 Experimental Databases

Each DS comprises $m \times n$ test sets where m is the number of degradations considered in that particular DS, and n the range of SNRs considered. For instance, DS1$_{add}$ considers 8 different types of degradations at 31 different SNRs (from -10dB to 5dB at 0.5dB intervals), totalling to 8×31=248 test

[2]Courtesy of Dr Tuan V. Pham from the Signal Processing & Speech Communication Laboratory at Graz University of Technology in Austria.

Type	Test Set	Descriptions
Additive	$DS1_{add}$	additive noises of diverse characteristics including both speech-like and more stationary noises.
	$DS2_{add}$	additive noises, mostly fairly stationary.
Coding	$DS3_{cod}$	car noise and tandemings of single coding schemes
	$DS4_{cod}$	various $DS1_{add}$ noises and tandeming of mixed coding schemes
Enhancement	$DS5_{enh}$	car noise and different speech enhancement processes
	$DS6_{enh}$	various $DS1_{add}$ noises and different configurations of NLSS

Table 3.1: Brief descriptions of test sets DS1 to DS6.

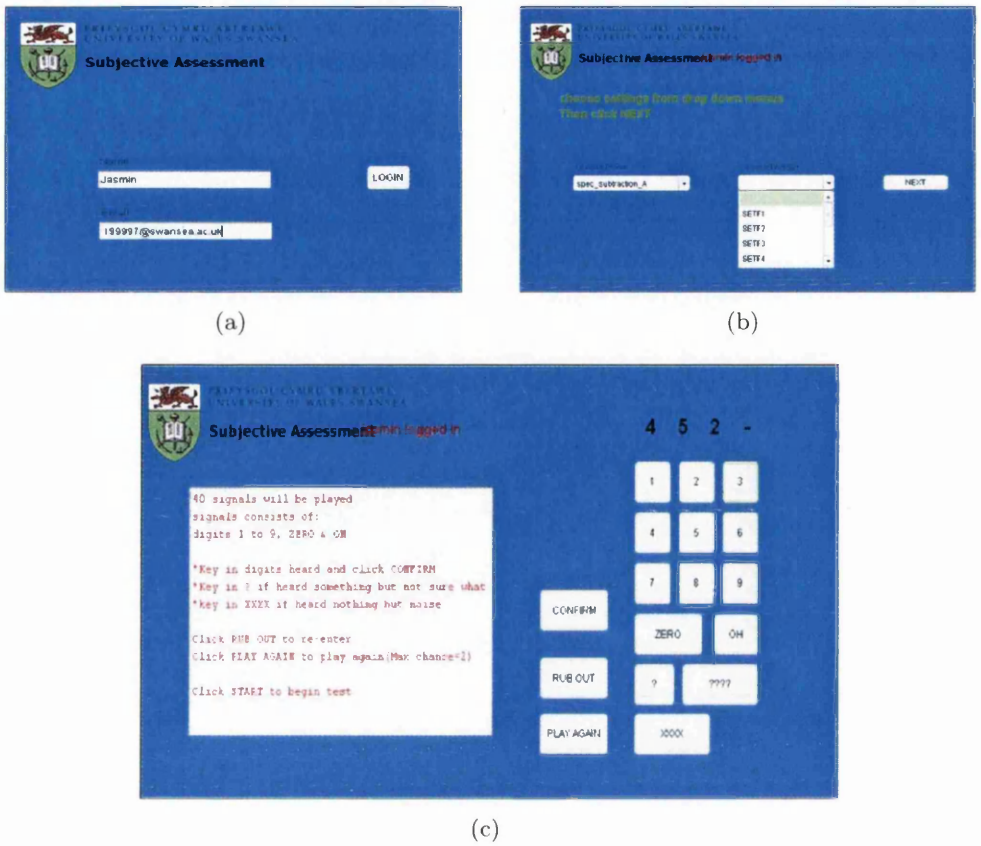
sets in $DS1_{add}$ (one for each degradation condition). Each test set consists of 566 four-digit signals degraded by the same degradation condition. These 6 DSs are used for evaluating objective measures throughout part I of this thesis and last section of Part II.

Databases for the listening tests are subsets of respective complete test sets which consist of 566 test utterances per condition. A signal is chosen for each SNR to make up test set for a particular condition under test. For example, to obtain human scores for $DS1_{add}$, the listener needs to perform listening tests for 8 test sets (8 noise types are considered in $DS1_{add}$) where each consists of 31 signals (31 SNRs are considered in $DS1_{add}$, i.e., -10dB to 5dB at 0.5dB intervals). The test set is designed such that there is no repeating of test utterances within one test and minimal repeating across tests of the same degradation category to avoid the possibility of the subject memorising the test signals. For every condition 10 test sets are prepared, i.e. 5 male-spoken and 5 female-spoken.

3.3 Procedure

3.3.1 Online Test Interface

In an attempt to compensate for the lack of formality in the listening tests carried out for this research, it is desirable to collect as many responses as possible and to compute the average. An online test has been implemented in order to recruit a large number of listeners. Figure 3.2 shows the interface of the online test. Figure 3.2(a) shows the login page and (b) shows the selection page where the listener chooses the test set, i.e., degradation under test (left pull down menu) and 1 random test set out of the 10 sets available for each degradation (right pull down menu). Figure 3.2(c) shows the test page where upon clicking the 'start' button, the signals in the test set would be played one by one. The panel on the left shows instructions for the test. The number pad allows the user to key in the digits heard as the signals are being played. The '?' key may be hit when a part of the utterance is incomprehensible while the '????' key (equivalent to hitting the '?' key 4 times) is hit when the whole utterance is incomprehensible. The 'XXXX' key is hit when no deterministic signal can be heard at all. The web address for the online test is http://eeceltic.swan.ac.uk/subj.



(a) (b)

(c)

Figure 3.2: The interface of the online listening test: (a) login page; (b) degradation and test set selection; (c) test page where test signals are played and the listener key in digits heard by tapping the number pad; instructions are given in the text box to the left of the number pad.

3.3.2 Test Arrangement

All tests under the same DS are performed by the same group of listeners, for example the airport test set and babble test set must be tested with the same people since both airport and babble belong to $DS1_{add}$. This is so that performance between different degradations under the same DS can be directly compared. There are 50 listeners in any one group. A test set consisting of 31 signals take approximately 3 to 4 minutes. To avoid fatigue and learning, the listeners are requested to keep each session below 15 minutes. For each setting every listener perform a test on a male-spoken and a female-spoken test set (randomly choosed from the 5 male-spoken and 5 female-spoken test sets available). Hence with 50 listeners 100 sets of responses are obtained for each degradation.

Real answer	Human's answer	$SUBJcorr$
1234	1256	2
1234	OO12	2
1234	OO14	1
1234	1348	3
1234	1?3	2
1234	1?2	1
1234	1?23	3

Table 3.2: Examples of how $SUBJcorr$ is scored where the correctly identified digits are underlined.

3.3.3 Human Subjects

All human subjects are university students, the majority of whom are in the age range of 20 to 30. They are of mixed genders and nationalities. Instructions are given on webpage and no training is provided. Besides, there is no supervision since the tests are conducted online. Apart from the use of headphones, no other specification on equipment is imposed. Given the objective of ranking (rather than absolute score) and given that it is a simple digit database, it is thought that the effects of these factors are minimal.

3.3.4 Deducing Ground truth

The level of intelligibility as indicated by the human subjects is quantified simply by the number of digits identified correctly, referred to as $SUBJcorr$. The $SUBJcorr$ takes shifting into account, for instance, 3 $SUBJcorr$ is scored when 'O123' is keyed by the listener whereby in fact the signal played is '1234'. Some examples of how $SUBJcorr$ is scored are shown in Table 3.2. Other indicators such as the number of '?' responses, deletion and others could be investigated. The investigation in [65] shows that these different indicators agree well with each other, hence only $SUBJcorr$ is used here.

Figure 3.3 shows example result from a single listening test for airport noise degraded signals from a random listener. The scale of the Y-axis ranges from 0 to 4 and corresponds to the number of digits correctly answered out of the 4-digit string played. Ideally the profile is expected to decline monotonically as the SNR decreases along the X-axis. In reality such a trend is only vaguely seen through the 'zigzag' response as shown in Figure 3.3. Notice that most 'zigzagging' occurs at lower SNRs while scores at higher SNRs are almost consistently high. Such an unruly response is largely attributed to limitations of the vocabulary since certain digits are more confusable, for example, '2' and '3'; while others are more distinctive and can be more easily guessed, for example, '7' with its lengthy proounciation is easily identifiable. Besides, the 'zigzagging' could also be due to the fact that the corrupting noises (when additive noise is considered) are random chunks of a long continuous signal recorded from real situations, thus different chunks differ in their degrading effect. For instance,

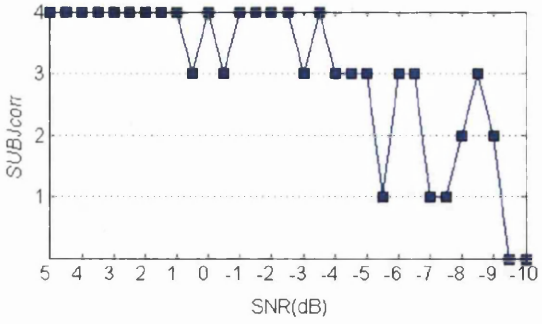


Figure 3.3: An illustration of the 'unruly' responses from the listening tests. Figure shows example response of one listener for airport noise degraded signals under $DS1_{add}$. Along the X-axis is decreasing SNR, along the Y-axis is the number of digits correctly recognised by the listener from each 4-digit utterance. Monotonic responses are expected but zig-zag responses are generally observed.

a chunk of street noise with the beeping of a car alarm would be considerably more degrading than those with just cars or people passing by. We accept this as a weakness of the vocabulary chosen and noise adding process and hope to overcome this by averaging a large number of responses from different listeners and sessions. Figure 3.4 shows an example of how the 'unruly' response smooths out as more and more responses are added. The blue profile shows 1 response from 1 listener and the black profile is the average of 40 responses from 20 listeners.

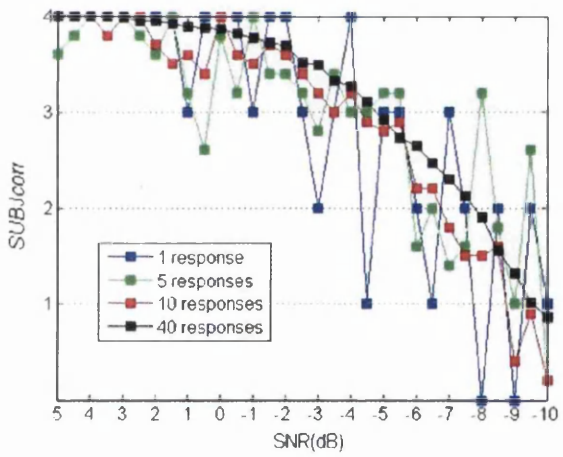


Figure 3.4: An illustration of obtaining a smoother listening test results by averaging large number of responses across listeners and test sets. The blue profile is a single response, whereas the black profile corresponds to an averege across 40 responses. Result are for the airport noise condition from test set $DS1_{add}$.

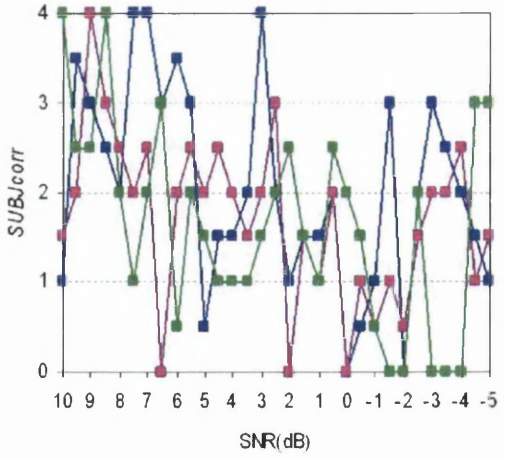


Figure 3.5: Example response of an unattentive listener. Each profile is for a single response for a degradation condition for e.g., airport profile, babble profile, etc. Profiles should be decreasing monotonically as SNR decreases along the X-axis, however, such a trend is not observed. Low recognitions are observed even at high SNRs.

Note that since there is no supervision while the test is being conducted it is hard to control the quality of the response obtained. Outliers are sifted out manually while analysing the results. For example, Figure 3.5 shows example responses of an unattentive listener where each profile is the human

response for a degradation condition under test. Note that a high-to-low trend as SNR decreases is not obvious in the figure, reflecting the lack of concentration while the listener was carrying out the test. Such responses are discarded manually by observing plots such as the one shown in Figure 3.5. Such intervention is deemed justifiable since the training provided for listeners of formal listening tests serves similar purpose, that is, to reduce outliers and increase reliability of the data. For each set of listening tests (for DS1 to DS6), the best 100 responses that give no more than 0.5 standard deviation are used.

3.4 Human Scores

This section presents the listening test results obtained for test sets DS1 tp DS6. Scores are averaged across the listeners (after unreliable responses from inattentive listeners are eliminated) and normalised to be between 0% and 100%. Figure 3.6 shows the scores obtained for $DS1_{add}$ where the 8 profiles represent ground truth for the 8 degradation conditions considered in $DS1_{add}$. Profiles for the other test sets can be found in Appendix B. For ease of observation, the intelligibility ranking of the different degradation conditions in a DS are presented in the form of bar plots as shown in Figure 3.8. Each bar represents the scores obtained for a degradation condition averaged across listeners as well as the SNR range, in other words, scores integrated along the x-axis in Figure 3.6. The bars are presented in ascending order of intelligibility. An example ground truth deduced from Figure 3.8(a) is that for the same SNR range babble noise degraded signals are less intelligible than the subway noise degraded signals, as shown by the lower babble bar and higher subway bar.

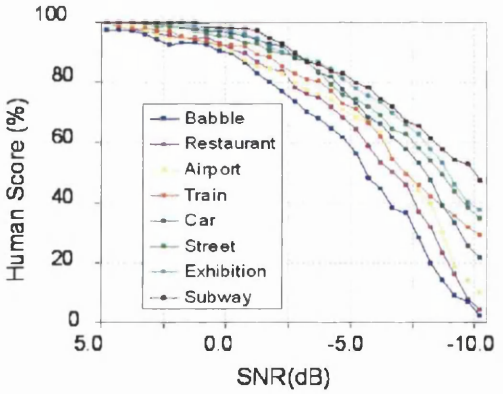


Figure 3.6: Human scores for $DS1_{add}$ averaged across listeners.

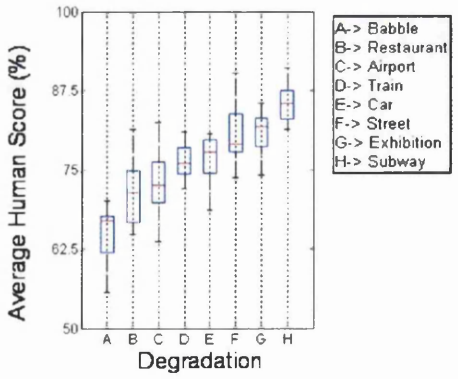


Figure 3.7: Human scores for $DS1_{add}$ averaged across listeners and SNRs for each degradation condition.

Errors in human listening tests are unavoidable. As mentioned in Section 3.3 the first 100 acceptable responses (seemingly reliable responses from attentive listeners) that give no more than 0.5 standard deviation for each condition in each DS are used. Clearly within that variation there might

be overlapping regions between the tolarance of different conditions under comparison. An example is illustrated for $DS1_{add}$ using a boxplot [3] as shown in Figure 3.7. Notice that while the medians (red line in the middle) for each condition are distinctively separated, the spreads overlap with one another. The bigger that overlapping region, the bigger the possibility for errors. For example, restaurant and airport conditions are largely overlapped. Given another group of listeners the ranking might well be the opposite. We acknowledge the possible existence of such errors with the human scores collected, but hope that the scores are sufficient to provide an indication of the relative performances of the objective measures.

Subsequent chapters are dedicated to the assessment of objective measures. The usefulness of those measures are judged by the correlations between their objective scores and the human scores. Note that the correlation of interest is on trend rather than absolute values, hence an objective measure is deemed useful if its output scores for a given set of degradation conditions are ranked in the same order as indicated by the human scores.

[3] A boxplot is a convenient way of depicting spread of data and identifying outliers. It summarises the following statistics: median, lower and upper quartile, minimum and maximum data values

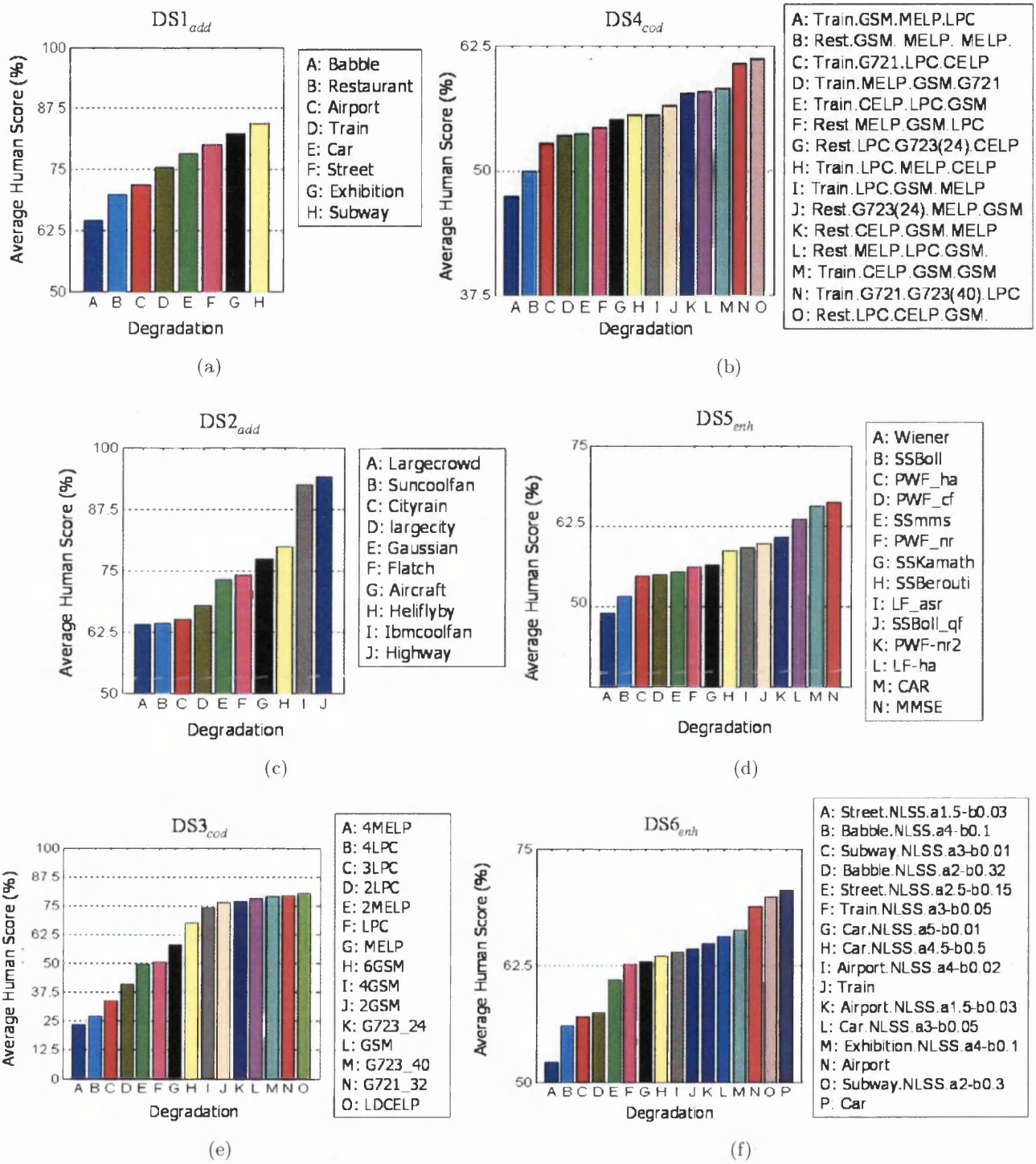


Figure 3.8: Human scores (averaged across listeners and SNRs) for the 6 test sets, namely $DS1_{add}$, $DS2_{add}$, $DS3_{cod}$, $DS4_{cod}$, $DS5_{enh}$ and $DS6_{enh}$. The bars are arranged in ascending order of intelligibility with degradation condition at far left being the least intelligible and vice-versa.

Part I

Existing Measures

Chapter 4

Quality Measures for Intelligibility Assessments

Whilst speech quality can be described as how good a speech signal sounds, speech intelligibility is concerned with how well understood the signal or message is. Some descriptions of the relationship between quality and intelligibility are as follows (text in italic are direct quotes):

(1) Hansen [28] stated that *"intelligibility can be viewed as one aspect of quality, since high quality speech generally implies good intelligibility. However, the converse needs not be true"*;

(2) Chong et al [20], who gives a list of evidence to demonstrate that speech quality encompasses a broader scope in which intelligibility is included, stated that *"there exists a strong correlation between speech quality and intelligibility in that the level of intelligibility relates to determination of quality,..., intelligibility possesses a narrower scope and can be considered as a dimension of speech quality"*;

(3) In a study which compares intelligibility scores deduced from DRT and quality scores from Diagnostis Acceptability Measure (DAM), Voiers [66] concluded that *"quality is heavily but not totally dependent on measured intelligibility"*;

(4) In an experiment which investigates the effect of amplitude distortion upon quality and intelligibility Licklider [67] observed that intelligibility is significantly less affected than quality by the distortion;

(5) Beerends [68] considers the potential of PESQ for assessing intelligibility of speech processed by vocoders, states that *"One should be aware of the fact that one can improve speech quality while decreasing intelligibility"*.

Based on the descriptions listed above, it is generally accepted that intelligibility is a dimension of the multidimensional entity of overall quality [20]. This is confirmed in two experiments performed by Preminger and Van Tasell [27] in an attempt to quantify the relationship between the two. The experiments require human subjects to rate 5 attributes of overall quality, namely intelligibility, pleasantness, loudness, ease of listening and total impression as a function of change to the frequency response of a listening system. In the first experiment, intelligibility of the test signal varies over a wide range of 25% to 100%; in the second experiment, signal intelligibility is held constant. First experiment reports high correlation between intelligibility and other quality components. This suggests strong link between

intelligibility and quality. Lower correlation is reported by second experiment and this suggests that intelligibility is the key contributor to the high correlation in the first experiment.

Despite the postulation in Section 2.1 that measurement of intelligibility using quality measures would result in a relatively small section of the full dynamic range of scores from quality measures hence making intelligibility ranking more difficult than it already is, it is unsure how useful these quality measures might be even within this constrained dynamic range. To quote directly from Kaga et al [69]: *"Since we can assume that speech quality is correlated with intelligibility, it should be possible to estimate the intelligibility from the estimated opinion scores or some of its derivatives."* The experiments reported by Priminger and Van Tasell [27] clearly suggest the existence of strong link between intelligibility and quality. For this reason, plus the ready availability of large selection of objective quality measures, it is worth investigating the potential of quality measures in the context of intelligibility estimation.

Nine prominent objective quality measures are considered here. The measures include the perceptual-based: PESQ, MBSD, MNB; the spectral-based: IS, LAR, LLR, WSS, and the SNR-based: Classical SNR and Segmental SNR (see Table 4.1 for expansion). The performance of these measures are judged by Kendall and Pearson correlations computed by comparing their objective estimates with human results. Section 4.1 presents an overview of quality measures, making a contribution by reviewing the measures in the context of intelligibility assessment. Section 4.2 describes the experimental procedures including the correlation methods employed. Lastly, Section 4.3 presents the performance of the measures in terms of correlations with ground truth provided by human listeners.

4.1 Quality Measures

4.1.1 An Overview

Generally the conventional quality measures can be categorised into: time domain, spectral domain and perceptual domain. Some prominent measures are shown in Table 4.1. Unless specified otherwise, all correlation values quoted in this section are referring to Pearson correlations between objective scores and human perceived quality.

Time-domain Measures

The simplest approach is the time-domain approach, which is basically a comparison of the sampled original and degraded waveforms. This also means that alignment or synchronisation of the two waveforms is crucial in order to obtain any meaningful measurement. Common measures under this

Domain	Measures
Time	Classical SNR (CSNR) Segmental SNR (SegSNR) Frequency-weighted segmental SNR (FW-SegSNR)
Spectral	Log area ratio (LAR) Log likelihood ratio (LLR) Itakura-Saito (IS) Cepstral distance (CD) Weighted spectral slope (WSS) (1988)
Perceptual	Bark spectral distortion (BSD) (1993) Modified BSD (MBSD) and enhanced modified BSD (EMBSD) (1998) Measuring Normalizing Blocks (MNB) Perceptual speech quality measurement (PSQM) (1998) Perceptual evaluation of speech quality (PESQ) (2001)

Table 4.1: Table shows some prominently used quality measures. The measures are categorised into 3: time, spectral and perceptual domain.

category are Signal-to-Noise ratio (classical SNR, CSNR) measure and its various variations. One variation is segmental SNR (SegSNR) which computes SNR over short segments of typically 15-20ms the scores of which are averaged to indicate the overall signal quality. Frequency weighted segmental SNR (FW-SegSNR) further enhances SegSNR by imposing weighting on frequency bands. Quackenbush et al [5] reports correlations 0.24, 0.77 and 0.93 for CSNR, SegSNR and FW-SegSNR respectively when tested with distortions introduced by waveform coders. The SNR-based measures are generally known to be poorly correlated in the context of other degradations types [5], for example one of vocoders. Despite poor correlation the SNR-based measures are ironically still rather widely used in quality assessment [70] due perhaps to its simplicity.

Spectral-domain Measures

Alternative to the SNR-based measures, the spectral-domain measures are more widely applicable and reliable [5, 71, 72]. These measures claim to assess not only linear but also non-linear distortions such as those introduced by coding. Unlike time-domain measures, spectral-domain measures utilise second-order properties of the waveforms, such as the autocorrelation or the spectral models. Many use parameters of speech production models, which explain why they are more applicable to distortion introduced by vocoders based on similar model. For example, many spectral-domain measures are based on parameterizations of the LPC vocal tract models of original and distorted speech. The parameters used can be the prediction coefficients, or transformation of the predictor coefficients such as log area ratio coefficients. Each set of parameters provides a different way of quantifying the differences between the vocal tract models of the original and distorted speech. Common LPC parameter-based measures include the log likelihood ratio measure (LLR, also known as Itakura distance measure) [5], the log area

ratio (LAR) measure [5], the Itakura-Saito (IS) measure [73, 74]. After comparison studies of several kinds of objective measures, Kitawaki et al [71,72] concluded that spectral domain measures correspond better to human scores than time-domain measures. Quackenbush et al [5] tested the measures with a large range of distortions: eleven types of coding distortion including both pitch-excited vocoders and waveform coding techniques; and fourteen classes of controlled distortions including filtering, additive noise, echo, various types of clipping, interruptions and frequency variant distortions. Correlations of the LPC-based measures reported by Quackenbush et al [5] range from 0.06 by LPC to the highest of 0.62 with LAR.

The spectral-domain measures can be divided into those that measure spectral distortion or spectral envelope distortion. Kitawaki et al [71,72] observed that the latter corresponds better to human scores, and of several such measures, the LPC cepstral distance measure (CD) achieved a correlation of 0.87 and is strongly proposed as an accurate quality estimator for low-bit rate coding systems and other non-linear distortions alike. Another well-reported spectral envelope-based measure is the weighted spectral slope (WSS) by Klatt [75,76], which estimates distortion using the difference between spectral slopes (rather than absolute spectral distances) of original and distorted signals in each of the 36 overlapping frequency bands. The frequency bands approximate the critical bands of human auditory system and different weightings are assigned for each band. In 1988, Quackenbush et al's [5] thorough investigation into objective assessments concludes that the best predictors are those based on auditory criteria; of those, WSS gave the best correlation at 0.74.

Perceptual-domain Measures

Much of the development since mid 1990s has followed a perceptual-based approach. Explicit models for some of the known attributes of human auditory perception are quantified and incorporated in the quality estimators. The motivation for this perceptual-based approach is to create assessors that operate on speech signals following similar transformations through a human ear. Some perceptual features that have been identified as useful include the critical band concept, noise masking effects and loudness level. The critical band concept reflects the non-linear frequency response of the human hearing system. Each band corresponds to an actual section of the cochlea (approx. 1.3mm). With this concept, the inner ear can be simulated by having a series of band-pass filters with their centre frequencies following the critical band rates, as is the case for WSS. Noise masking is another effect that directly affects speech quality. This is the occlusion of one sound by another louder sound, which reflects of humans' inability in distinguishing two signals that are close in time or frequency. Masking may happen preceding or following the occurrence of the loud sound or when the two sounds are simultaneous. The perceived loudness is also a very important perceptual feature of the human hearing system. The ear is not equally sensitive to all frequencies in that it is less sensitive to low frequency sounds than mid to high frequencies, with the most sensitive range being approximately 1 - 5kHz [77]. These characteristics have become the basis of many perceptual-based measures and

have proved beneficial in improving correlations (between objective quality scores and true quality as perceived by humans) across a range of degradations.

Bark spectral distortion (BSD) measure proposed in 1992 by Yang [78, 79] was one of the first objective quality measure to incorporate psychoacoustic features. BSD is based on the assumption that speech quality is directly related to speech loudness whereby in order to calculate loudness, the speech signal is processed using the results of other psychoacoustic measurements, which includes critical band analysis, equal loudness pre-emphasis, and intensity-loudness power law. Finally the distortion is the mean-squared loudness error computed from the average Euclidean distance between the two signals in loudness domain. This rather simple judgement model is criticised in [7, 70] for not being a good measure for matching errors between two speech spectra. In [79] BSD is reported to have a correlation score at 0.8976 when tested with signals corrupted by Modulated Noise Reference Unit (MBRU) and coders (4.8kbps to 32 kbps). Yang [80] introduces two extensions to BSD, namely modified BSD (MBSD) and enhanced MBSD (EMBSD). MBSD incorporates noise masking threshold into the original BSD algorithm to differentiate between audible and inaudible distortions. EMBSD uses a more sophisticated cognition model which is refined using feedback from experiments on MBSD. Yang carried out an evaluation [80] that compares his two BSD extensions and two other appealing perceptually-based measures, namely the MNB by Stephen Voran [70, 81] and the ITU-T P.861(the predecessor of PESQ) by Beerends et al [8,82] with two different databases. One database is corrupted by coding distortions, while the other one is corrupted by distortions encountered in real network environments. While testing with the first database, all measures (i.e., EMBSD, P.861 and MNB) score 0.98 (Pearson correlation) apart from MBSD with score at 0.95. Whilst using the second database, MNB at 0.74 is the lowest score, followed by MBSD at 0.76, P.861 at 0.83 and the EMBSD at 0.87, hence concluding that EMBSD seems to be the best measure.

Most of the perceptual measures follow a similar structure where the auditory system is splitted into two phases, namely the hearing phase (perceptual transformation model: modelling the ear) and the judgement phase (judgement model: modelling the brain). The hearing phase is concerned with transforming signals into a perceptually relevant domain, whilst the judgement phase emulates auditory judgement by comparing the two perceptually transformed signals. The judgement model can be a simple process such as calculation of average Euclidean distance or it can be sophisticated, for instance, involving fuzzy logic. However, whilst most perceptual-based measures have an extensively researched hearing model, interestingly in almost all cases only simple distance-based judgement models have been considered. A notable exception of Measuring Normalizing Blocks (MNB) introduced by Voran [70,81] for US Department of Commerce around 1997. He takes a somewhat opposite approach to others whereby he gives more emphasise to the judgement model instead of the perceptual transformation model, as he observed that a simple distance metric cannot cover the wide range of distortions encountered in modern voice communication systems. Voran [70,81] also boldly claims that the many perceptual features such as outer-middle ear transfer function, absolute hearing thresholds, equal loudness curves and masking do not appear to be helpful in estimating quality of telephony bandwidth

speech. However, these are the features that form the foundation of many perceptual-based measures. MNB claims to correlate significantly better than other measures in degradation cases involving low bit rates, bit errors and frame erasures [70, 81]. Voran assessed MNB with codecs, tandeming of single/mixed codecs, frame erasures and bit errors. All tests achieved correlations as high as 0.929 and above apart from bit error condition which score 0.795. MNB is shown to outperform BSD by as much as 0.55 [70, 81]. Interestingly, Yang [80] also claims that MBSD outperforms MNB. Though MBSD is supposedly an enhanced version of BSD, only modest enhancement is reported in [78, 80]. Nonetheless, though both MBSD and MNB are considered in the context of coding distortion, obviously the experiments may well not be directly comparable. An agreement is that both measures are reported to achieve better correlation than the ITU-T P.861 perceptual speech quality measurement (PSQM), which is the predecessor of PESQ.

PESQ (Perceptual Evaluation of Speech Quality) is developed by Beerends et al [8] is standardised as the ITU-T standard P.862 in 2001 and is the successor of P.861 PSQM (1996). PESQ has an improved delay compensation module. The PESQ scores are calibrated using a large database of subjective tests. The performance of PESQ has been verified by strict evaluation of ITU-T standardisation process and is widely acknowledged as the state-of-the-art providing fast and reliable estimation of MOS over wide range of distortions. Further enhancement introduces the PESQ-LQ mapping which attempts to better correlate the raw PESQ scores with subjective listening quality (LQ) MOS, hence the name PESQ-LQ. The non-linear mapping modifies the raw PESQ scores at high and low ends of the scale where they are found to be less accurate. Rix [83, 84] investigates PESQ-LQ over a wide range of network distortions as well as languages and suggest that PESQ-LQ provides good estimation of MOS.

Non-Intrusive Approach

All measures discussed above take an intrusive approach where reference signals are needed during measurement and scores obtained relate to differences of quality between the reference and degraded test signals. The alternative is a non-intrusive approach where measurement is based on the degraded test signals alone. A well-known non-intrusive measure is the Single-sided Speech Quality Measure (3SQM) which is released in 2004. It is the combination of three winning systems in an open ITU-T competition that searches for non-intrusive quality measurement. The measure is standardised by ITU-T as P.563. Across the 18 ITU subjective databases it is evaluated on, a minimum of 0.80 correlation is achieved. In fact, and in some cases its accuracy competes that of PESQ which is an intrusive approach [9].

Unconventional Approach

The latest trend in quality assessment seems to follow a statistical data-driven approach pioneered largely by the works of Falk and Chan [25,85–89] since 2004. Whilst all measures introduced previously deduce quality estimation according to known rules of human hearings, for instance, high SNR suggests high quality, Falk et al propose measures that deduce quality estimation based on experience gained from large amount of examples given during training, the approach of which can be considered as data-driven.

One of Falk et al's latest developments uses Gaussian mixture model (GMM) for non-intrusive speech quality measurement [88]. The GMM is trained with well-chosen features extracted from clean and undistorted speech, serving as a reference model of normal behaviour of clean speech. This is done first by identifying frame features as belonging to voiced, unvoiced or inactive and then by modelling a GMM for each class. Indication of speech quality is obtained by matching features extracted from the test signals against each GMM model. The system claims to outperform the current 'state-of-the-art' P.563, the ITU-T standard for non-intrusive quality [9]. The system is further enhanced by adding a reference model of the behaviour of speech degraded by different transmission and/or coding schemes [89]. Improved robustness and accuracy is reported with the addition of this degraded reference model, alongside the existing clean reference model. Prior to this work Falk et al also proposed a similar GMM-based approach for intrusive speech quality assessment [85], the performance of which is reported to surpass the state-of-the-art PESQ.

In summary, the development in the area of objective quality assessment can be considered as consistently active. Higher and higher correlations are reported as more and more sophisticated measures are introduced. In short, objective quality assessment is a popular area of research and accomplishments are evident, quite in the contrary to intelligibility assessment.

4.1.2 Quality Measures in the context of Intelligibility Assessment

There are very few published studies on employment of quality measures for intelligibility assessment, with most of the related material published only in recent years, and mostly related to PESQ. Possible reasons could be that PESQ is the state-of-the-art quality measure hence should naturally be the one with the most potential for this task.

Perhaps the earliest reported work on the employment of PESQ for intelligibility assessment is by Beerends et al [90]. The degradations considered are interfering talkers at 4 SNRs (0, -3, -6 and -9dB). Experiments were conducted with diotic (identical signals at the ears) and dichotic (different signals at both ears) using four different beamforming algorithms as used in hearing devices, aimed at enhancing intelligibility by improving the SNR of the target signal. Experimental results show high correlations for both diotic and dichotic presentations, at 0.99 and 0.91 respectively.

Another investigation reported by Beerends et al looks at the applicability of PESQ in assessing intelligibility of speech degraded by low bit rate vocoders [68]. The database used are CVC words from NATO and correlation obtained is 0.86, which Beerends et al [68] deemed to be a distance away from reliable intelligibility estimation. Correlation of 0.95 is obtained after incorporating inspirations from both AI and STI where frame-by-frame Bark power spectrum differences between input and output are computed to indicate modulation difference. The paper concludes that further investigations are needed to check if the improvements apply over a wider range of distortions.

Very recently, Yamada et al [91, 92] propose to estimate intelligibility of noise-reduced Japanese speech using PESQ. Four noise reduction algorithms are considered. Vocabulary of the test signals are divided into four levels of difficulty where word difficulty is related to how familiar the word is in daily usage. Signals are degraded by car and subway noise at SNRs ranging from clean, 20, 15, 10, 5, and 0dB. An equation is empirically formulated to estimate word intelligibility as a function of PESQ MOS [91]. The transformed PESQ MOS is shown to estimate word intelligibility accurately without distinguishing the noise reduction algorithms or the SNRs. However, results are reported for each level of word difficulty which encompasses all levels of SNRs considered. It is unclear how the correlation and RMSE are calculated. It is possible that the high correlation is largely contributed by inter-SNR correlation, rather than inter-degradation type correlation (correlation across different types of noise reduction algorithms). It would be interesting to find out if similar high correlation still applies when considering just the noise reduction algorithms while holding other parameters constant, for example, at a fixed SNR and noise type.

Chong et al [20] investigates whether consonantal and tonal intelligibility of Chinese speech are accounted for in PESQ MOS. Experimental results find that correlation between subjective intelligibility and PESQ scores is rather low, with the highest Pearson correlation at only 0.013. A consonant amplification method which increases the ratio of SNR of consonant-vowel combination is proposed and found to improve correlation. The potential of PESQ in this context is also reported by Manohar and Rao in [93] who investigate speech enhancement in non-stationary noise environments where a post-processing scheme is proposed to improve enhancement due to traditional short-time spectral attenuation (STSA). Intelligibility of the original (signals degraded by factory, machine gun and train noise, without enhancement) and 'enhanced' signals are evaluated using human listening tests, which give 60.7, 51.7 and 50.6 respectively for original signals degraded by factory noise, signals 'enhanced' by Berouti spectral subtraction (BSS) and signals 'enhanced' by BSS plus the proposed post-processing(BSS+PP) respectively. Whilst humans perceive decreasing intelligibility as the signals configuration goes from no enhancement to BSS to BSS+PP, conversely PESQ consistently gives higher scores for BSS signals over the original noisy signals at all SNRs; at lower SNRs PESQ gives the highest scores to BSS+PP signals, which is exactly opposite to the human scores obtained from the listening tests. Therefore Manohor and Rao's studies (2006) seem to suggest that PESQ is wholly unsuitable for intelligibility estimation in the context of degradation caused by enhancement schemes.

Rather than investigating the direct relationship between quality measure scores with human perceived intelligibility, Jose Frage et al [94] and Sun et al [95] investigate the relationship between PESQ and speech recognition rates of an ASR in noisy environments. Sun et al [95] investigate the relationship between PESQ MOS and speech recognition rates using six additive noises at SNRs ranging from 5dB to 25dB and concluded that the latter can be mapped to the PESQ MOS by a simple polynomial rule. Even more impressive is that the ranking of quality of differently degraded signals as estimated by PESQ and ASR seems to agree with each other. For instance, both measures predict white noise degraded signals being the lowest and car noise degraded signals being the highest quality. Jose Fraga et al [94] further investigate this relationship and propose a logistic function that claim to better map PESQ MOS to ASR compared to the simple polynomial fitting proposed in [95]. Both papers are motivated by the fact that ASR performance falls in noisy environments hence a measure to predict its performance is desirable in order to avoid spending time and money on carrying out extensive speech recognition tests. Rather differently, the motivation in this thesis is that if ASR is sufficiently well linked to intelligibility, and if speech recognition can be reliably predicted by quality measures such as PESQ, then by induction theory intelligibility can be estimated by PESQ too.

Apart from PESQ, one other quality measure that is found to have been investigated in this context is WSS. Whilst PESQ gives totally opposite indications in [93] when compared to human intelligibility, the WSS distance measure which is investigated alongside PESQ consistently indicates increase of distortion (i.e., decrease in intelligibility) from original signals with no enhancement to BSS, and from BSS to BSS+PP across all noise conditions considered. This trend agrees with the subjective intelligibility indicated by SUS results, putting this relatively traditional measure in new light as a potential estimator of intelligibility [93]. This finds agreement with works reported in Section 4.3 of this chapter and Chapter 8 in Part II.

No conclusive comment can be made regarding the potential of quality measures for intelligibility estimation since the works reported above are not directly comparable and the amount of literature in this area is rather modest.

4.1.3 Brief Descriptions of the Quality Measures Considered

This section briefly describes the 9 quality measures to be investigated in the experiments in Section 4.2. Again all correlation values quoted in this section refer to correlations between scores given by the objective quality measures and human perceived quality. Note that among the 9 measures: CSNR, SegSNR, MNB and PESQ give scores in terms of quality (used as intelligibility indications in this research) while the other measures give scores in terms of distortion. In the first case: higher scores imply higher quality and lower distortion; whereas in the second case: higher scores imply more distortion and hence lower quality. In this section the '$q_{measure}$' denotation refers to the first case while '$d_{measure}$' refers to the second case. The source codes for CSNR, SegSNR, WSS, LAR, LLR and IS are

obtained from the Robust Speech Processing Laboratory (RSPL)[1].

Classical Signal-to-Noise Ratio(CSNR)

The CSNR measure is simply the power of signal over noise as shown in Equation 4.1. Quackenbush et al [5] quotes correlation at 0.24 for this measure. Though generally agreed to be an incompetent quality estimator, it does give some indication of the quality of stationary, non-adaptive systems [5]. Despite its weak correlation, it remains rather widely used especially for testing of new systems due possibly to its simplicity [70]. The measure is defined as:

$$q_{CSNR} = 10 \log \frac{\sum_n S_\Phi^2(n)}{\sum_n [S_d(n) - S_\Phi(n)]^2} \tag{4.1}$$

where $S_\Phi(n)$ is the reference speech, $S_d(n)$ the processed version, and n is the sample index.

Segmental Signal-to-Noise Ratio(SegSNR)

The SegSNR is similar to the CSNR measure except that SNR is calculated over each short segments of the speech and summed over all segments for the final quality score. The measure has proved to perform better than the classical SNR (CSNR) [96, 97]. A correlation of 0.77 is quoted from Quackenbush's studies [5]. The measure is defined as

$$q_{SegSNR} = \frac{10}{M} \sum_{m=0}^{M-1} \log \frac{\sum_{n=N_m}^{N_m+N-1} S_\Phi^2(n)}{\sum_{n=N_m}^{N_m+N-1} [S_d(n) - S_\Phi(n)]^2} \tag{4.2}$$

where M is the number of segments in the speech signal and N is the segment length. In this thesis all segment durations are 30ms with 10ms overlap.

Weighted Spectral Slope (WSS)

WSS by Klatt [75, 76] is based on weighted differences between the spectral slopes in each of 36 overlapping frequency bands. Bandwidth sizes approximate critical bands in order to give equal perceptual weight to each band. One main difference between WSS and other spectral-based measures is that rather than measuring absolute spectral distance per band, a weighted difference between the spectral slopes in each band is used. The measure is designed to be sensitive to differences in spectral peak locations and insensitive to differences in the heights of those peaks as well as in the neighbourhood of spectral valleys or the spectral tilt. Quackenbush et al's [5] thorough investigation into objective

[1] Freeware downloadable from http://clsr.colorado.edu/rspl/rspl_software.html courtesy of the Robust Speech Processing Laboratory (RSPL) in the University of Colorado at Boulder.

assessments concludes that the best predictors are those derived based on auditory criteria; of those, WSS gave the best correlation at 0.74.

In this measure, the spectral slope in each critical band is first computed as follows:

$$S_\phi(k) = V_\phi(k+1) - V_\phi(k)$$

$$S_d(k) = V_d(k+1) - V_d(k) \tag{4.3}$$

where $V_\phi(k)$ and $V_d(k)$ are the original and distorted spectra in decibels, $S_\phi k$ and $S_d k$ the first order slopes of these spectra and k the critical band index. Next, the per-frame spectral distance measure is defined as:

$$d_{WSS}(n) = K_{spl}(K_\phi - K_d) + \sum_{k=1}^{36} w_a(k)(S_\phi(k) - S_d(k))^2 \tag{4.4}$$

where K_ϕ and K_d are related to overall sound pressure level of the reference and processed signals; K_spl is an optimizing parameter to increase overall performance; n the frame index and w_a a weight ranging from 0 to 1 which aims to place emphasis on spectral peaks rather than valleys. The final distortion score is obtained by summing d_{WSS} from all frames. Note that the WSS used in this thesis considers only 25 critical-band filters spanning the 4kHz bandwidth, whereas Klatt's original measure uses 36 filters.

Log Area Ratio (LAR)

The LAR measure is based on dissimilarity of linear predictive (LP) coefficients between reference and processed speech signals. The log-area-ratio parameters are obtained from the P^{th} order LP reflection coefficients for the reference and processed signals. Quackenbush's studies [5] shows that LAR gives the highest correlation at 0.62 among many LPC-based measure. The objective measure is formulated as

$$d_{LAR} = \left| \frac{1}{M} \sum_{i=1}^{M} \left[\log \frac{1 + r_\phi(n)}{1 - r_\phi(n)} - \log \frac{1 + \hat{r}_d(n)}{1 - \hat{r}_d(n)} \right]^2 \right|^{\frac{1}{2}} \tag{4.5}$$

where $r_\phi(n)$ and $r_d(n)$ are the reflection coefficients for the reference and processed signals respectively for frame n.

Log likelihood Ratio (LLR)

LLR is also known as the Itakura distance measure. It is based on the dissimilarity between all-pole models of the reference and processed speech signals. The measure is defined as

$$d_{LLR}(\vec{a}_d, \vec{a}_\Phi) = \log\left(\frac{\vec{a}_d R_\Phi \vec{a}_d^T}{\vec{a}_\Phi R_\Phi \vec{a}_\Phi^T}\right) \tag{4.6}$$

where $\vec{a}_\Phi$ is the LPC coefficient vector $(1, -a_\Phi(1), -a_\Phi(2),.., -a_\Phi(P))$ for the reference speech $x_\Phi(n)$; and $\vec{a}_d$ the LPC coefficient vector for the distorted speech $x_d(n)$p; $\vec{a}^T$ refers to transpose of $\vec{a}$, and R_Φ is the autocorrelation of matrix $x_\Phi(n)$.

Itakura-Saito Distortion Measure (IS)

The difference between IS measure and the Itakura distance measure (LLR) is that IS measure incorporates the gain estimate using variance terms which directly influences how each measure emphasises differences in general spectral shape as opposed to an overall gain offset. IS measure is defined as

$$d_{IS}(\vec{a}_d, \vec{a}_\phi) = \left[\frac{\sigma_\phi^2}{\sigma_d^2}\right]\left[\frac{\vec{a}_d R_\phi \vec{a}_d^T}{\vec{a}_\phi R_\phi \vec{a}_\phi^T}\right] + \log\left(\frac{\sigma_d^2}{\sigma_\phi^2}\right) - 1 \tag{4.7}$$

where $\vec{a}_\phi$ and $\vec{a}_d$ are the LPC coefficient vector for the reference and degraded speech respectively; σ_ϕ^2 and σ_d^2 represent the all-pole gains for the reference and degraded speech frame respectively.

Modified Bark Spectral Distortion (MBSD)

MBSD [78, 80] assumes that speech quality is directly related to speech loudness. The measure involves three major processing steps: loudness calculation, noise masking threshold computation and finally computation of MBSD. The measure first transforms energies to the Bark frequency scale, the Bark coefficients are then transformed to dB in an attempt to model perceived loudness. This transformation process involves main steps of critical band analysis, equal loudness preemphasis and intensity-loudness power law. Finally, a masking threshold is incorporated where distortion below the threshold is excluded from the calculation. MBSD gives correlation at 0.96 when tested on a modulated noise reference unit(MNRU) and a large range of coding distortions [80].

Measuring Normalising Blocks (MNB)

MNB was introduced by Voran [70, 81] in 1995. It is somewhat distinctive from other perceptual-based measures in that it employs a simple perceptual transformation module. A sophisticated cognition

module follows which consists of a hierarchy of measuring normalising blocks for emulating human patterns of adaptation and reaction to spectral deviations that span different time and frequency scales. In [70, 81] this measure is reported to outperform CD, BSD and ITU-T Rec. P.861 (PSQM) with an average correlation coefficient of about 0.97 when tested on 219 different degradation conditions.

Perceptual Evaluation of Speech Quality (PESQ)

PESQ [8] compares two perceptually-transformed signals and generates a noise disturbance value to estimate the perceived speech quality. The measure is standardised as ITU-T Recommendation P.862 in 2001 replacing PSQM (ITU-T Rec. P.861) with an improved time alignment module which makes it more robust for use in real networks with varying delays. Widely accepted as the state-of-the-art for objective speech quality measurement, the PESQ scores are calibrated using a large database of subjective tests and aims to give quality indications which mimic the Mean Opinion Score (MOS) measured using panel tests according to ITU-T P800 [8]. .

4.2 Experimental Setup

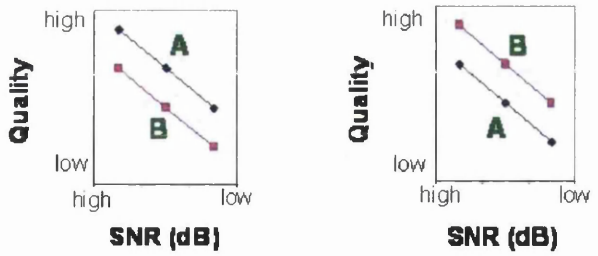
The experiments reported here aim to evaluate how well quality measures estimate comparative intelligibility (see objective (i) in Section 1.1). The usefulness of each measure is determined from correlation analysis on scores given by human and scores given by the quality measures for the same set of test signals. The most widely used correlation method is Pearson product moment correlation which is further described later.

4.2.1 Correlation Analysis

It is sometimes unclear how comparisons between human and objective scores are carried out and it can be difficult to fully comprehend the significance or meaning of the correlation scores given. For instance, Voran [81] evaluates the MNB technique using 9 test sets. 0.928 correlation is reported for Test set 1, which consists of 22 conditions including many types of codecs and 6 levels of modulated noise reference unit (MNRU). The 0.928 correlation reported could be the result of correlating all conditions at once (consider both the varying codecs and varying MNRUs at the same time); or the correlation could be performed among just the varying codecs at fixed MNRU then averaging over all MNRUs to give an overall correlation. These two approaches yield very different outcomes because it is postulated that most if not all objective measures are sensitive towards changes in levels of noise and hence would yield high subjective correlation while assessing degradations of varying MNRUs or SNRs. Therefore by considering both varying codecs and varying MNRUs in one assessment score,

a reasonable portion of the high correlation might be contributed by inter-MNRU correlations rather than inter-codec correlations.

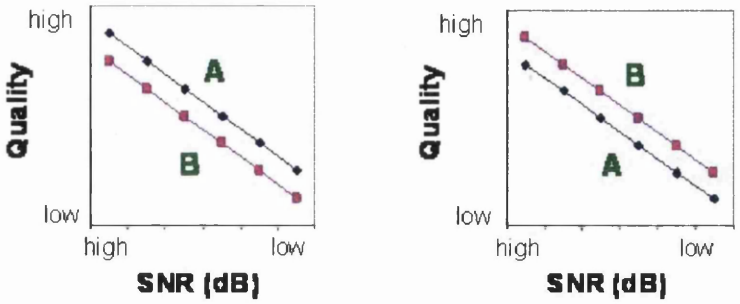
A simple illustrative example is given in Figure 4.1 comparing 2 systems at a range of 3 SNRs. Let Figure 4.1(a) shows human scores and Figure 4.1(b) the objective scores. As shown humans rank system A as having better intelligibility than system B at all SNRs, but the objective measure predicts the opposite. The correlation should be negative if the question askes whether A is more intelligible than B, since clearly the objective measure is consistently wrong at all SNRs. However, Pearson correlation of 0.45 indicates positive correlation. Higher correlation can be obtained if investigation spans even larger range of SNRs. For instance, if the same example shown in Figure 4.1 now spans over 6 SNRs as shown in Figure 4.2(a) and 4.2(b), correlation obtained is now 0.84 even though the quality between system A and system B are still wrongly estimated at every SNR. In short, degradations of the nature of varying MNRUs or SNRs are rather straightforward and not as challenging as degradations of diverse range of different characteristics. Therefore a test set containing large range of degradations of the nature of varying SNRs could lead to artificially high correlation score. In this research such inter-SNR correlations are not considered, instead scores are integrated across the whole SNR range hence producing only one score per degradation type. This is hoped to achieve a more meaningful and realistic evaluation of the measures.



(a) human results (b) objective results

Figure 4.1: An illustration of how misleading high correlation could be obtained if a test set consists of large portion of degradations of the nature of varying SNR. (a) shows human results while (b) shows objective estimates. Objective measure successfully predicts lower quality as SNR decreases as shown by monotonous profiles in (b). However, objective measure deems system B as having higher quality, which is in contrast to human opinions and should lead to poor or negative correlation. Yet positive correlation is obtained at 0.45 due to accurate estimations for inter-SNR.

Two correlation methods are used in the experimental works reported here, namely the more widely used Pearson product-moment correlation (referred in short here as Pearson correlation) and the less commonly used Kendall tau distance (referred in short here as Kendall correlation). Both take in 2 vectors and produce a score reflecting how well the vectors correlate with each other. The first vector is intelligibility ratings given by humans (the so-called ground truth) for a range of differently degraded signals; the second vector is scores given by quality measures for the same signals.



(a) (b)

Figure 4.2: The same as in Figure 4.1 except that larger range of SNRs are considered. The profiles of system A and system B are still wrongly ranked. However, higher Pearson correlation is now obtained at 0.84.

Pearson Correlation

The Pearson correlation is widely used as a performance parameter for the evaluation of speech quality measures. It is defined as

$$r = \frac{\sum_{i=1}^{n}(V1_i - \overline{V1})(V2_i - \overline{V2})}{(n-1)S_{V1}S_{V2}} \tag{4.8}$$

where V1 and V2 are the subjective and objective scores, with means $\overline{V1}$ and $\overline{V2}$, and standard deviation S_{V1} and S_{V2} respectively, while n is the number of degradations considered. The value of the coefficient ranges from -1 to 1 with 1 being the highest-correlated and -1 the opposite.

The primary objective of the research described in Section 1.1 is to determine comparative intelligibility between two signals (which is more intelligible?). Therefore as far as this research is concerned, the key word is ranking. The Pearson correlation is highly sensitive to outliers and would sometimes fail to give correlation that reflects ranking. For instance, assuming V1 = [1 2 3 4 5] and V2 = [5 1 2 3 4], Pearson correlation coefficient is 0 indicating zero correlation between the 2 vectors. However, the last 4 components of both vectors are ranked correctly indicating strong correlation. Nonetheless, given its widespread use, Pearson correlation is employed here along with an alternative correlation approach.

Kendall tau distance

The Kendall tau distance (referred to as Kendall correlation) is a form of Kendall tau rank correlation. It counts the number of pairwise disagreements between 2 vectors. It is also called a bubble-sort distance as it is analogous to the number of swaps that bubble-sort algorithm needs to make in order

to re-arrange one list to be in the same order as the other list. A vector of m components would yield ${}^{m}C_2$ different combinations of pairs. Note that C here refers to a form of permutation where order of element does not matter. Absolute values are disregarded and only rankings are of interest. For instance, V1 of [20, 30, 55, 61, 62] and V2 of [0.2, 0.1, 0.12, 0.9, 0.3] would be transformed into [1, 2, 3, 4, 5] and [3, 1, 2, 5, 4] respectively before pairwise comparisons are carried out. The Kendall tau distance ranges from 0 to 1 and is subtracted from 1 in order to represent pairwise agreements rather than disagreements. Therefore 0 means that none of the pairs agrees in their rankings and 1 means every pair is ranked accordingly. Considering the same example given in Section 4.2.1 where V1 of [1 2 3 4 5] and V2 of [5 1 2 3 4] yield 0 Pearson correlation, Kendall correlation obtained is 0.6 since there are 6 correctly ranked pairs out of the total of 10 pairs (${}^{5}C_2$=10). This correlation method is directly relevant to the objective of the thesis namely comparative intelligibility. Note that since comparisons are carried out on paired data, guessing would have achieved 50% correlation. Therefore a score of 0.5 is by chance and a score of 1 is 100% correlation, i.e., the ranking is perfectly correct.

4.2.2 Databases

Databases used in testing of objective quality measures are the same databases described in Section 3.2 (Details available in Appendix A.1). Here Table 4.2 briefly describes the databases.

Type	Test Set	Descriptions
Additive	$DS1_{add}$	additive noises of diverse characteristics including both speech-like and the more stationary ones.
	$DS2_{add}$	additive noises, most fairly stationary.
Coding	$DS3_{cod}$	car noise and tandeming of single coding scheme
	$DS4_{cod}$	various $DS1_{add}$ noises and tandeming of mixed coding schemes
Enhancement	$DS5_{enh}$	car noise and different speech enhancement processes
	$DS6_{enh}$	various $DS1_{add}$ noises and different configurations of the NLSS process

Table 4.2: Brief descriptions of the 6 test sets, i.e., DS1 to DS6.

4.2.3 Procedures

All 9 objective measures considered here are based on a so-called intrusive approach in that a reference signal is needed in order to compute intelligibility difference between the reference and test signal. References used were the corresponding clean signals [2]. Intelligibility associated with a particular degradation at a particular SNR is the mean score across all 566 signals for that SNR. Those results given in terms of distortion indication, namely IS, LAR, LLR, WSS and MBSD are inverted to reflect

[2] taken from the Aurora2 database, they are the original 20kHz signals from the TIDigits downsampled to 8kHz and filtered with the G.712 characteristic [26].

level of intelligibility rather than lack of intelligibility. Inversion is done simply by subtraction from the maximum score obtained from signals in the same test set.

Correlation analysis is performed for every test set by comparing a subjective vector with a corresponding objective vector. A subjective vector contains a list of intelligibility opinions given by human assessment for signals in the test set; the objective vector contains scores given by a chosen objective quality measure for the same signals. The subjective and objective vectors are deduced in two different ways from every test set:

Approach_I: Averaging across the whole SNR range (integrated approach)

The first approach integrates scores across the range of SNRs. Therefore objective intelligibility associated with a particular degradation is the mean score across all 566 signals at every SNR and subsequently averaged across all SNRs considered. Subjective vectors are deduced in the same manner.

Approach_II: at fixed SNRs (thresholded approach)

Taking the average across a large range of SNRs hides detailed information such as crossing-over of profiles at a particular SNR region. The second approach takes 3 sets of scores at 3 fixed SNRs. The SNRs are chosen such that the subjective scores correspond approximately to 75%, 62.5% and 50% intelligibility respectively (equivalent to 3, 2.5 and 2 digits correctly recognised out the 4 digits in an utterance), as illustrated in Figure 4.3, using human results of $DS1_{add}$ as an example. 3 objective vectors are also obtained at SNRs a, b and c. Then correlation is performed for the 3 pairs of subjective and objective vectors. The final correlation is then average of the 3 correlations.

In conclusion, 3 correlations are to be computed for every test corpus. The correlations are:

(i) Pearson: computed by correlating subjective and objective vectors obtained with Approach I using Pearson correlation;

(ii) Kendall_1: computed by correlating subjective and objective vectors obtained with Approach I using Kendall correlation; and

(iii) Kendall_2: computed by correlating subjective and objective vectors obtained with Approach II using Kendall correlation.

Though integration in Approach_I tends to hide details at instantaneous SNRs, it can be seen as computing the overall performance which might be more reliable given that understandably the human scores may not be 100% reliable and accurate (crossovers might not be meaningful). On the other hand,

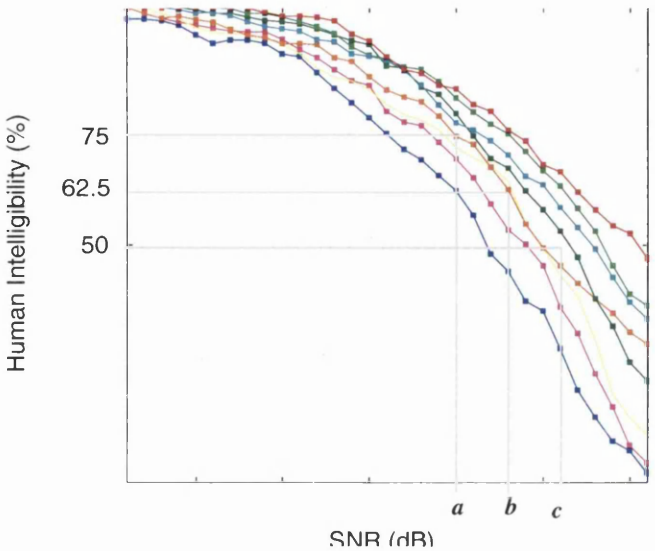


Figure 4.3: An illustration of how score vectors are deduced for computation of correlation using Approach_II. Example shown is for $DS1_{add}$, profiles are human scores for the different degradation conditions. SNRs corresponding to approximately 50%, 62.5% and 75% subjective intelligibility are determined. Then correlations are performed on vectors of subjective and objective scores taken at these SNRs. Final correlation is the mean of the 3 correlations.

Approach_II is subjected to the reliability and accuracy of the human scores at each instantaneous SNR. It is desirable that the correlations computed using both approaches agree at least in trends (which objective measure might be better). Correlation performance of the objective measures in Part I are mainly presented in the form of Kendall_1, while the corresponding Kendall_2 and Pearson correlations are shown in Appendix C. Correlations obtained using Approach_II are computed mainly for the purpose of comparing the performances of existing measures in Part I and the new system proposed in Part II, which is a data-driven system that understandbly requires more test samples in order for the performance score to be more statistically significant.

4.3 Results and Discussion

As is discussed in Section 2.1 and the beginning of this chapter, quality and intelligibility have different operational ranges and hence measuring intelligibility using quality measures would most probably result in scores that cover only a relatively small section of its full dynamic range. Prior to presenting the Pearson and Kendall correlations for the different measures and test sets, it is interesting to see how far apart the operational ranges of these measures are from that of human intelligibility.

4.3.1 Intelligibility Response of Quality Measures

Figure 4.4 plot quality scores against SNRs for a car noise-degraded signal using the 9 measures. SNRs considered span over the wide range of 50dB to -20dB to observe the changes of the quality scores from region where signals are totally intelligible (50dB) to region where signals are totally unintelligible (-20dB). All scores are obtained from one test utterance which is a male-spoken digit string saying 'four-six-two-five' (4625). As usual those results given in terms of distortion indication (scores given by

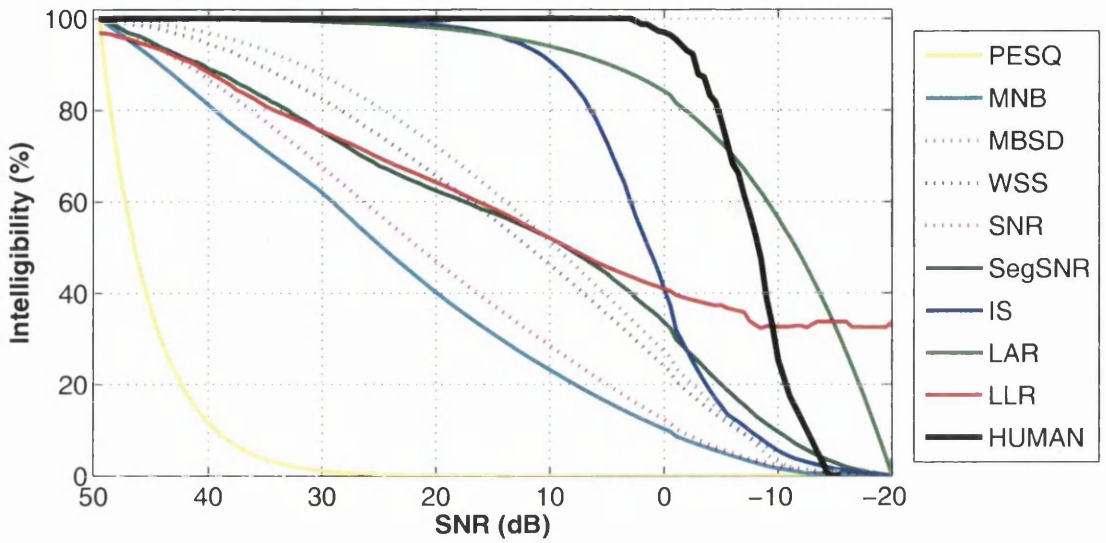


Figure 4.4: A comparison of the intelligibility response of the quality measures. Black profile is human intelligibility for car noise degraded signals plotted against SNRs; other profiles are normalised objective scores given by the quality measures. All profiles decrease monotonically as SNRs decreases. The dynamic intelligibility region (where intelligibility begins to fall from 100% down to 0%) is at approximately 5dB to -10dB in this example. The scoring range of different measures at this region differs.

IS, LAR, LLR, WSS and MBSD) are inverted to an intelligibility indication simply by subtracting the scores from respective maximum score. Normalisation is then performed by scaling the scores to 0% and 100% to enable side-by-side comparison. In all figures the black bold profiles are the corresponding human scores (data taken from Chapter 3). The other profiles are produced by respective quality measures for the same signal. Human profiles for regions below -10dB are estimated because it is not advisable to play such highly degraded signals to human listeners due to health concern.

One thing to take note is that among all measures only PESQ and MNB give scores in fixed range, i.e., 0 to 1 for MNB and -0.5 to 4.5 for PESQ. The rest of the measures give scores in 'free' scale where a low quality (intelligibility) signal and a high quality (intelligibility) signal could score as low and as high as possible. For presentation purpose it is simply assumed that at 50dB, the signal is of excellent quality hence a measure should give the maximum score of its dynamic score range; similarly at -20dB, the signal is of very poor quality hence a measure should give its minimum possible score. The scores obtained at these two SNRs are taken as the dynamic range of scores for that particular measure, corresponding to 0% and 100% on the y-axis. Therefore when CSNR scores seems to have saturated from 30dB onwards in Figure 4.4, this does not mean that constant scores are obtained throughout the regions where the profile is seemingly saturated, but simply means that the rate of changes of scores at high SNR region is much greater compared to the changes at low SNR region.

Several observations can be made from Figure 4.4. First of all, though at different rates, all

measures seems to give monotonous response against changes in SNR, as expected. One exception to this is perhaps the PESQ profile which shows slight ripple at -5dB downwards. Next, notice that while human intelligibility remains at 100% throughout the region of 50dB to around 0dB despite decreasing SNRs, all measures with the exception of perhaps MBSD start giving decreasing scores as SNR decreases. At around 0dB where human intelligibility begins to degrade from 100% (threshold of intelligibility), all measures apart from MBSD indicate intelligibility of less than 50%. In the shaded region where human intelligibility changes dynamically from 100% to 0%, the score range of most measures span less than 50%, for instance, MBSD's score range span approximately from approximately 50% to 85%.

MBSD is more sensitive to changes at lower rather than higher SNR region. The exact opposite is CSNR which is more sensitive to changes in high SNR region and relatively insensitive at low SNR region with a seemingly saturated profile at lower SNRs. As a whole both SNR-based measures are the measures that give the lowest profiles in the figure; while two out of the three perceptual based measures, namely MBSD and MNB give the two highest profiles. In fact MNB is the only profile that seem to exhibit the s-curve trend of the human profile, that is, giving consistently high scores (notionally 100%) at high SNR region where there is little change in intelligibility, despite decreasing SNRs; scores then fall off rather quickly during the dynamic intelligibility region; followed by constant zero scores at very low SNRs where signals are unintelligible. Unlike MNB, the spectral-based measures give a near linear decreasing response as SNR decreases, showing no such s-curve trend. PESQ gives a near linear response too at SNRs above 0dB but starts to saturate at around after 0dB. Notice that PESQ saturates not at 0% but at around 30%, which is equivalent to about 1.11 raw PESQ score out of the scale of -0.5 to 4.5. This is perhaps because PESQ is not meant to operate at such low SNRs.

The same figure is now produced for signals degraded by coding distortion. The signal is firstly degraded by car noise then en-decoded by MELP coder twice. Comparing Figure 4.4 with Figure 4.5 notice that most profiles are now no longer monotonic due to the non-linear degradation introduced by MELP. Again MBSD is the exception because of the smoothness of its profile. Other reasonably smooth profiles include the spectral-based measures IS, LAR and LLR. PESQ profile is somewhat smooth until 0dB after which the scores swing wildly. Another observation is that while human intelligibility still ranges from 0% to 100%, the dynamic range of the quality measure scores has shrunk (apart from MBSD). For example, the maximum PESQ score is only around 60% eventhough the signal is perfectly intelligible at 50dB. The same goes with MNB, whose maximum score is only around 70 to 80%. The scores for WSS and both SNR-based scores at those region are also greatly reduced. This is not ideal but is expected since the measures are essentially measuring quality and there is obvious quality reduction due to the en-decoding process eventhough a 50dB signal is understandably intelligible.

Viewing purely from the examples given here, it interesting to observe that in overall the SNR-based measures consistently give the lowest profiles; followed by the spectral-based measures; and lastly the perceptual-based measures (with the notable exception of PESQ) always gives the highest profiles.

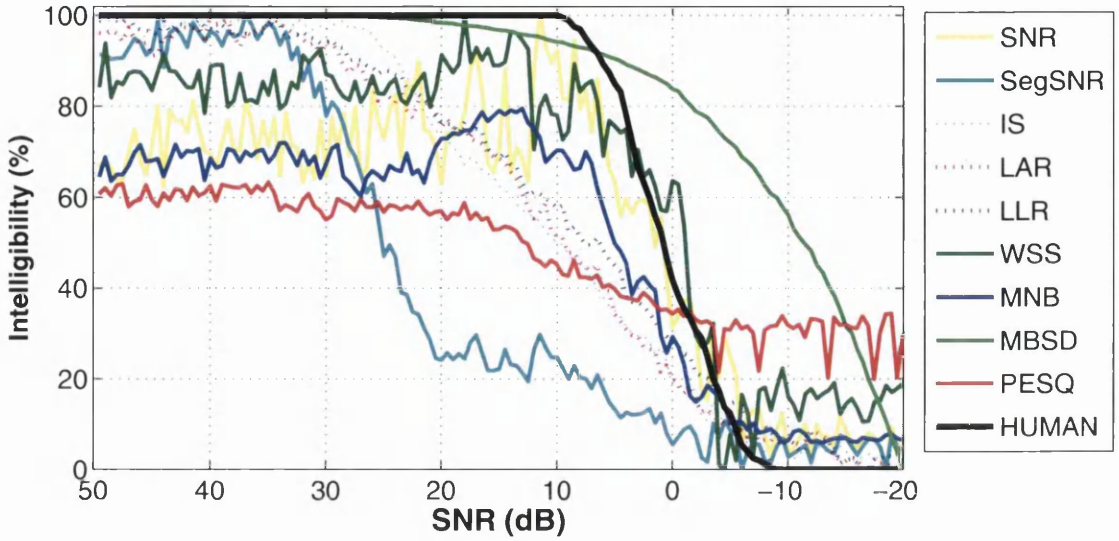


Figure 4.5: The same as in Figure 4.4 except that the signals are now degraded by car noise followed by twice en-decoding of MELP codec. Notice that most objective profiles are no longer smooth due to the non-linear distortion. The scoring range of the quality measures are decreased not only at dynamic intelligibility region but also at higher SNRs, for e.g. at 50dB PESQ's maximum scores is only 60% of its full scale though humans deem the signal as intelligible.

MBSD is the only measure considered here giving scores in relatively large dynamic score range in low SNRs. However, compared to other measures MBSD seems to be less sensitive towards degradations of different nature with little change in its scores as the degradations changes from pure car noise in Figure 4.4 to the severe 'car noise plus endecodings' in Figure 4.5. PESQ gives a very noisy profile for this arbitrary choice of non-linear degradation. At low SNRs where human intelligibility is zero, PESQ continues to give fluctuating and possibly meaningless scores while scores from most other measures have already saturated.

Obviously, none of these observations is directly conclusive of the potential of the quality measure in intelligibility assessment. A measure giving scores in more constrained ranges of intelligibility may, but not necessarily lead to poorer performance in intelligibility estimation, and vice-versa. One concern is that under high noise condition if the scores are saturated then the measure could not differentiate low quality signals with lower quality signals. Another concern is that under high noise fluctuating scores such as that of PESQ in Figure 4.5 might be obtained. However, if a measure remains sensitive (not fully saturated) and reliable (not fluctuating) even at this very end of its scoring range, then its potential in intelligibility assessment is a matter of how closely related quality and intelligibility are, in other words, how representative the quality score is of intelligibility. As mentioned, no conclusive comments regarding the potential of the measures in intelligibility assessment can be made yet. The real potential is to be judged by looking further at their correlation with human scores.

4.3.2 Correlations

Table 4.3 and 4.4 present Kendall$_1$ and Pearson correlations computed for all 9 quality measures and 6 test sets. The Kendall$_2$ which is not shown here are always slightly less than Kendall$_1$ but show the same trend of performance (included in the Appendix C). Recall that the test sets can be divided into 3 types based on the degradation involved: environmental or additive noise (DS1$_{add}$ and DS2$_{add}$), coding schemes (DS3$_{cod}$ and DS4$_{cod}$) and enhancement processes (DS5$_{enh}$ and DS6$_{enh}$). Again, bear in mind that since Kendall correlation relates to pairwise comparison, pure wild guessing would give a 0.5 hence any correlation below 0.5 means that it is worse that guessing. Pearson correlations are presented here for comparison purpose since it is commonly-used in assessments of quality measures.

	Waveform		Spectral				Perceptual		
	CSNR	SegSNR	IS	LAR	LLR	WSS	MNB	MBSD	PESQ
DS1$_{add}$	0.60	0.60	0.32	0.25	0.24	0.72	0.4	0.32	0.21
DS2$_{add}$	0.32	0.41	0.74	0.32	0.63	0.57	0.71	0.76	0.86
DS3$_{cod}$	0.88	0.91	0.75	0.86	0.91	0.93	0.9	0.89	0.87
DS4$_{cod}$	0.51	0.53	0.63	0.44	0.57	0.67	0.61	0.59	0.49
DS5$_{enh}$	0.61	0.6	0.6	0.6	0.56	0.75	0.54	0.46	0.54
DS6$_{enh}$	0.36	0.33	0.4	0.61	0.55	0.68	0.45	0.37	0.40
Average	0.55	0.56	0.57	0.51	0.58	0.72	0.60	0.57	0.56

Table 4.3: Kendall$_1$ Correlations (range from 0 to 1) obtained for the six test sets using the 9 quality measures.

Looking across the measures, some of the observations are:

- the perceptual measures such as PESQ, MNB and MBSD do not seem to give better performance. On the contrary, in some cases such as DS1$_{add}$ and perhaps DS5$_{enh}$ the primitive SNR-based measures seems to show more correlattions than the perceptual-based measures.

	Waveform		Spectral				Perceptual		
	CSNR	SegSNR	IS	LAR	LLR	WSS	MNB	MBSD	PESQ
DS1$_{add}$	0.33	0.33	-0.69	-0.82	-0.76	0.65	-0.32	-0.7	-0.74
DS2$_{add}$	-0.46	-0.03	0.71	-0.04	0.5	0.09	0.57	0.76	0.79
DS3$_{cod}$	0.81	0.79	0.52	0.69	0.74	0.82	0.88	0.77	0.79
DS4$_{cod}$	0.33	0.29	-0.21	0.18	0.2	0.04	0.18	-0.38	0.3
DS5$_{enh}$	0.16	0.27	0.54	0.28	0.24	0.64	0.4	-0.08	0.4
DS6$_{enh}$	-0.31	-0.39	-0.13	0.34	0.11	0.38	-0.11	-0.38	-0.12
Average	0.14	0.21	0.12	0.21	0.17	0.44	0.27	-0.01	0.24

Table 4.4: Pearson Correlations (range from -1 to 1) obtained for the six test sets using the 9 quality measures.

- WSS gives outstanding correlation at 0.72 for $DS1_{add}$ which considers background noises of diverse characteristics including the speech-like and the more stationary ones. PESQ gives a mere 0.21 for the same test set.

- However, it is interesting to note that the outcome for $DS2_{add}$ is rather opposite. All three perceptual measures give relatively high $Kendall_1$ correlations for this test set. PESQ in particular gives 0.86. The WSS on the contrary obtains only 0.57 though is still slightly better than guessing. The reasons for this discrepancy across the two test sets (despite both belonging to background noises category) could be that $DS1_{add}$ contains more varieties of degradations: for instance, noises such as babble is more speech-like, almost like multiple speakers; airport, train station, and restaurant are also somewhat speech-like though 'speech' occurs less frequent than in babble noise; meanwhile subway noise is adverse yet periodic, street noise is impulsive (occasional car alarms, vehicles and pedestrians passing by); lastly car noise and exhibition noise are fairly stationary. Noises such as the speech-like babble noise is 'confusing' as the degraded signals could appear reasonably clean though components crucial to recognition of the words are damaged. Most measures obtain poor correlations because they deem babble noise degraded signals as having higher intelligibility as they may appear less noisy than signals degraded by car-noise for instance, which seems to corrupt the whole bandwidth. On the other hand, all noises under $DS2_{add}$ are rather stationary, even the largecrowd noise does not contain distinctive speech-like noise but more like stationary noise faded into the background. Hence it is postulated that for $DS2_{add}$ there might be less difference between the impact that the noises have on quality and on intelligibility. This perhaps explain the better performance of the perceptual-based measures as they are known to be better quality measures.

- All measures obtain good correlations for $DS3_{cod}$ with the lowest at 0.75 by IS and the highest at 0.93 by WSS. The perceptual-based measures obtain better correlations than the spectral-based ones (apart from WSS) in general, but it seems that the simple SNR-based ones are just as reliable in this context with SegSNR giving 0.91 correlation. High correlations obtained for this test set could be due to the fact that most degradations consists of tandeming of the same codec, for instance, LPC en-decoded once, twice and thrice. The same way that the measures are generally sensitive to changes in SNR or MNRU, it is thought that they are sensitive to changes in the number of tandeming too. Worth noting is the exceptional good performance of WSS over that of perceptual measures at 0.93.

- In contrary to $DS3_{cod}$, all measures obtain rather poor correlations for $DS4_{cod}$. This is perhaps expected since the degradation context here offer no such 'bonus points' given by tandeming of the same codec. Again the best correlation is reported by WSS at 0.67.

- As expected poor correlations are obtained for $DS5_{enh}$ and $DS6_{enh}$ since speech enhancements are meant to enhance quality scores but could have reverse effect on intelligibility. All three

perceptual measures obtain poor correlations for both $DS5_{enh}$ and $DS6_{enh}$, in fact poorer correlations are reported for $DS6_{enh}$ where degradations involved different configurations of one type of enhancement technique, namely the non-linear spectral subtraction (NLSS). $DS6_{enh}$ in particular serves as an acid test for the measure as it mimics real life application where different configurations and parameters of a new system under development needs to be evaluated in search for optimal performance.

- WSS is again the exception giving reasonably good correlations for the enhancement test sets compared to other measures, with 0.75 Kendall_1 is reported for $DS5_{enh}$ and 0.68 for $DS6_{enh}$.
- The best measure in overall is the spectral-envelope-based WSS with 0.72 average Kendall_1 correlation as well as the highest average Pearson correlation at 0.44, while PESQ the state-of-the-art gives 0.56 and 0.24 respectively.

4.4 Overall Observations

Figure 4.6 is reproduced from Figure 4.4 showing human and PESQ scores for car noise degraded signals over a large SNR range. One of the first observation is that human intelligibility response exhibits a distinctive s-curve trend where, despite increasing degradation, signal intelligibility remains at the notional 100% prior to reaching the threshold where intelligibility begins to be threatened; and remains at 0% once pass the threshold of unintelligibility since once rendered unintelligible, more degradations do not degrade the signal further. However, the response of most quality measures towards increasing degradations seems to be largely linear and show no obvious sign of stationary regions at 100% or 0% intelligibility, as illustrated by the PESQ profile in Figure 4.6.

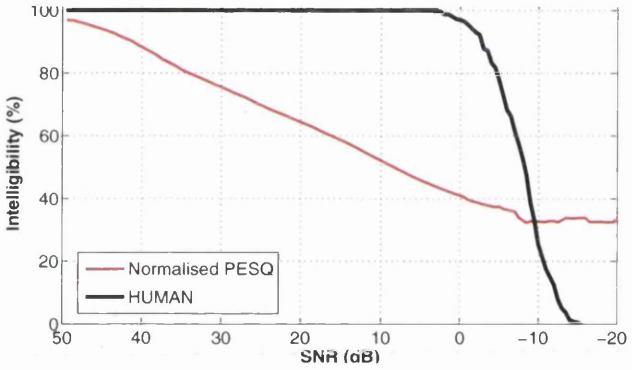


Figure 4.6: A comparison of the intelligibility response of human and quality measures in general.

In overall, all quality measures with the notable exception of WSS, give scores that correlate poorly with human intelligibility. Poor correlations are especially obvious when dealing with enhancement processes and challenging environmental noises such as speech-like noises. In the first case, enhancement processes tend to improve quality hence quality scores but degrade intelligibility; in the second cases, signals degraded by speech-like noises are often deemed to be of good intelligibility although such noises are in fact more impairing than the more stationary ones. As a whole, among all measures considered here WSS seems to be the most promising intelligibility assessor with overall Kendall_1 correlation at 0.72. It also appears as the only measure that is able to distinguish the intelligibility of signals degraded by challenging, speech-like noises in DS1_{add}. More impressive is its correlations for the enhancement test sets at 0.72 on average. These observations compliment the findings of Manohar and Rao [93].

One other interesting observation is that the primitive SNR-based measures are not as poorly correlated as thought would be based on poor comments regarding the measures in the literature [5]. In fact, in many categories such as DS1_{add} their correlations are higher than those of PESQ. Lastly, poor correlations of the perceptual-based measures are perhaps because they are designed and optimised for measurement of the more general quality and hence perform less well in the more extreme circumstances when intelligibility is at threshold and quality is low.

Chapter 5

ASR System for Intelligibility Assessments

Chapter 4 considers intelligibility assessment using measures intended for quality assessment. The approach is motivated by the ready availability of objective quality measures; also the observation that intelligibility is an integral part of quality. However, the only notable performance is 0.72 Kendall_1 correlation obtained with Weighted Spectral Slope (WSS), all other measures give poor correlations including the state-of-the-art PESQ at 0.56. Perhaps the idea of estimating intelligibility through such quality measures is too idealstic.

While a poor quality speech needs not be of poor intelligibility, a word that fails to be recognised must be. Similarly, a word that can be successfully recognised is by definition of good intelligibility. Afterall Fletcher [21] the founder of Articulation Index (AI) defines intelligibility as the probability of correct recognition of words. This suggests that word recognition rather than overall quality, is more closely related to intelligibility. In this sense an automatic speech recogniser (ASR) emerges as the obvious potential solution for intelligibility assessment not only because mature ASR technology is readily available but most importantly because word recognition is the fundamental task of ASR systems. In this context Odette [18] states that ASR is closely related to human speech recognition since the central issue of both is word recognition.

Following the above postulations, this chapter considers ASR in the context of intelligibility assessment. The objective is to assess how well ASR scores correlate with human intelligibility where correlation is associated with trend rather than absolute values. First of all, the motivations as well as some underlying problems are presented in Section 5.1; Section 5.2 presents a review on related work in the literature while the remaining sections report the experimental work describing the correlation experiments equivalent to those presented in Chapter 4. Note that for direct comparison purposes the databases and correlation approaches are the same as in Chapter 4.

5.1 Motivations

The main motivation for the use of ASR in intelligibility assessment is obvious, namely word recognition is intelligibility. As discussed in Section 4.4 and as illustrated in Figure 5.1, three regions can be identified forming an s-curve: (i) the wholly intelligible region where signal intelligibility remains at the notional 100% despite increasing degradation; (ii) a dynamic region where intelligibility falls from

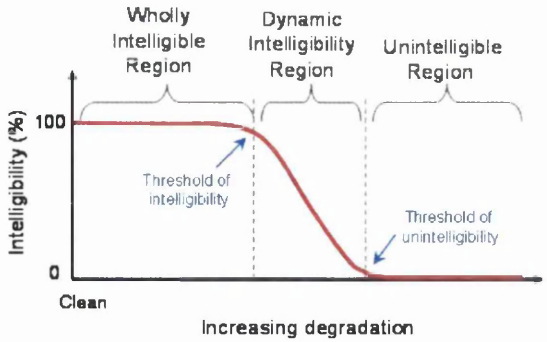


Figure 5.1: An illustration of the human intelligibility s-curve. From left-to-right: (i) Wholly Intelligible Region: notionally 100% intelligibility despite increasing degradation; (ii) Dynamic Intelligibility Region: intelligibility falls from 100% towards 0%; (iii) Unintelligible Region: intelligibility remains 0% despite increasing degradation.

100% towards 0%; and (iii) unintelligible region (i.e., 0% intelligibility) where signal is already rendered unintelligible and more degradation does not degrade its intelligibility further. Here it is interesting and somewhat promising to observe that the ASR word accuracy also exhibits this similar s-curve trend. Figure 5.2 shows human and ASR word accuracy for car-noise degraded signals over a large SNR range. As shown on the same graph the PESQ profile shows no distintive stationary regions at 100% and 0%. On the other hand, the ASR and human profiles portray similar s-curve trends, though with a distinct SNR offset of about 15dB.

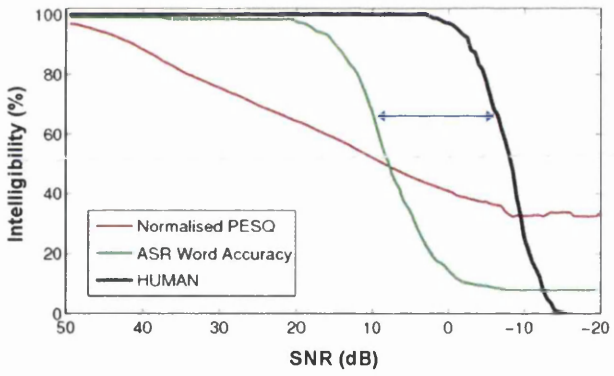
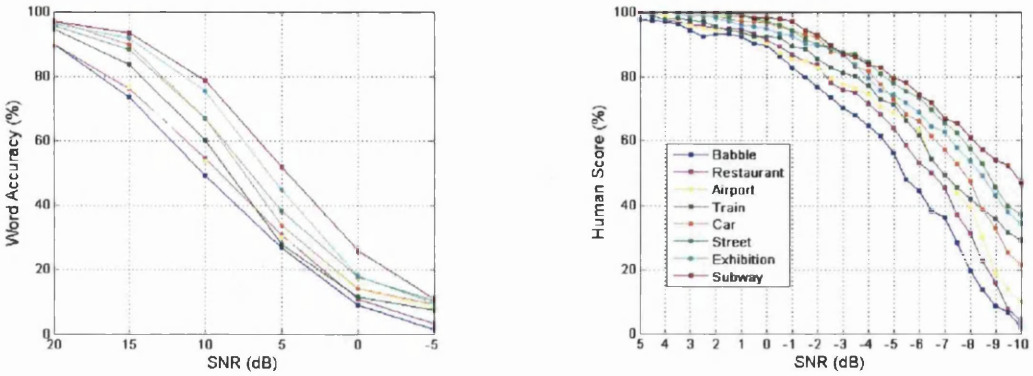


Figure 5.2: Figure shows scores obtained from human, ASR and PESQ for car noise degraded signals over the SNR range of 50dB to -20dB. Notice that PESQ profile seems to decrease almost linearly as SNR decreases along the x-axis; both human and ASR profiles exhibit similar s-curve trend though with a dB offset of about 15dB (indicated by arrow).

Figure 5.3 show profiles of word accuracy for the Aurora2 test set side-by-side with profiles of human scores for test set $DS1_{add}$. Note that both test sets consider the same 8 noise types but at different SNR ranges: Aurora2 test set at 20dB to -5dB; $DS1_{add}$ at 5dB to -10dB. As shown the two sets of profiles show striking resemblance both in terms of trend and ranking. Notice that in both figures the subway profile is the highest and babble the lowest, whereas all quality measures (with the exception of WSS) rank these 2 conditions incorrectly. These preliminary observations indicate ASR's potential in intelligibility assessment, certainly for these additive noise cases.

Whilst the motivations for use of ASR in intelligibility assessment is obvious, the challenges may be equally obvious. Notice that the PESQ and ASR profiles in Figure 5.2 aproach saturation at about -8dB and -5dB respectively (ASR saturates at about 8% rather than 0% due possibly to the random guessing option in the ASR scoring process). These two levelling trends occur in the dynamic intelligibility region defined in Figure 5.1 and thus indicate some cause for concern. Nonetheless, it is



(a) ASR word accuracy

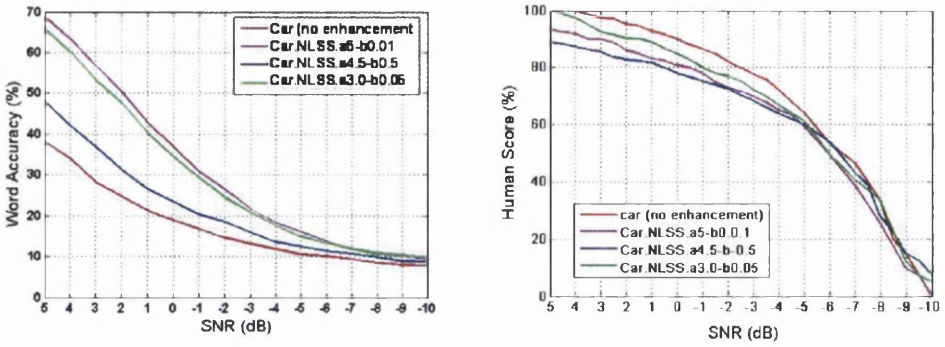
(b) Human

Figure 5.3: A comparison of intelligibility response of human and ASR. Figure (a) shows ASR word accuracy for the Aurora2 test set (data taken from [98]) where SNR ranges from 20dB to -5dB; (b) shows human scores for test set $DS1_{add}$ where SNR ranges from 5dB to -10dB. Notice the resemblance between the two sets of profiles in terms of trend and ranking.

the correlation with trends that is important and this is assessed in this chapter. Therefore so long as the profiles ranking are in order then the ASR word accuracy is deemed useful as an intelligibility measure.

Moving from additive noise to enhancement systems presents yet another challenge, namely the discrepancies caused by enhancement processes which improve ASR scores without actually improving intelligibility. Figure 5.4 shows ASR word accuracy and corresponding human scores for a subset of conditions from test set $DS6_{enh}$ which considers non-linear spectral substraction (NLSS) processs. The conditions chosen for illustration are car noise degraded signals and three NLSS-processed versions. As shown, while humans deem all three processed versions to be of lower intelligibility, the ASR word accuracy indicates otherwise, and the correlation (profiles ranking) seems poor.

Such challenges are encountered in Chapter 4 when quality measures assessed. Though the same problems seem to remain here when ASR is assessed for the same task, it is anticipated that better correlation might exist given the close link between intelligibility and word recognition; as well as the positive remarks gathered from literature, which would be reviewed in the next section.



(a) ASR word accuracy (b) Human

Figure 5.4: Figure (a) show word accuracy for car noise degraded signals and 3 NLSS-processed versions; Figure (b) shows corresponding human scores. While humans deem all 3 NLSS-processed versions as of lower intelligibility than the car noise degraded version, ASR word accuracy indicates otherwise, indicating potential for poor correlation.

5.2 Background

5.2.1 Brief Introduction of ASR Technology

ASR is a form of data-driven classification (DDC). It recognises spoken sounds by comparing an observed speech signal pattern to signal patterns previously learnt during training. Like all DDC it consists of the general stages of feature extraction, modelling, and search mechanism. In the so-called feature extraction or ASR front-end, the continuous speech waveform is segmented into frames in the order of typically 10-30ms. Each short-term segment is transformed into a vector of parametric representations called features. Typical parametric representations are smoothed spectra or linear prediction coefficients plus their variants. The sequence of discrete parameter vectors then are used to form acoustic models for each recognition unit, commonly phoneme or whole word. Acoustic modelling is most commonly realised with the use of hidden Markov Models (HMM) [99, 100] which can be seen as a finite-state machine, viewing each short-term speech segment as stationary. A sequence of HMM states model the signal variations along the time course. In each state a statistical distribution called a mixture of diagonal covariance Gaussian will give a likelihood for each observed feature vector. Given sufficient training data then a HMM can be constructed which implicitly models all of the many sources of variability inherent in real speech. The Viterbi algorithm is generally the underlying technique used for decoding the observed vectors. It finds the best path through the HMM states leading to a decision of the decoded recognition unit.

5.2.2 Related Work

The idea of using ASR to estimate intelligibility is relatively recent and related work in the literature is limited. Unless otherwise stated, all correlation values quoted in this section refer to Pearson correlation between ASR and human intelligibility scores.

One among the very few examples found is the work reported by Hicks et al [24] where an ASR system is used to assess intelligibility of speech corrupted by stationary noises and co-channel speech (two people talking at the same time). An objective method is needed to compare the performance of various reconstruction algorithms, aiming to reconstruct the speech of each speaker first by detecting usable segments from the co-channel speech. The TIDigits digit strings database is used for simplicity. Results shows good correlation between human and ASR as both give decreasing recognition scores as the level of interference increases. Furthermore, results seem to suggest that there exists a consistent offset of approximately 18dB between human word recognition and ASR word recognition (here we show a similar offset of approximately 15dB with digit recognition in car noise, see Figure 5.2). For instance, human target-to-interferer ratio (TIR) tests give 54% recognition rate at -12dB TIR while the ASR systems yield approximately the same recognition rate at 6dB, which is an offset of 18dB from -12dB; again, humans give 73% recognition rate at -6dB TIR and the ASR system gives similar recognition at 12dB, which is also an offset of 18dB. This suggests that ASR system could be a reliable detector of usable speech segments in co-channel speech and eventually be used to evaluate the effectiveness of reconstruction algorithms. Furthermore, the almost-consistent offset seems to imply possible estimation of absolute intelligibility.

In 1999, Chernick et al [22, 101] investigated the potential of ASR as a quality and intelligibility estimator in the context of signals degraded by DoD-CELP codec with bit errors. Unfortunately, the term 'quality' and 'intelligibility' are used inter-changeably in the paper. Specifically, the authors might be looking for a quality score that reflects intelligibility at the same time. Tests are conducted on the TIMIT database. Machine scoring is based on phoneme recognition but it is not very clear how intelligibility is scored by humans, presumably on a scale of understability, for example, 'understandable and good', 'understandable but poor quality', etc. As the bit error rate rises from 0.1% to 5%, human-derived scores range from 'understandable and good' to 'practically unintelligible'. The reported correlations are 0.816 for the recogniser with matched speakers training and 0.745 for an unmatched recogniser. The paper concludes that the correlations are sufficiently encouraging and further investigation are warranted. However, strangely no further reported work with new result can be found since this original publication in 1999.

In 2002, Jiang and Schulzrinne [23] also investigated the potential of ASR as a MOS[1] and intelligibility predictor but this time in the context of signals degraded by a G.729 codec over packet loss rates at 0, 2, 5, 10 and 15% respectively. Human listening tests are conducted where the listeners

[1]MOS refers to Mean of Opinion Score, which is a scale of 1 to 5 used for measurement of quality.

are played a test sentence and asked to rate the quality in term of MOS as well as to transcribe the sentence. The latter is to measure speech intelligibility by obtaining human-based word recognition which is then compared with machine-based word recognition. The paper concludes that ASR word recognition score is a reliable predictor of MOS. As for intelligibility correlation, the paper concludes that human and machine-based word recognition are correlated though the relationship is not linear. One interesting point to note is that when packet loss probability increases from 10% to 15%, word recognition by humans is found to fall only slightly, much less when compared to the fall caused by changes of packet loss probability at other lesser rates, for example, from 0% to 2%. The paper explains that this is because signal intelligibility at 10% packet loss rate is already very much degraded hence 15% packet loss rate does not make the signal much less intelligible, presumably already approaching the end of the intelligibility s-curve. However, machine-based word recognition scores are found to continue falling with almost the same rate as at 0% packet loss.

A more comprehensive study is published very recently in 2006 by Teng and Kubichek [40]. Three types of speech recognition systems are assessed for their ability to predict human intelligibility for a variety of medium to low bit-rate codecs with channel impairment. Speech data used are consonant-vowel-consonant (CVC) word pairs with confusable leading consonants. Evaluation is performed on recognition results using three approaches: (i) word recognition; (ii) phoneme recognition and (iii) phoneme group recognition. During word recognition, the word network for the recogniser is restricted to the pair of words under test in order to mimic human listening test. The phoneme group approach uses phoneme recognition obtained from each phoneme category (for eg, stops, affricatives, etc) as a metric to predict intelligibility. Higher correlation is reported for vocoders compared to waveform coders, with correlations ranging from 0.80 to the perfect 1.00. Despite these apparently rosy presentations, a close look finds that whilst high correlations are reported for comparisons within the same type of codec across different rates of bit errors, inter-codec comparisons are not well reflected in the machine scores. For instance, humans rate the two waveform coders, namely G.711 and G.726 as having higher intelligibility than the three vocoders at all rates of burst bit errors; however, both word recognition and phoneme recognition by ASR systems indicate the exact opposite. Besides, humans rate random bit error as having significantly more adverse effects on intelligibility of LPC10 signals compared to burst bit error, yet machine word recognition implies negligible difference between the two. Nonetheless, ASR is given positive verdict in the paper as an intelligibility predictor.

As a whole, relatively little work has been published on the application of ASR systems for intelligibility assessments. All related works discussed in this section investigates intelligibility assessment in the context of just one type of degradation over varying degrees. For instance, Chernick et al [22, 101] looks into signals degraded by CELP codecs over varying degrees of bit errors; Jiang and Schulzrinne [23] look into a G.729 codec with varying degrees of packet loss. It has not been possible to find reported work on ASR systems in assessing intelligibility across different types of degradations. Besides, all the reported works reviewed here consider only clean-trained ASR and investigate only the potential of word accuracy, though there are other ASR statistics such as deletion and insertion

errors that might also prove indicative of human intelligibility. The remaining sections in this chapter investigate the potential of various ASR statistics in intelligibility assessment in the context of various degradation types.

5.3 ASR with Clean Training

There are obviously many ways to train an ASR system. Here, baseline clean training seems to be a sensible first step and is in line with the works of others [22,24,40]. Furthermore, a clean-trained ASR provides common treatment towards all kinds of differently degraded test data. This subject about the make-up of the training data is considered again later in the chapter.

5.3.1 ASR Configuration

Unless stated otherwise, the European standard ETSI W1007 front-end [102] and Aurora2 HTK reference recogniser [26,98] are used for all experimental work reported in this chapter. The recogniser is a hidden Markov model (HMM) based word recogniser designed using the HTK toolkit. An HMM model is built for each digit with simple left-to-right topology. There are 16 active states per word and mixture of 3 components per state. The features consist of 12 cepstral coefficients, log energy plus the corresponding delta and double delta coefficients forming a 39th order vector for each 25ms frame with 10ms overlap.

The clean training set used here is defined by the Aurora2 framework [98]. The database is a subset of the TIDigits which are downsampled to 8kHz and subsequently filtered with G.712 characteristic without noise added. It contains the recordings of 55 male and 55 female adults. In total there are 8440 digit-strings the length of which ranges from 1 to 7 digits. A relatively low recognition rate is expected since the training and testing are severely mis-matched in that the train data are clean while test data are heavily corrupted. Nonetheless, the interest of this research is correlation in ranking rather absolute values or matched recognition rates.

5.3.2 ASR Statistics

Recognition performance of an ASR system is often measured by the word accuracy. It is defined as the total number of tests minus all recognition errors, namely deletion, substitution and insertion. Another less commonly used measure is the percentage correct, defined as the total number of tests minus deletion and substitution. All these recognition statistics be it scores of recognition success (such as word accuracy and percentage correct) or scores of recognition error (such as substitution, deletion

and insertion) could be indicative of intelligibility. The definition and the scenarios under which each score is incurred are as follows:

(1) Insertion (Icns): falsely recognise a non-existing word

(2) Deletion (Del): fail to detect and recognise an existing word

(3) Substitution (Subst): mis-recognise a word as a different word

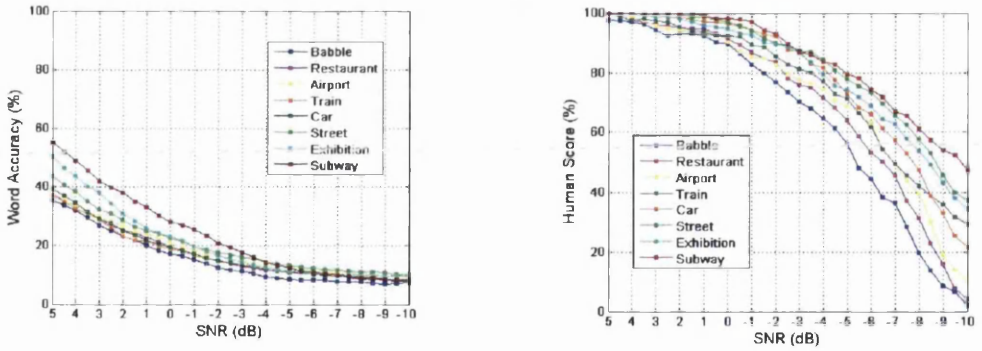
(4) Percentage Correct (Corr):

$$\frac{\text{Total tests - Del - Subst}}{\text{Total tests}} \tag{5.1}$$

(5) Word Accuracy (WordAcc):

$$\frac{\text{Total tests - Del - Subst - Ins}}{\text{Total tests}} \tag{5.2}$$

5.3.3 Intelligibility Correlations



(a) Human scores

(b) Human scores

Figure 5.5: (a) shows ASR WordAcc for the 8 conditions under $DS1_{add}$; (b) shows corresponding human scores for the same data set and noise conditions.

The objective of the experimental work reported in this chapter is to assess the correlation between ASR scores and human scores for given set of conditions. For instance, the 8 profiles in Figure 5.5(a) are ASR word accuracy for the 8 conditions considered in test set $DS1_{add}$; while Figure 5.5(b) shows the corresponding human profiles. The aim is to assess how well these two sets of profiles agree with each other in terms of ranking. The same correlation approaches used in Chapter 4 are repeated here where the scores for each condition are integrated across the SNR range. Hence a vector of 8 components is deduced from each set of profiles shown in Figure 5.5. The Kendall correlation is then computed to evaluate the ranking correlation between the two vectors; in addition Pearson correlation is computed to reflect the overall correlation. The $DS1_{add}$ example shown in Figure 5.5 yields 0.92 Kendall_1 and

0.88 Pearson correlation which compares favourably with the 0.72 Kendall_1 and 0.65 Pearson obtained with WSS (the best quality measure as suggested by findings in Chapter 4).

Though word accuracy is the most commonly employed measurement for the performance of ASR, in the context of intelligibility assessment all other recognition statistics could also be indicative of intelligibility. Here the above experiment is repeated using all 5 ASR statistics, namely Word Accuracy (WordAcc), Percentage Correct (Corr), Substitution (Subst), Insertion (Ins) and Deletion (Del). Note that recognition *errors* are inverted in order to indicate recognition success (hence intelligibility). The resultant Kendall_1 correlations are shown in Table 5.1.

ASR Training: Clean						
Category	Test Set	ASR Statistic				
		WordAcc	Corr	Del	Subst	Ins
Environmental	DS1_{add}	0.92	0.36	0.24	0.84	0.76

Table 5.1: Kendall_1 Correlations obtained for test set DS1_{add} using all 5 ASR statistics namely Word Accuracy (WordAcc), Percentage Correct (Corr), Deletion (Del), Substitution (Subst) and Insertion (Ins).

Notice that the correlations obtained with different ASR statistic varies significantly. For instance, Word Accuracy (WordAcc), Substitution and Insertion yield relatively good correlations at 0.92, 0.84 and 0.76 respectively; whereas Percentage Correct (Corr) and Deletion yield poor correlation at 0.36 and 0.24. This is thought to be due to the characteristics of the degradations under consideration. Recall that test set DS1_{add} considers the 8 noise types of diverse characteristics which include the more speech-like (babble, airport and restaurant); the more stationary (car and exhibition); periodic (subway) as well as impulsive ones (street). While humans deem signals degraded by speech-like noises as the least intelligible, almost all quality measures investigated in Chapter 4 predict the exact opposite possibly because unlike noises that are more stationary, speech-like noises do not corrupt the whole spectrum hence signals may seem less degraded hence are thought to be more intelligible). Here 3 ASR statistics namely WordAcc, Substitution and Insertion give relatively good correlations for this particular test set. The reasons for the good and bad of the correlation performance of each ASR statistic are postulated as follows:

- Substitution gives good correlation because speech-like noises could cause spectral distortion which affects pronunciation of words; this leads to substitution errors when words are mistaken as other words.
- Insertion errors are incurred when speech-like noises are mistaken as speech. Good correlation is obtained with Insertion since higher degradation leads to more insertion errors.
- Deletion error are rarely incurred in the context of speech-like noises. The correlation is weak since higher degradation leads to more insertion and substitution errors rather than deletion error.

- Percentage Correct gives poor correlation at 0.36 while WordAcc gives good correlation at 0.92. This could be due to the exclusion of insertion as an useful intelligibility assessor in the computation of Percentage Correct (see Equation 5.1).

The above reasonings for the correlation performance shown in Table 5.1 suggests that each recognition statistic is in fact linked to human speech recognition. Different statistics correlate differently depending on the nature of the degradations considered. Now the experiment is extended to other test sets namely $DS2_{add}$, $DS3_{cod}$, $DS4_{cod}$, $DS5_{enh}$ and $DS6_{enh}$ (See Section 3.2 or Appendix A for descriptions). The resultant correlations are shown in Table 5.2 and the corresponding bar plot is given in Figure 5.6. The highest correlations achieved by the quality measures for each test set are plotted alongside for comparison. For example, the highest correlation achieved in Chapter 4 for $DS1_{add}$ is given by Weighted Spectral Slope (WSS) at 0.72 (light blue bar at leftmost). Note that correlation value below 0.5 is worse that guessing hence is considered as negative, and vice-versa. The followings can be observed from Table 5.2 and Figure 5.6:

(1) Highest correlation is given by Word Accuracy (WordAcc) for $DS1_{add}$ at 0.92. This marks significant improvement when compared to that given by WSS at 0.72. However, though both $DS1_{add}$ and $DS2_{add}$ considers additive, environmental noises, the correlation obtained with WordAcc for $DS2_{add}$ is much lower at 0.68.

(2) Conversely, Percentage Correct (Corr) gives very poor correlation for $DS1_{add}$ at 0.36 but relatively good correlation for $DS2_{add}$ at 0.82.

(3) Both WordAcc and Corr give positive correlations for the coding test sets namely $DS3_{cod}$ and $DS4_{cod}$ with averages of 0.81 ($DS3_{cod}$: 0.91; $DS4_{cod}$: 0.70) and 0.73 ($DS3_{cod}$: 0.86; $DS4_{cod}$: 0.60) respectively. Both indicators give lower correlations for $DS4_{cod}$ compared to $DS3_{cod}$, presumably because $DS4_{cod}$ which considers degradations introduced by tandeming of mixed codecs is more challenging than $DS3_{cod}$ which considers only tandeming of a single codec. Worth noting is the correlation for $DS4_{cod}$ by WordAcc at 0.70 which is higher than that achieved by any quality measure in Chapter 4 as illustrated in Figure 5.6.

(4) Deletion gives negative correlation for $DS1_{add}$ at 0.24 but relatively good correlation for $DS2_{add}$ at 0.76. This is due to the different nature of the noises considered in these 2 test sets. $DS1_{add}$ consists of speech-like noises such as babble, airport and restaurant; meanwhile the $DS2_{add}$ noises are fairly stationary. In the first case, the recogniser is prone to mistaking noise as speech hence higher degradation leads to more insertions and substitutions, not deletions. Therefore predictably deletion is not a useful indicator here since less deletion does not imply higher intelligibility. In the second case, higher degradation leads to masking of the signal which may incur firstly substitution then eventually deletion when the signal is totally masked beyond recognition. Therefore, deletion is deemed more relevant in the context of $DS2_{add}$ but less relevant in the context of $DS1_{add}$.

ASR Training: Clean						
Category	Test Set	ASR Statistic				
		WordAcc	Corr	Del	Subst	Ins
Additive	$DS1_{add}$	0.92	0.36	0.24	0.84	0.76
	$DS2_{add}$	0.68	0.82	0.76	0.42	0.32
Coding	$DS3_{cod}$	0.91	0.86	0.50	0.82	0.43
	$DS4_{cod}$	0.70	0.60	0.49	0.60	0.48
Enhancement	$DS5_{enh}$	0.57	0.67	0.58	0.44	0.35
	$DS6_{enh}$	0.53	0.45	0.34	0.64	0.54
Average		0.72	0.63	0.48	0.63	0.48

Table 5.2: $Kendall_1$ Correlations obtained for the six test sets. Note relatively good correlations obtained with Word Accuracy (WordAcc).

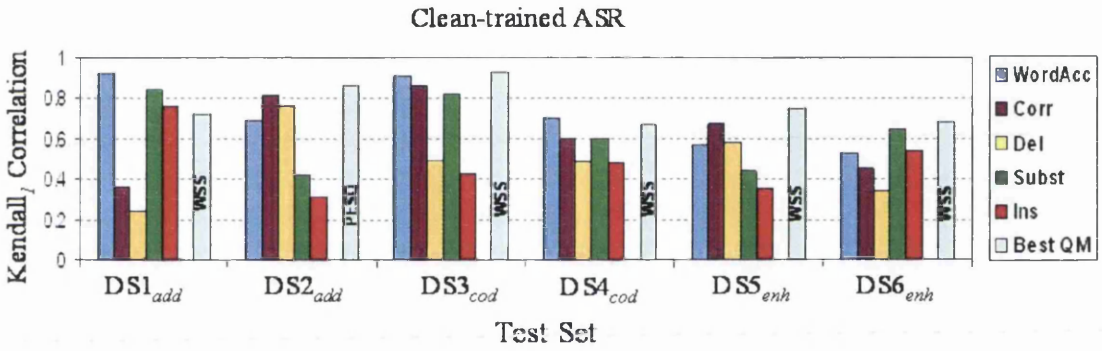


Figure 5.6: Bar plot showing $Kendall_1$ correlations obtained with the 5 ASR statistics produced by the clean-trained ASR system. Labelled bars refer to the best-performing quality measure for each test set.

(5) Both Substitution and Insertion are potentially good intelligibility indicator for $DS1_{add}$ with correlation at 0.84 and 0.76 respectively. This is presumably because the speech-like noises in $DS1_{add}$ lead to substitution and insertion errors when noises are mistaken as speech. The same indicators give negative correlations for $DS2_{add}$ at 0.42 and 0.32 respectively since substitution and insertion errors are less relevant in the context of stationary noises considered in D_{add2}.

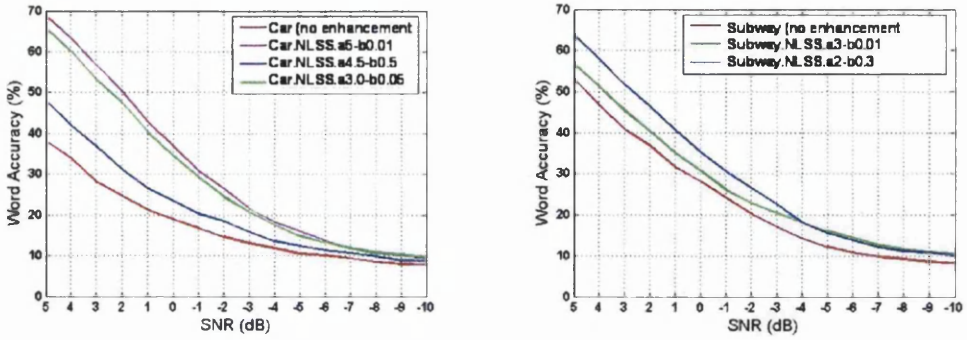
(6) The absence of Insertion in the computation of Percentage Correct (Corr) (see equation 5.1) could possibly explain the poor correlation given by Corr for $DS1_{add}$ at 0.36. Similarly, the inclusion of insertion in the computation of word accuracy explains the good correlation given by WordAcc at 0.92.

(7) Among 3 recognition errors Substitution is the best correlator for the coding test sets namely $DS3_{cod}$ and $DS4_{cod}$ with joint $Kendall_1$ correlation at 0.71 (D_{cod1}: 0.82; D_{cod2}: 0.60). Insertion is less useful here perhaps because the corrupting noise considered in $DS3_{cod}$ is car noise which is not speech-like hence does not incur insertion errors; meanwhile Deletion is also less useful

perhaps because the SNR ranges considered for the coding test sets are relatively high (at -5dB to 10dB, compared that of $DS1_{add}$ at -10dB to 5dB), therefore it has not reached the stage where signals are totally masked by noise and failed to be detected by the recogniser. It is thought that in this case the recogniser always detects the signals (i.e., low deletion) but might fail to identify the signals (i.e., substitution) due to increasing non-linear distortions introduced by the coding algorithms that effectively cause pronunciation of the words to change.

(8) Apart from Substitution and Corr which might be useful for $DS5_{enh}$ and $DS6_{enh}$ respectively, all indicators correlate poorly with the enhancement test sets namely $DS5_{enh}$ and $DS6_{enh2}$. For example, the correlation obtained with WordAcc are merely 0.57 and 0.53, which are less than those of WSS at 0.75 and 0.68 respectively (see Figure 5.6). This is perhaps because the enhancement processes are optimised using ASR during their development stages. They are also competing against other processes using improvement of ASR scores as the yardstick of usefulness. In other words, many such systems are designed to boost ASR scores and, in doing so, whether intelligibility is improved or not is irrelevant. Figure 5.7(a) and 5.7(b) illustrate this using two subsets of signals from $DS6_{enh2}$. Figure 5.7(a) shows word accuracy of car noise degraded signals and three versions of NLSS-processed signals (the same car noise-degraded signals as are for the car profile but further processed by various NLSS configurations); Figure 5.7(b) shows that same for signals corrupted by subway noise and two NLSS-processed versions. In both cases humans deem the original noisy signals (without NLSS) as more intelligible. However, in both figures the ASR Word Accuracy profiles for these signals (car profile in Figure 5.7(a) and subway profile in Figure 5.7(b)) are shown to be the lowest, while the NLSS-processed signals are deemed to be more intelligible. This indicates that WordAcc and ASR statistics in general are not well correlated with human scores when considering degradations introduced by such enhancement processes.

In summary, despite simple clean training, WordAcc seems to be a potential indicator of intelligibility especially for the additive or environmental noises test sets ($DS1_{add}$ and $DS2_{add2}$) and coding test sets ($DS3_{cod}$ and $DS4_{cod2}$). WordAcc is also a more consistent indicator considering that the usefulness of other ASR statistics such as Deletion, Substitution and Insertion depend very much on the nature of the degradations being dealt with. For instance, Insertion is more suitable for assessment involving speech-like noises, Deletion for stationary noises, and Substitution is for non-linear distortions introduced by coding algorithms which do not necessarily make the signals sound noisier but could change the pronunciation of the speech signals leading to substitution errors. An independent work that further investigates the potential of WordAcc in intelligibility assessment has been carried out by a contemporaneous research colleague, Mr K. Jellyman. Findings of the investigation would be briefly presented in the next section.



(a) (b)

Figure 5.7: An interpretation of why low correlations are obtained for the enhancement test sets namely $DS5_{enh}$ and D_{enh2}. Both figures show word accuracy versus SNRs: (a) car noise degraded signals and three 'enhanced' version using NLSS; (b) subway noise degraded signals and two 'enhanced' versions. In both cases humans deem original noisy signals (red profile) as more intelligible than the 'enhanced' versions, but ASR WordAcc shows the opposite.

5.3.4 Further Investigations on WordAcc

This section reports on an independent work carried out by a fellow research colleague, K. Jellyman, that investigates the potential of Word Accuracy (WordAcc) from a clean-trained ASR for intelligibility assessment. The motivation is to find a reliable intelligibility assessor for testing of an in-house image-based enhancer in search for optimal configurations. This serves as a real-world scenario of a new system under development. Here firstly brief introduction of the system is given, followed by experimental results.

The enhancer is based on a morphological spectrogram segmentation procedure [34, 103–105] which divides the speech spectrogram into regions deemed likely to contain deterministic signal information and regions of noise. Two parameters under investigation are the sizes of the segmented shapes which are controlled by spectrogram filter sizes and are changed in an attempt to optimise intelligibility. Some examples are shown in Figure 5.8 where Figure 5.8(a) shows the original noisy spectrogram of a 4-digit utterance and Figure 5.8(b) shows the segmentation outcome where regions deem to contain deterministic signals are shown in red, whilst the complement is shown in blue. Signal regions are is set to one while noise regions to zero so that when multiplying with the original spectrogram, noise is suppressed and enhancement is achieved. The size of this segmented shape can be expanded or reduced horizontally and / or vertically in an attempt to optimise signal intelligibility. Shown in Figure 5.8(c) and (d) are examples of manipulation of shape size where Figure 5.8(c) illustrates reduction of the original shape while Figure 5.8(d) shows expansion.

Here the speech enhancer is assessed under 2 chosen noises, namely babble -2.5dB and subway -5.0dB the SNRs of which reflect approximately the region of 75% intelligibility based on the humans

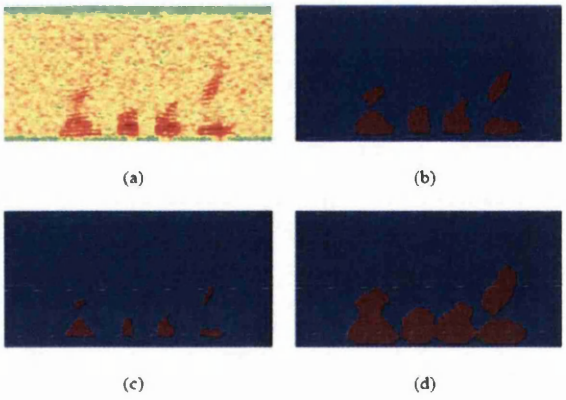


Figure 5.8: Spectrogram segmentation examples with (a) 5 dB AWGN SNR noisy original spectrogram, (b) Spectrogram segmentation binary map result, (c) Horizontal and vertical reduction of spectrogram map by 5 pixels, (d) Horizontal and vertical expansion of spectrogram map by 5 pixels. Horizontal axis (0-1.5 seconds, 1-180 pixels), vertical frequency axis (0-4 kHz, 0-129 pixels).

opinions. Objective measures considered are WordAcc and PESQ. Figure 5.9 shows the subjective intelligibility results alongside objective results for both noises. (note that the PESQ score has been directly mapped to a percentage range). Both subjective tests are shown to peak at a filter size of 0 (i.e. no change in original segmentation) with approximately 80% for babble and 83% for subway. The peak value in the case of babble could be interpreted as part of a large flat peak with little difference between filter sizes -2 to 2.

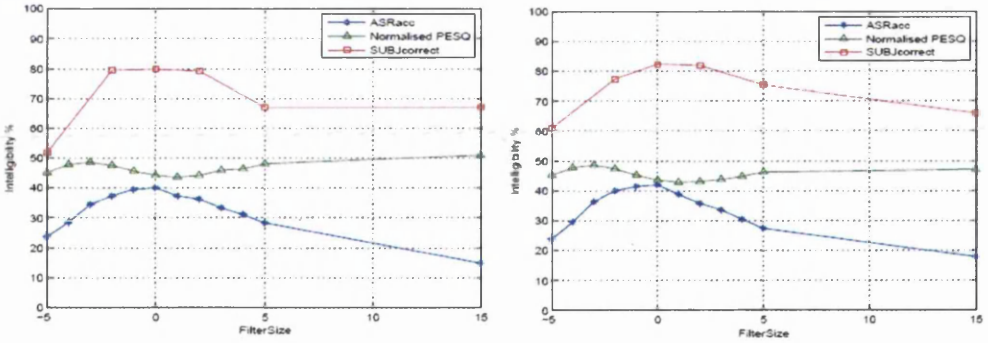


Figure 5.9: Objective and human scores versus filter sizes for (a) babble noise degraded speech at -2.5dB, (b) subway noise degraded speech at -5.0dB.

Notice that the peaks of the WordAcc profile are also seen to be at a filter size of 0, correlating remarkably well with human scores. As the filter size is decreased or increased from this optimum size the predicted intelligibility of the WordAcc measure is seen to decrease for both noise conditions. Though this observation does not quite match the human curve which is seen to flatten out beyond filter size of 5 for babble noise condition, it correlates promisingly with the subjective curve of subway noise condition. In contrast the PESQ measure on the other hand is seen to predict almost the reverse to the subjective intelligibility results. In the case of both noise conditions PESQ dips to its lowest point around the filter size of 0 and 1. Furthermore, PESQ curve for the babble noise condition is seen to increase beyond the filter size of 5 though human deem those signals as having constant intelligibility.

In short, WordAcc has been shown to be a good intelligibility estimator in the degradation context considered here. These findings suggest that ASR-based scores do not necessarily correlate poorly with intelligibility of signals processed by enhancement algorithms. This again support the idea that no measure works universally well. It is emphasised that before choosing a measure for system evaluation, the suitability should be assessed and confirmed first with subjective results.

5.4 ASR for Noisy Conditions

Whilst ASR systems perform reasonably well under relatively clean, controlled and matched train-test conditions, as soon as they are deployed in more hostile environments, for example, at SNRs approaching 0dB, their performance tends to deteriorate. In Teng and Kubichek's studies [40] where an ASR system is investigated for its potential in intelligibility estimation, one further research proposed is to improve recognition performance using a model adaptation approach. It is of the opinion that by improving the recognition rate, the speech recogniser will better mimic human performance hence give better correlation to subjective intelligibility [40]. It is interesting to investigate this postulation here, bearing in mind that it is the correlation of ASR scores and human scores that is critical, not the absolute performance.

5.4.1 Missing Data Techniques

One approach to improve ASR peformance for corrupted speech is the so-called 'missing data' techniques. The approach is motivated by the assumption that some spectral-temporal regions of a degraded signal remain uncorrupted [106,107] . Therefore the approach is to classify features as 'missing' or 'reliable' (hence the name 'missing data') and then use only features from those less corrupted, more reliable regions for recognition. Speech recognition with missing data techniques is expected to be more like human and hence more robust, after all, we humans are able to perform 'auditory scene analysis' (ASA) to pick out and pay selective attention to individual sound sources [108] while ignoring sounds that are not informative.

Here whole word digit models are trained using the Aurora clean training set as in Section 5.3. The auditory scene analysis is performed using Barker's Computational Auditory Scene Analysis (CASA) Toolkit (CTK). Instead of passing all features of the test data to the recogniser, they are firstly classified as reliable or unreliable (missing) components. The recogniser is supplied with the features as well as a time-frequency mask reflecting the locations assessed as reliable or unreliable parts of the features. Two system variants are investigated here: (i) a discrete mask is used where each element of the mask is labelled as either 0 (unreliable component) or 1 (reliable component), features having a local SNR greater than 7dB are deemed reliable [106]; (ii) a fuzzy mask is used where the each element of the mask is a real value between 0 and 1 reflecting the reliability/usability of the corresponding features.

The mel-frequency cepstrum coefficient (MFCC) which most traditional ASR systems are based on is not approapriate for recognizing separated sounds from simultaneous speech signals, hence note that the features used here is mel-scale filter bank coefficients instead of MFCC.

5.4.2 Multi-condition (Mixed) Training

Apart from the missing data approach, an even more common approach to have noise-robust ASR is the use of multi-condition training, which means training on mixtures of clean and degraded speech (also called mixed training). The motivation for mixed training is to again investigate the possibility of improving intelligibility correlation by improving recognition rate of the ASR system.

The Aurora2 framework [98] has two training sets: the first set consists of 8440 clean utterances selected from training part of the TIDigits (this training set is used in section 5.3); the second set is the same 8440 utterances but are splitted into 20 subsets and are corrupted by 4 different noises at 5 SNRs. Specifically, the noises are babble, car, exhibition and subway; and the SNRs are clean, 20, 15, 10 and 5dB. Better recognition rates are anticipated especially for the environmental test sets (i.e., $DS1_{add}$ and $DS2_{add}$) since training and testing are now better matched compared to the case of clean training in Section 5.3. Obviously, improvement of recognition rates would be more apparent when training and testing are done in the same degradation conditions; equally, such improvement is less likely when dealing with a different degradation condition to that seen during training.

5.4.3 Results and Discussion

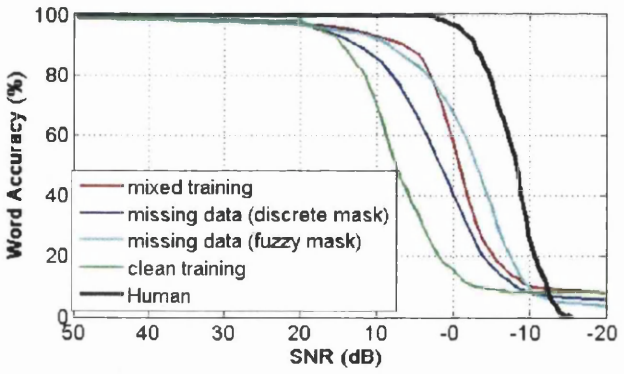


Figure 5.10: A comparison of word accuracy profiles from clean-trained, mixed-trained and missing data systems; and the corresponding human derived profile.

Figure 5.10 shows word accuracy profiles from clean-trained, mixed-trained and missing data ASR systems for a set of car noise degraded signals, and the corresponding human profile. Note that the clean-trained and human profiles shown are the same as in Figure 5.2. As expected the recognition rates are significantly improved with both missing data and mixed training approach. In all 3 cases (i.e., missing data with discrete mask, missing data with fuzzy mask and mixed training) the improvement in recognition rate over that of the clean-trained system is obvious. Notice that whilst the gap between the clean-trained ASR profile and human profile is relatively large (approx. 15dB), the profiles of the missing data sys-

tems and the mixed-trained system are approaching the human profile; and the gap between ASR and human effectively decreases. It is possible that these improved recognition rates could lead to improvement in intelligibility correlations. In particular better correlations are anticipated from recognition statistics given by the missing data systems since the approach attempts to mimic the sound separation abilities of human listeners hence should potentially better reflect humans' judgement. Section 5.4.3.1 and 5.4.3.2 report the intelligibility correlations obtained with recognition statistics from the missing data systems and the mixed-trained system respectively.

5.4.3.1 Correlations with Missing Data ASR Systems

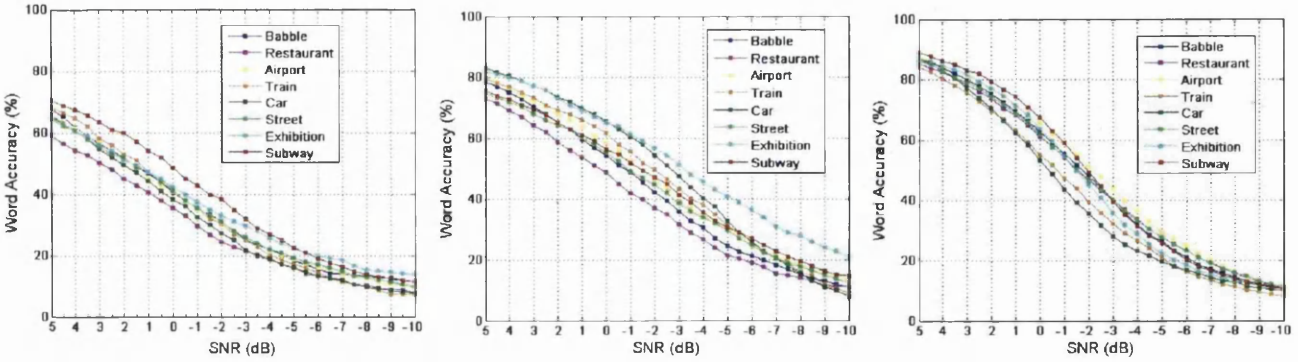
Correlations for the missing data ASR system are shown in Table 5.3 together with corresponding results from clean-trained ASR taken from Section 5.3. Despite the impressive increment in recognition rates, disappointingly the correlations achieved with the statistics coming from the missing data ASR systems are worse than those obtained in when the simple clean training is employed. As shown the average correlations obtained with WordAcc are only 0.48 and 0.52 for the use of discrete and fuzzy mask respectively; poor correlations are also reported by percentage Correct at 0.50 and 0.53 respectively. Both are inferior compared to 0.72 and 0.63 achieved by the same statistics from the clean-trained system. Furthermore, there seems to be no meaningful difference between the correlations obtained when discrete or fuzzy mask is used, though the latter gives bigger improvement in recognition rates as illustrated in Figure 5.11.

Despite the extended dynamic score range, intelligibility correlation has not benefited and possible reasons for these poor correlations are unclear. Obviously the mask could be manipulated in attempt to obtain not only higher recognition rates but scores that reflect intelligibility. However, this idea is not examined further here.

5.4.3.2 Correlations with Mixed-Trained ASR System

Table 5.4 shows Kendall_1 correlations obtained with the 5 recognition statistics from the mixed-trained system (ASR with multi-condition training). The corresponding bar plot is shown in Figure 5.12 where the best correlations achieved with statistics from the clean-trained ASR in Section 5.3 are plotted alongside for easy comparison (light blue bars). Several observations can be drawn from Table 5.4 and Figure 5.12:

(1) The recognition statistic giving the highest overall correlation in Section 5.3 (clean training) is Word Accuracy (WordAcc) with average correlation at 0.81 if the enhancement test sets are excluded. However, WordAcc from the mixed-trained system here gives relatively poor correlation at the average of 0.57.



(a) (b) (c)

Figure 5.11: ASR Word accuracy profiles for test set $DS1_{add}$ produced by (a) mixed-trained system, (b) missing data system with discrete mask, and (c) missing data system with fuzzy mask.

(a)

ASR with missing data					
		Discrete mask		Fuzzy mask	
Category	Test Set	WordAcc	Corr	WordAcc	Corr
Environmental	$DS1_{add}$	0.71	0.61	0.57	0.64
	$DS2_{add}$	0.73	0.69	0.76	0.73
Coding	$DS3_{cod}$	0.72	0.64	0.78	0.77
	$DS4_{cod}$	0.25	0.23	0.26	0.25
Enhancement	$DS5_{enh}$	0.35	0.41	0.41	0.42
	$DS6_{enh}$	0.12	0.39	0.36	0.37
Average		0.48	0.50	0.52	0.53

(b)

ASR with clean training	
WordAcc	Corr
0.92	0.36
0.68	0.82
0.91	0.86
0.70	0.60
0.57	0.67
0.53	0.45
0.72	0.63

Table 5.3: Kendall_1 correlations obtained with recognition statistics from the missing data system (ASR with missing data techniques applied). The first 2 result columns are obtained when the discrete mask is used, the last 2 columns when the fuzzy mask is used. Table to the right shows correlations from the same statistics given by the clean-train ASR system investigated in Section 5.3.

(2) the best overall correlations are obtained with Substitution and Insertion at 0.64 and 0.61 respectively. Especially worth noting are correlations obtained for the enhancement test sets at 0.70 ($DS5_{enh}$: 0.65; $DS5_{enh}$: 0.75) and 0.71 ($DS5_{enh}$: 0.67; $DS5_{enh}$: 0.75) respectively on average. This contrasts the poor correlations obtained with the same statistics given by the clean-trained system for these two test sets ($DS5_{enh}$: 0.44 (Subst) and 0.35 (Ins); $DS6_{enh}$: 0.64 (Subst) and 0.54 (Ins) in Table 5.2). As illustrated in Figure 5.12, both statistics show much better correlations for these two enhancement test sets. For instance, the correlations obtained with WordAcc in the same context are significantly lower at 0.30 and 0.60 respectively.

(3) Possible reasons for the superior performance of Substitution and Insertion for the enhancement

ASR Training: Mixed						
Category	Test Set	ASR Statistic				
		WordAcc	Corr	Del	Subst	Ins
Additive	$DS1_{add}$	0.60	0.48	0.40	0.68	0.52
	$DS2_{add}$	0.33	0.64	0.83	0.17	0.17
Coding	$DS3_{cod}$	0.88	0.85	0.17	0.91	0.94
	$DS4_{cod}$	0.71	0.64	0.41	0.67	0.63
Enhancement	$DS5_{enh}$	0.30	0.29	0.18	0.65	0.67
	$DS6_{enh}$	0.60	0.57	0.36	0.75	0.75
Average		0.57	0.58	0.39	0.64	0.61

Table 5.4: Kendall_l correlations obtained for the six test sets using recognition statistics from the mixed-trained system. The 5 ASR statistics are Word Accuracy (WordAcc), Percentage Correct (Corr), Deletion (Del), Substitution (Subst) and Insertion. Note the relatively good correlations obtained with Subst and Ins, though with the notable exception of $DS2_{add}$.

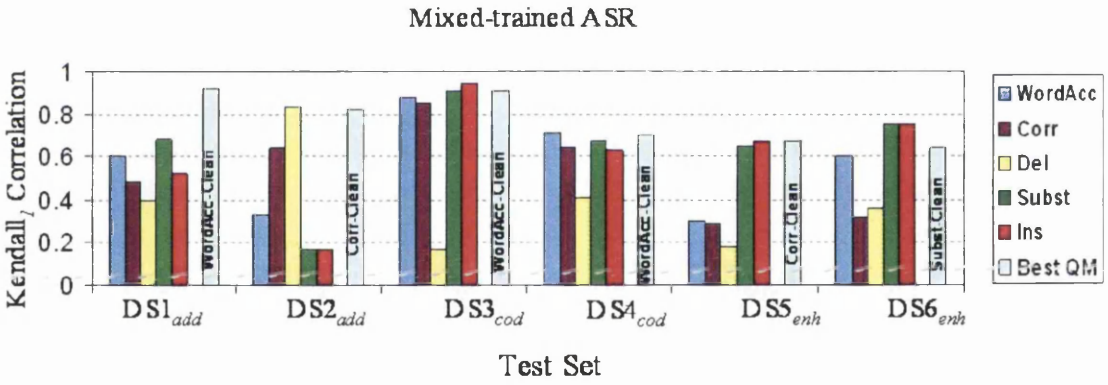


Figure 5.12: Bar plot showing Kendall_l correlation obtained with the 5 recognition statistics from a mixed-trained ASR system. The 6th bar (light blue bar) of every set of bars for each test set shows best correlation obtained with statistics from the clean-trained ASR system in Section 5.3.

test sets could be that with clean training, the substitution and insertion errors are contributed not only by distortions caused by the enhancement algorithms but also by the background noise. This introduces an offset making the measurement less reflective of the real impairment factor of interest in that particular test category, namely distortion caused by enhancement processes. With multi-condition training, the recogniser is already 'familiar' with the background noise and hence is more sensitive to degradations caused by the enhancement processes.

(4) On the other hand, Deletion correlates poorly most probably because very rarely do deletion errors occur. Most degradations are in the form of clipping and distorted pronounciations due to excessive attenuation, leading more to occurance of substitution and insertion rather than deletion errors. The poor performances of Word Accuracy and Percentage Correct could be due to unreliability of Deletion as an intelligibility indicator.

(5) As shown in Figure 5.12 the correlation for $DS1_{add}$ by WordAcc at 0.60 is much lower compared to that obtained with WordAcc from the clean-trained system at 0.92. Such poor correlation is obtained despite the fact that the Aurora2 multi-condition training set considers some of the same degradations considered in test set $DS1_{add}$ [2], which means that some of the degradation conditions are seen during training.

Regarding observation (5), the poor correlation of WordAcc for test set $DS1_{add}$ is perhaps expected due to two reasons:

(i) certain noises are seen during training while others are not (babble, car, exhibition and subway are seen; airport, restaurant, street and train are unseen) hence introducing bias in the system whereby signals degraded by seen noises could more readily be recognised;

(ii) the amount of 'tolerance' given to each noise type is not according to humans' perception of intelligibility, hence signals degraded by noises that are given more tolerance could more easily be recognized.

Factor (ii) is caused by training the recogniser with different types of degraded signals at identical SNR range. Such an arrangement causes the ASR to assume equal 'tolerance' for all babble, car, exhibition and subway noise, which proves not to reflect human opinion, since humans could tolerate more of certain noises, such as subway noise as opposed to other more adverse ones, such as babble. This is illustrated in Figure 5.13 which shows human scores for $DS1_{add}$, for example, humans could tolerate around -8dB of subway noise but only -5dB of babble noise for the signal to be 60% intelligible. The term 'tolerance' here refers to the amount of degradation a signal can take prior to arriving at a given intelligility level.

For illustration, Figure 5.14 shows that if the ASR is trained on babble noise and subway noise degraded signals both at -5dB, the ASR system is equipped with knowledge of babble noise degraded signals at human intelligibility of as low as 60%, but such training only equips it to recognise subway noise degraded signals at human intelligibility of about 80%, as indicated by the arrows. In other words the recogniser has been trained to be more robust towards babble noise degraded signals. This perhaps explains why the profile ranking of these two conditions are almost reversed when the ASR system employed switches from clean training to mixed training. This is shown in Figure 5.15 with profiles of word accuracy for babble, restaurant and subway noise degraded signals. The bottom three profiles are produced by the clean-trained ASR system while the top three by the mixed-trained system. Notice that the babble profile is very much below the subway profile when the ASR is trained on clean . However, when the ASR is trained on signals ofi multi-condition (mixed training), the babble profile is

[2] the multi-condition training set consists of 25% of clean utterances, and 75% of signals degraded by 4 noise types namely car, babble, exhibition and subway. The same 4 noise types are considered in test set $DS1_{add}$, on top of which are airport, restaurant, street and train noise

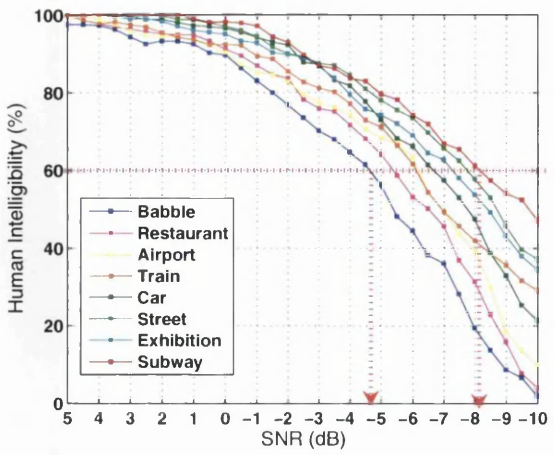


Figure 5.13: An illustration that humans have different level of tolerance towards different noises. For e.g., at 60% intelligibility, humans can tolerate only about -5dB of babble noise but up to -8dB of subway noise.

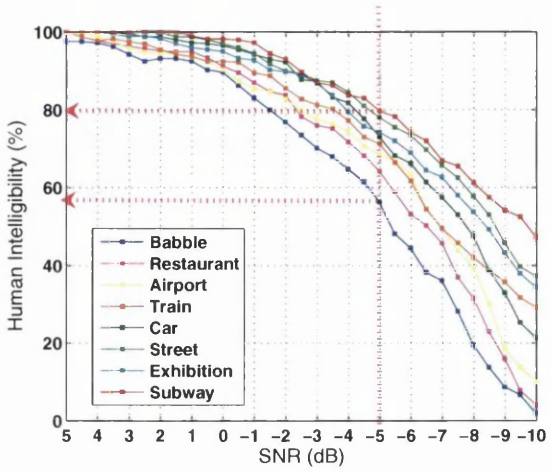


Figure 5.14: An illustration of bias resulted from mixed training the ASR with different noises at identical SNR. For e.g., if an ASR is trained on babble and subway signals both at -5dB, the ASR is trained to recognise babble signals with intelligibility as low as 60% but is only trained to recognise subway signals of $> 80\%$ intelligibility. More tolerance has been given to babble noise.

only slightly below the subway profile and eventually is above the subway profile, below -3dB. Notice also that even though restaurant noise is unseen during training, comparable recognition is obtained when tested against the mixed-trained ASR due perhaps to its similarity to babble noise. The absence of such bias in clean training conditions perhaps is the key factor leading to higher correlation of WordAcc from the clean-trained system. Obviously multi-condition or mixed training could be configured to give better correlations, however to do so without introducing bias is deemed to be difficult. Section 5.5 reports further work on mixed training.

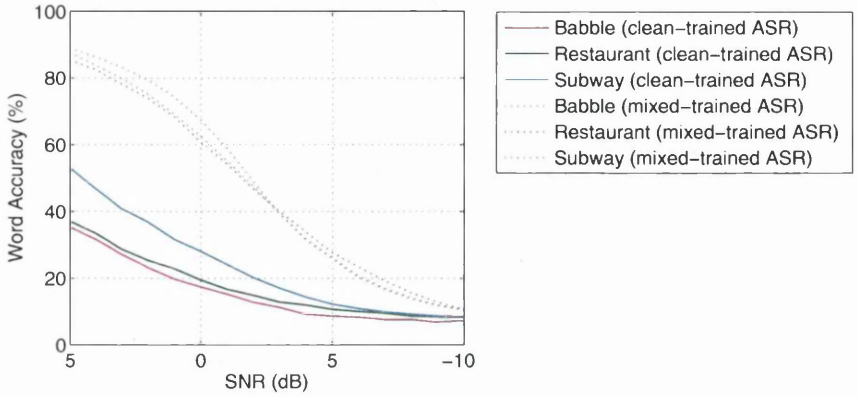


Figure 5.15: An illustration of the difficulty in improving correlation through mixed training. Solid profiles (given by clean-trained ASR) are ranked correctly while dotted profiles (given by mixed-trained ASR) are ranked incorrectly as the mixed-trained recogniser is more robust towards babble noise-degraded signals.

5.5 Preliminary Experiments of ASR with Input of Human Intelligibility

The performance of any data-driven system including ASR depends on the information given during training. The performance of an ASR system in distinguishing intelligibility levels is likely to be influenced by its knowledge regarding intelligibility, that is, some form of intelligibility labels rather than just word labels during training.

The previous section shows that indicators of mixed-trained ASR gives generally poor correlations especially for the environmental noise test sets ($DS1_{add1}$ and $DS2_{add}$) despite the fact that the training set consists of signals degraded by the same degradations. As has been discussed in Section 5.4.3.2, poor correlations are postulated to be due to an inappropriate level of tolerance given for each noise type during training. To avoid such bias the proportion of tolerance given to each degradation that is involved in the training set should perhaps reflect human opinions of intelligibility. One way to achieve this could be by fixing a desired intelligibility threshold and then identify the corresponding SNRs of each degradation crossing that threshold. This is illustrated in Figure 5.16 using human results for $DS1_{add}$. A threshold is fixed at 70% intelligibility (horizontal red dotted line) and the corresponding SNRs for each degradations in that category are identified, as shown by the labels **a** to **g**. In this particular example shown in Figure 5.16, the training set would consist of babble signals at **a**dB, restaurant signals at **b**dB and so on. Obviously, this information only becomes available with human listening tests, which to an extend negates the motivation for an objective measure. However, it is hoped that with enough reference points, intelligibility of signals degraded by degradations unseen during training can be interpolated.

This section investigates the possibility of improving intelligibility correlation of ASR by imparting human knowledge of intelligibility during the training of the recogniser. Three preliminary experiments

are performed on selected test sets: the first and second experiments are performed on $DS1_{add}$; and the third on $DS6_{enh}$. Experiments conducted here are mainly for proof of concepts and may not be statistically significant.

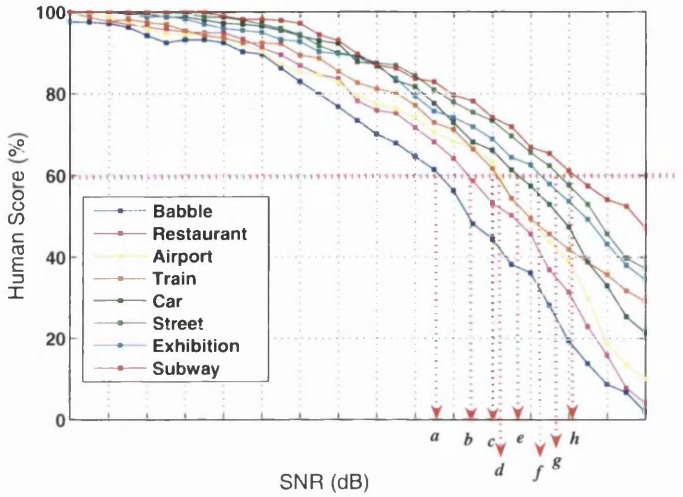


Figure 5.16: An example of how balanced levels of data based on human perception of intelligibility can be given to each degradation type. The training set should consist of each degradation at SNRs corresponding to the same level of human scores as shown by the horizontal dotted line.

5.5.1 Experiment Procedures

5.6.1.1 Experiment I

The first experiment, Experiment I use a training set that has the similar structure to that of the Aurora2 multi-train set. The Aurora2 multi-train set consists of signals degraded by babble, car, exhibition and subway noise at SNRs clean, 20, 15, 10, and 5dB. The new training set use the same degradations except that, rather than fixed identical SNRs, the training signals are degraded at SNRs corresponding to 4 fixed intelligibility thresholds, namely 62.5%, 75%, 87.5% and 100% (corresponding to 2.5, 3, 3.5 and 4 digits respectively). The training set consists of 16 sets (4 noise types $\times$ 4 SNRs) of 566 utterances.

5.6.1.2 Experiment II

The second experiment, Experiment II trains the recogniser in the same way as in ExperimentI except that instead of using just 4 noise types, all 8 types of noises under $DS1_{add}$ are involved in the training. Similarly the SNRs of each type of degraded training data correspond to the 4 fixed intelligibility

thresholds of 62.5%, 75%, 87.5% and 100%. This approach is no doubt impractical and infeasisble since the human scores for all degradations under test are needed beforehand. However, this rather extreme experiment serves as a comparison to result of ExperimentI. Better correlations are anticipated.

5.6.1.3 Experiment III

The third experiment, Experiment III attempts to improve correlation for $DS6_{enh}$ where degradations considered are of the same characteristics namely distortions introduced by the NLSS algorithm. The new training set is created by processing the Aurora2 multi-raining set with 2 different configurations of the NLSS. The configurations are noise over estimate of 0.1 and 0.001 respectively both with the noise floor set to 3. The configurations are chosen such that one configuration causes more adverse degradation than the other (noise-over-estimation of 0.001 generally causes worse degradation as there is more attenuation). Signals degraded by the first setting cover lower range of SNRs (clean to -5dB) so that more tolerance is given; and signals degraded by the second setting cover higher SNR range (clean to 5dB).

5.5.2 Observations and Discussion

Figure 5.17 shows bar plot of average word accuracy obtained for 8 $DS1_{add}$ degradations using recogniser of Experiment I. The bar plot is presented in such a way that human perceived intelligibility increases from left to right. Green bars represent noise types that are seen during training while for the blue bars the noise is unseen. First of all, notice that the 4 noise types that are involved in the training, namely babble, car, exhibition and subway are ranked correctly, as shown by the green bars with increasing heights from left to right. However, this knowledge does not seem to equip the recogniser with the ability to rank signals degraded by unseen noise types. This is shown by the blue bars which do not show increasing monotonous trend from left to right. Besides, babble noise degraded signals should be the least intelligible but in this case is higher than all the blue bars. This preliminary proof-of-concept experiment suggests that it might prove difficult to train an ASR system for it to be able to rank intelligibility. Training an ASR system to recognise word is perhaps very different to training it to evaluate intelligibility.

Figure 5.18 shows bar plot of average word accuracy obtained for 8 $DS1_{add}$ conditions using recogniser of ExperimentII. ExperimentII involves all 8 $DS1_{add}$ noises in its training at tolerances reflective of human intelligibility. In other words, intelligibility ranking of the test signals have already been made known to the recogniser during training. Since Experiment I gives perfect ranking for the conditions that the recogniser has seen during training, it is thought that this Experiment II's recogniser could also give good ranking for all 8 conditions under $DS1_{add}$. As shown in the Figure 5.18 the trend of increment from left to right is now more obvious compared to that of Figure 5.17. In fact, the

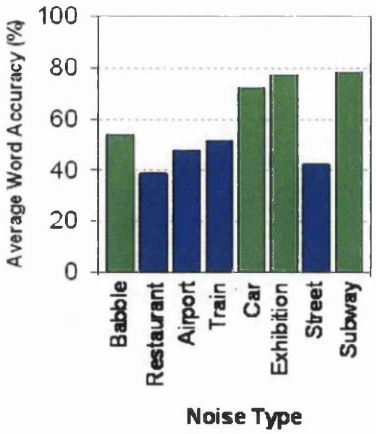


Figure 5.17: Bar plot shows average word accuracy obtained for 8 $DS1_{add}$ using ASR system of Experiment I where the training set consists of signals degraded by babble, car, exhibition and subway noise (green bars) at SNRs corresponding to human-defined intelligibility. The x-axis components are arranged such that human-defined intelligibility increase from left to right. Notice that green bars (seen degradation) increases from left to right, however blue bars (unseen degradation) and all bars as a whole do not show such trend.

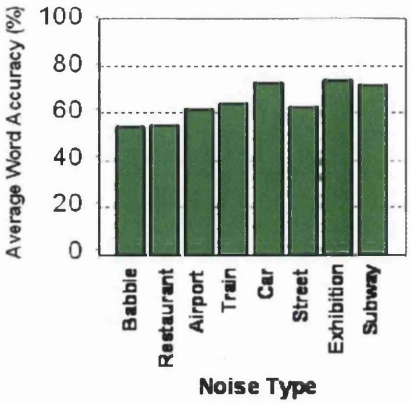


Figure 5.18: The same as in Figure 5.17 except that ASR system is trained on signals degraded by all degradation conditions under $DS1_{add}$ at SNRs corresponding to human-defined intelligibility. The trend of increment from left to right is now more noticeable.

trend is almost monotonic. Possible reason that perfectly monotonous ranking is not obtained could be that among these 8 $DS1_{add}$ noise types there are some belonging to the same nature/characteristic, hence unintentional, extra tolerance could have been set for group of degradations of that nature. For instance, if degradation A and B are very similar, though each trains the recogniser at SNR reflective of human intelligibility respectively, grouping together could unintentionally equip the recogniser more in recognising signals degraded by degradation of this characteristic. By comparing Figure 5.17 and 5.18, this preliminary proof-of-concept experiment shows that it is possible to tune an ASR system for the task of intelligibility ranking, though it also suggests that it could be difficult to predict the interaction between the training data of different conditions and hence the behaviour of the recogniser.

Experiment III attempts to equip the ASR system to rank intelligibility involving only degradation of one specific characteristic, namely those introduced by the NLSS process. Figure 5.19 and 5.20 show bar plot of average word accuracy obtained for $DS6_{enh}$ conditions using the clean-trained ASR system from Section 5.4 and the newly trained recogniser from Experiment III respectively. Similarly the bar plot is presented in such a way that human perceived intelligibility increases from left to right. As shown this trend of left-to-right increment does not exist in Figure 5.19 where the clean-trained ASR is used, but can be seen in Figure 5.20 where the newly trained recogniser is used. The Kendall_1 correlation is improved from 0.53 as presented in Table 5.4 to the current 0.65. This suggests while the scores from clean-trained ASR might be poor intelligibility estimators of signals degraded by enhancement

processes, through targetted training the recogniser learnt the particular patterns associated with good and poor intelligibility, hence leading to better correlation.

This preliminary experiment also suggests that it is possible to configure an ASR system for intelligibility testing in specific application dealing with specific types of degradation. Such a measure is useful for testing of new systems under development where optimal parameters need to be found empirically. One good example is the image-based enhancer discussed in Section 5.3.4.

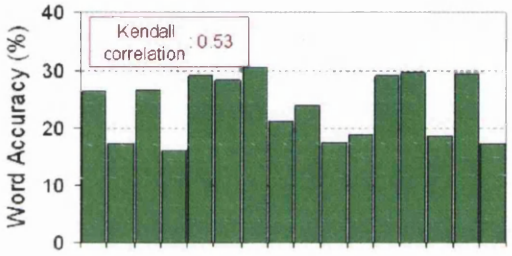


Figure 5.19: Bar plot of average word accuracy obtained for the 16 $DS6_{enh}$ degradations using clean-trained ASR. Data is obtained from Section 5.3.2.

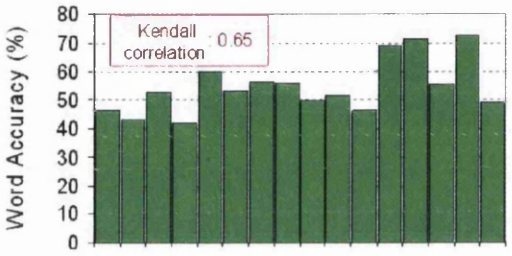


Figure 5.20: The same as in Figure 5.19 except that recogniser is trained on recogniser of Experiment III. A trend of increment from left to right is now noticeable.

5.6 Concluding Remarks

This chapter investigates the potential of ASR systems in intelligibility assessment. Several configurations of ASR are considered including clean and mixed training, and application of the state-of-the-art missing data techniques. The experiments carried out here aim to highlight some fundamental strengths and limitations of ASR for the task of intelligibility assessment. The main points covered in this chapter can be summarised as follows:

- Five recognition statistics namely Word Accuracy, Percentage Correct, Deletion, Substitution and Insertion are assessed for their potential as indicators of intelligibility.
- Different statistics/indicators can be used in different context for different applications depending on the nature of degradations involved. This contrasts the quality measures considered in Chapter 4 where good and bad performances of the measures are not easily comprehensible in such straightforward and logical manner.
- Table 5.5 lists best-performing ASR statistic for each test set, namely DS1 to DS6. Possible reasons for each correlation performance are discussed earlier in this chapter. Main observations are stated here in brief:

	ASR-based Indicator	Best Quality Measure
$DS1_{add}$	0.92 (WordAcc_clean)	0.72 (WSS)
$DS2_{add}$	0.82 (Corr_clean)	0.86 (PESQ)
$DS3_{cod}$	0.91 (WordAcc_clean)	0.93 (WSS)
$DS4_{cod}$	0.71 (WordAcc_mixed)	0.67 (WSS)
$DS5_{enh}$	0.67 (Ins_mixed)	0.75 (WSS)
$DS6_{enh}$	0.75 (Subst_mixed, Ins_mixed)	0.68 (WSS)

Table 5.5: Kendall correlations for the 6 test sets obtained with ASR statistics in comparison to the best quality measures for each test set. ASR statistics are in ther form of *statistic_training*, for e.g., Ins_mixed refers to Insertion from the mixed-trained system.

(i) Noises that are more stationary (e.g., $DS2_{add}$) can be more effectively assessed by Deletion rather than Substitution or Insertion since such noises, when at a high level, rarely incur substitution and insertion errors but simply mask the signal beyond recognition.

(ii) Noises that are more speech-like (e.g., $DS1_{add}$) are more effectively assessed by Substitution and Insertion rather than Deletion since such noises are often mistaken as speech.

(iii) Degradation caused by the coding processes can be more effectively assessed by Substitution rather than Insertion or Deletion since such spectral distortion often affects pronunciation of words leading to substitution errors.

(iv) Almost all ASR statistics correlate poorly with the enhancement test sets since such processes artificially improve ASR scores without improving intelligibility.

(v) Excessive attenuation by the enhancement processes could cause changes in pronunciation of words leading to substitution error and introduce clicking sound leading to insertion error when mistaken as speech. Whilst statistics from the clean-trained system correlate poorly in this context, Substitution and Insertion from the mixed-trained system are potentially more useful presumably because mixed training causes the system to be more sensitive to such degradation rather than distracted by the underlying background noise which in this case, is not the main factor degrading the intelligibility.

(vi) In overall Word Accuracy from the clean-trained ASR (WordAcc_clean) is shown to correlate best with average correlation at 0.72 and 0.81 if the enhancement test sets namely $DS5_{enh}$ and $DS6_{enh}$ are excluded.

- Improving recognition does not necessarily lead to better intelligibility correlation.
- Being data-driven, there is flexibility of training the system for specific applications. In particular, input of human knowledge of intelligibility trains the ASR towards different level of tolerance for different degradations, forming yardsticks of intelligibility levels in the recogniser.

Input of human-derived intelligibility could potentially tune an ASR system into an intelligibility measure as suggested by the preliminary findings reported in Section 5.5. The difficulty is how to

impart this knowledge into the recogniser such that it is trained not just for recognising word, but also for distinguishing level of speech intelligibility. Another difficulty relates to the fundamental difference between ASR and human speech recognition. Not only that machine performance is very much below that of humans [17,18,29,30], but human's superiority in speech recognition (and intelligibility assessment) is still largely unknown [17] implying that the processing in ASR does not necessarily reflect humans'. Therefore, while human word recognition is intelligibility, ASR word recognition is not (at least not yet).

Nonetheless, with the vast amount of research effort invested into bridging ASR and human speech recognition, future development might see ASRs approaching humans both in terms of performance (recognition rate) as well as behaviour/processing (intelligibility correlation). The missing data technique is one good example of the significant advancement made in this area. As shown in Section 5.4.3 word accuracy has doubled with the use of the missing data fuzzy mask (compared to traditional recognition). Though the intelligibility correlation obtained with this particular technique is generally poor, as ASR technology advances and as ASR processing better mimics humans' auditory, this correlation might improve. Though still a vision, such possibility is acknowledged.

Part II

Direct, Data-driven, Differential Intelligibility Classification (D^4IC)

Part II is titled "Direct, Data-driven, Differential Intelligibility Classifier (D^4IC)". The terminologies are used throughout the remaining chapters and are defined as follows:

- Direct: output score of the classifier is a direct indication of intelligibility levels (relative) . This contrasts approach in Part I where intelligibility is estimated indirectly through quality or ASR word recognition.
- Data-driven: the classifier is trained on data sets labelled with known intelligibility information. This contrasts the quality measures in Part I which are rule-driven but is equal in concept to ASR training.
- Differential: output score of the classifier indicates relative intelligibility between a pair of signals (i.e. which is the more intelligible?). This contrasts approach taken in Part I where intelligibility of each comparing signal is estimated explicitly and relative intelligibility deduced by comparing two separate, explicit estimations. The term 'differential' refers to the act of implicit rather than explicit signal differencing.

Part II consists of 4 chapters within which the 3 difficulties associated with objective intelligibility assessment as discussed in Chapter 2 are addressed. Briefly, the difficulties are:

(i) resultant of small dynamic score range due to the inevitability to operate under high degradation.

(ii) the presence of difficult or 'confusing' processes that artificially improve machine-based scores yet often without actually improving intelligibility.

(iii) the difficulty to establish substantial amount of reliable ground truth for development and evaluation of objective intelligibility measure.

The 4 chapters can be summarised as follows:

- Chapter 6 introduces the idea of direct differential intelligibility classification where the classifier would be trained on pairs of signals with known differential intelligibility and its output score is a direct function of differential intelligibility. A major contribution here is a practical strategy to generate large amount of ground truth needed to train such a D^4IC. This addresses difficulty item (iii).
- Chapter 7 presents the experimental framework with the goal to assess how good the D^4IC might be. Benchmark experiments are reported.
- Chapter 8 assesses potential features for the classifier where features are scores coming from quality measures and ASRs. In particular scores from various stages of the ASR word recognition process are used as features for the D^4IC.
- Chapter 9 proposes a novel feature for the D^4IC based on the concept of anchor models. Here the signal intelligibility is effectively characterised by the similarity between the degradation it underwent and a cohort of chosen degradations. This feature set addresses difficulty item (i) and (ii).
- Chapter 10 evaluates the D^4IC using the human-evaluated test sets used in Part I, namely DS1 to DS6.

Chapter 6

Introduction to D^{4}IC

In Chapter 4 intelligibility of an input signal is estimated through a function of quality and in Chapter 5 a function of ASR word-recognition. Both approaches can be seen as indirect, though the latter appears to be closer related. This chapter introduces the idea of a data-driven classifier the output score of which is directly related to intelligibility, and, being data-driven, is trained on data sets with known intelligibility information. Thus, the potential accuracy over the previous measures comes from the fact that information about intelligibility is instilled into the classification system.

6.1 Direct Intelligibility Classification

Two pre-requisites to any data-driven classification system are firstly a representative set of data labelled by its class and secondly a classifier unit as illustrated in Figure 6.1. To build an intelligibility classifier, the training set would be labelled by pre-determined intelligibility levels hence as the input is switched to train mode, these intelligibility levels would be 'learnt' by the classifier. Then, during test mode a signal of unknown intelligibility level would be classified according to the information uncalculated into the classifier during training. Such arrangements depend on factors such as features, normalisation, training data with the appropriate ground truth (class labels) and the classifier structure itself. Neural network, markov models, support vector machine (SVM) could all form such classifier. Meanwhile, features could be scores derived from measures considered in Part I as shown in Figure 6.2 where M is a measure such as PESQ, and s_1 is the output score from the measure which was taken directly as intelligibility indications of signal S_A in Part I of the thesis; now s_1 could serve as feature for the classifier. The output score, s_2 is a function of direct intelligibility, i.e. $s_2 = \text{fn}(\textit{Intelligibility})$. Intelligibility assessment is now via a statistical classification and this can be seen as an extension to word modelling in ASR.

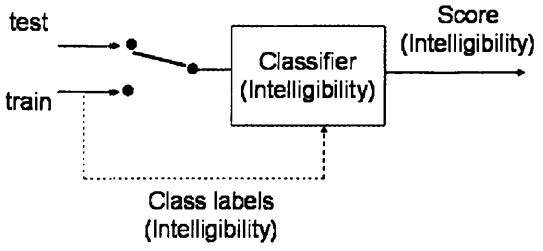


Figure 6.1: An illustration of direct intelligibility classification where the classifier is trained on signals with known intelligibility levels during train mode; during test mode the test set is classified accordingly.

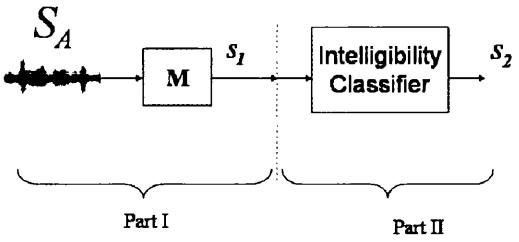


Figure 6.2: In Part I where M is a measure such as PESQ, s_1 is taken as intelligibility indication of signal S_A; in Part II this score could be used as feature for the intelligibility classifier.

6.1.1 Problem with Training

A major obstacle to realizing the arrangement shown in Figure 6.1 is sufficient amount of signals with known intelligibility levels to train the classifier. Interestingly and unfortunately, while there are extensive speech databases for data-driven tasks such as ASR (the Aurora2 database [26] used in Part I) and speaker recognition (National Institute of Standards and Technology (NIST) Speaker Recognition Evaluation (SRE) Series [109]), there are no intelligibility oriented databases with a variety of degradations and corresponding labelled intelligibility labels. The only exception is perhaps the FW03 database from NTT-AT (Nippon Telegraph and Telephone Advanced Technology Corporation) [110], however, the database is collected using Japanese word lists and is aimed for research into hearing aids.

Apart from the two 'free' sources of labelled data at (i) clean speech signals (i.e., 100% intelligibility); and (ii) very heavily corrupted signals (i.e., 0% intelligibility), labels of intelligibility are only obtainable through human listening tests. Establishing such ground truth for the classifier using humans would be extremely difficult if not infeasible due to the large amount of tests needed for signals across intelligibility range and across the essentially enormous, theoretically infinite variety of degradations (and their possible combinations and permutations). Due to the laboriousness and costs of listening tests the classifier could only be trained on limited amount of ground truth where such human efforts can be afforded, covering most probably only a specific intelligibility range and numbered degradations types. Therefore as the operational range of the classifier widens both in terms of intelligibility levels and degradation types, the classifier's performance is compromised as it becomes more challenged in its ability to extrapolate and to generalise. This is illustrated in Figure 6.3 where the vertical arrow refers to the requirement to operate at wider intelligibility range; profiles of dotted lines refer to degradations unseen during traning and the horizontal arrow refers to the requirement to generalise towards unseen degradations.

Critical to performance of any data-driven system is the quantity and quality of its training data. As an indication, the clean training set of Aurora2 [26] consists of 8440 digit strings spoken by over a 100 different male and female speakers; for every digit there are in excess of 2000 occurrences to ensure sufficient variations so that robust speaker-independent word recognition can be achieved. A major contribution of this research is proposing a strategy for acquiring potentially infinite amount of training data based on what we refer to as the Intelligibility Enhancement (IE) Hypothesis.

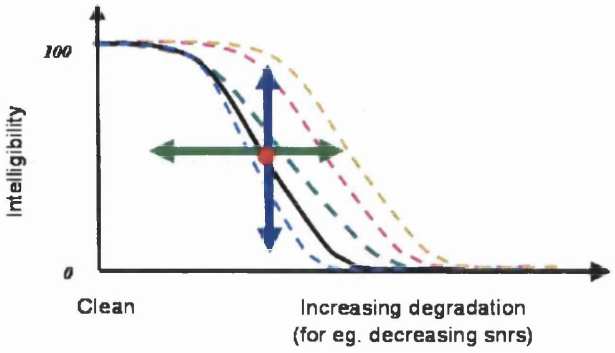


Figure 6.3: An illustration of the challenges that the classifier could anticipate as its operational range widens both in terms of intelligibility ranges and the variety of degradations encountered. Horizontal axis refers to the need to generalise towards unseen degradations (dotted profiles); and vertical axis the need to extrapolate to other intelligibility levels or regions.

6.2 Intelligibility Enhancement (IE) Hypothesis

"Speech processing rarely improves speech intelligibility and most processings either degrade or at most leave the intelligibility unchanged. In other words, under many practical circumstances where a speech signal is degraded to below 100%, it is impossible to enhance its signal to improve its intelligibility." This we refer here to as the Intelligibility Enhancement (IE) Hypothesis. A hypothetical illustration is shown in Figure 6.4 where intelligibility is plotted as a function of increasing undergoing processes. At zero process the signal S is clean hence the corresponding intelligibility is naturally 100%; S_1 is a processed version of S which is equally or less intelligible; S_1 is further processed resulting in S_{12} where the subscript '$_{12}$' indicates that S_{12} has 'accumulated' the degradation effects of both the 1^{st} and 2^{nd} process. This goes on progressively and after enough processes the signal would eventually be rendered unintelligible. The grey arrow highlights the monotonic profile expected as signals goes through more and more processes. Another illustration is given in Figure 6.5 where the solid profile is s-curve of an original set of signals, for example, signals degraded by airport noise at a range of SNRs. If these signals are further processed, the processed versions (dotted profiles in figure) would correspond to lower intelligibility hence the resultant s-curves could only go below the original s-curve.

This hypothesis is obviously true in simple cases where a speech signal is corrupted by increasing levels of additive noise. Presumably it can also be extended to more general form of processes such as speech coding since it is only logical that the act of compression and de-compression lead to reduction in signal intelligibility, or at best, leave the intelligibility unchanged. One 'grey' area is process that aims to enhance signals such as speech enhancement or de-noising algorithms. However, the hypothesis might be true even in this 'grey' area with the observation that it is exceedingly difficult to improve the intelligibility of a degraded signal. Some supporting evidences are given in the next section.

6.2.1 Supporting Evidences

Some supporting evidences for this hypothesis are discussed in Section 2.2. The main and some additional points are stated here.

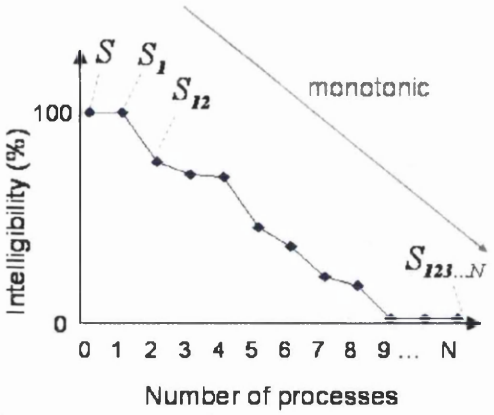


Figure 6.4: An illustration of the IE Hypothesis where signal intelligibility is expected to fall or remain constant as the number of processes it underwent increases. S is clean signal; S_1 is a processed version of S; S_{12} is a processed version of S_1; eventually $S_{12,,,N}$ is unintelligible when N is big enough. A monotonic profile is expected as indicated by arrow.

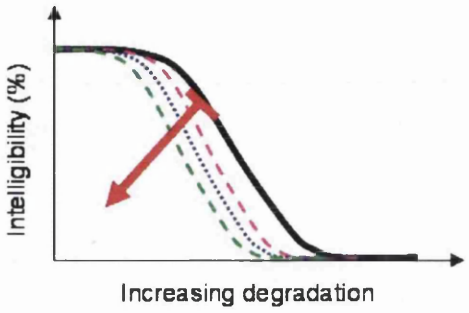


Figure 6.5: An illustration of the IE Hypothesis. Solid profile: intelligibility of an original set of signals; dotted profiles: the original set further processed in different ways. The hypothesis states that the dotted profiles could only move in negative direction as indicated by the arrow.

- In 1986 Lim published [36] a review on a large range of speech enhancement techniques that account specifically for additive noise, and reverberation respectively. It is observed that none of the approaches improves speech intelligibility, and some in fact decrease intelligibility.

- In 1987 [111], a study was conducted to assess noise suppression algorithms in the context of intelligibility in two application areas, namely two-way communications and transcriptions of recorded material. DRT indicated that none improves intelligibility.

- In 1991, S. F. Boll [56] commented "Why has no one found a way to eliminate noise from speech in order to improve intelligibility?". Boll is a pioneer in the field of speech enhancement. It is ironic for such statement to be made after a decade of active research in enhancing speech.

- Very recently in 2007, Hu et al [3] reported a comprehensive investigation on eight speech enhancement methods encompassing four classes of algorithms: spectral subtractive, subspace, statistical model based and Wiener-type algorithms in the context of their influence on speech intelligibility. Degradations considered are babble, car, street and train noise condition at 0dB and 5dB. In the opening statements of [3], it is mentioned that while the primary goal of speech enhancement is to improve quality, the secondary goal is to (at least) preserve intelligibility. It seems that intelligibility improvement is not even expected. Two forms of listening tests are conducted, namely consonants and sentence recognition. Of the 64 cases of noise conditions and algorithms considered, sentence test reported only one improvement case (car noise 5dB) while consonant test reports none. Among all algorithms considered, the Wiener-as algorithm

performed the best most probably because it introduces the least amount of attenuation. At the other extreme, the perceptual Karhunen-Loeve transform (pKLT) approach imposes significant noise reduction but impairs speech intelligibility. These two extreme cases highlight the difficulty of intelligibility improvement.

Though the hypothesis is bold and it is obligatory to consider the implications caused if it proves to be false, however, it is believed to be well-established considering that no speech intelligibility enhancement system has been standardised or widely published. This hypothesis would be the foundation for which the data generation strategy described in Chapter 7 is based on.

6.2.2 Information of Differential Intelligibility

Based on the hypothesis, it is possible to generate two signals where the intelligibility of one is known to be greater than or equal to the intelligibility of the other, even though the precise difference itself is unknown. This is done by putting one of the signal through one or more extra processes. For instance, while exact intelligibility of signals are unknown in Figure 6.4, it can be certain that the intelligibility of $S_{1,,n}$ is more than or equal to the intelligibility of $S_{1,,,n+m}$ where n and m are finite numbers.

Given that it is rather 'effortless' to go on generating large numbers, potentially infinite pairs of signals with known differential intelligibility, this leads to the idea of modelling differential intelligibility where models are trained on such signals pairs. The ground truth obtained in this way is considered reliable as long as we have confidence in the IE Hypothesis. This effectively addresses the difficulty discussed in Section 2.3, namely the need for and the scarcity of reliable ground truth for the development and evaluation of objective intelligibility measures. Of course there might well be some exceptions to the hypothesis such as when the noise is of a predictable nature (for e.g.,, train with constant speed over rail track giving constant frequency pulses). However it is possible to avoid such violation by constraining the choice of processes that the signals go through. In this research we consider such process as including environmental noises, coding algorithms, common speech processing techniques such as re-sampling and filtering, and even (majority if not all) enhancement algorithms as suggested by the literature listed in previous section. The process to generate ground truth signals pairs is elaborated in Chapter 7.

6.3 Data-driven Direct Differential Intelligibility Classification (D^{4}IC)

Since the ground truth provided is in terms of differential intelligibility between signals in pairs, the classifier is essentially a differential classifier, hence the name D^{4}IC which the 'D^4' stands for data-driven, direct and differential. Of course this also means that absolute intelligibility of each input signal is lost. In previous chapters the relative intelligibility between a pair of signals is deduced by comparing two separate estimations as shown in Figure 6.6(a) where M is a measure such as PESQ, scores s_A and s_B are intelligibility estimations given to signals S_A and S_B respectively; the scores are differenced to determine which of the signals is more intelligible. Here rather than estimating intelligibility of each signal explicitly, the D^{4}IC outputs a score that directly indicates their relative intelligibility. This can be thought of as bringing the differencing from score level earlier to model level in D^{4}IC. Since the primary objective of the thesis is comparative intelligibility, and bringing the differencing to an earlier stage in the process is thought to be more targeted towards achieving the objective.

The differencing process can also be brought even earlier to feature level as illustrated in Figure 6.6(b) where f_A and f_B are feature vectors representing signal S_A and S_B respectively; and f_B is subtracted from f_A resulting in f_{AB} of the same vector size which represents the differences between signal S_A and S_B. Note that the f_{AB} is the same s_{AB} in Figure 6.6(a) which is used directly for decision making in Part I. Here f_{AB} is used as feature for the classifier and the classifier's output directly indicate relative intelligibility between S_A and S_B. In both cases s_{AB} are thresholded to give the decision YES or NO corresponding to the question 'Is S_A more intelligible than S_B?'. Clearly this feature level differencing are not always applicable to other classification tasks as the comparing signals may not be directly comparable, for instance, the signals could have different length, speaker and content. In the experimental setup used in this research, S_A and S_B comes from the same source, only processed / degraded in different ways. This essentially allows feature differencing to be performed at various levels from frame to the whole utterance. This early differencing process causes the classifier to focus on signal differences and in turn, intelligibility differences which hopefully would assist in classification accuracy.

Worth noting is that here the D^{4}IC has the distinct benefit of balanced data sets for training the 2-class problem (i.e., YES and NO class). This is possible since it is simple to generate equal amounts of data for the 2 classes, and the data sets can be equivalent in all aspects, one complementary of the other. This means that the scores are normalised and decision thresholds are not needed hence a prior threshold determination is not required.

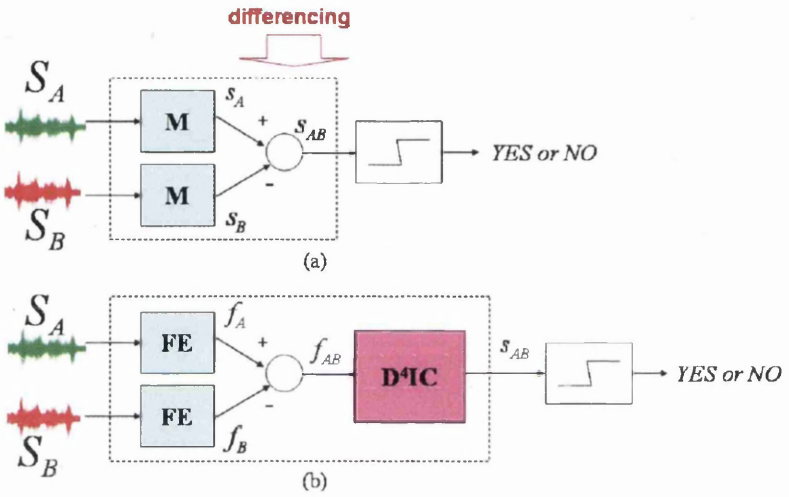


Figure 6.6: Figure (a) shows Part I approach where intelligibility of each comparing signal is estimated explicitly, the difference of two resultant scores is used for decision making. Figure (b) shows D^{4}IC structure where FE is the feature extraction unit where M could be a measure considered in Part I such as PESQ, the same s_A, s_B and s_{AB} in (a) are f_A, f_B and f_{AB} in (b). f_{AB} serves as feature into the D^{4}IC. In both approach the s_{AB} is thresholded to give the decision of YES or NO, corresponding to the question 'Is S_A more intelligible than S_B?'

6.4 Epilogue

The main points in this chapter are summarised as follows:

- Rather than estimating intelligibility indirectly through quality or ASR word recognition, we propose a statistical model that is trained on signals with intelligibility information and outputs scores that relate directly to intelligibility. The direct intelligibility classifier could take in as features the scores from measures considered in Part I.
- We overcome the main constraint in statistical modelling of intelligibility, i.e., the lack of training data, which, in this case, are speech signals labelled with intelligibility information.
- IE Hypothesis enables the generation of large amount of data with information of relative intelligibility, i.e., which is the more intelligible among a pair of differently processed signals? oa.
- A data-driven direct differential intelligibility classifier (D^{4}IC) is proposed where the output score relates to differential intelligibility between a pair of signal. Explicit estimation for each input signal is not needed. This is thought to be bringing the differencing (the act of comparison) earlier from score to model level, which could be seen as more targeted towards achieving the objective.
- Feature level differencing is proposed where features into the D^{4}IC are differences between features of the two comparing signals. This makes the classifier focuses on signals difference and subsequently intelligibility difference.

Chapter 7

Experimental Framework and Benchmark Results

Chapter 6 proposes a data-driven, direct, differential intelligibility classifier, referred in short to as the D^{4}IC. The classifier is trained on intelligibility difference and the derived score indicates intelligibility relationship between signals in pairs. The remaining chapters aim to assess the performance of the D^{4}IC which is likely to be influenced by factors including the (make-up/composition of) training data, normalisation strategies, features and the classifier itself. Here we choose to investigate feature sets, while keeping the other factors constant.

The goals of this chapter is to present the experimental framework and the D^{4}IC benchmark results. In the following chapters we consider different feature sets using otherwise similar D^{4}IC and data structures, with the aim to give an indication of how good the D^{4}IC might be. Lastly we compare the D^{4}IC performance with that of measures considered in Part I.

This chapter is structured as followed: as described in Chapter 6 the D^{4}IC is to operate on pairs of differently processed signals with assumed knowledge of the relative intelligibility (for e.g., $I_{S_A} \geq I_{S_B}$), hence this chapter begins by describing the creation of the data sets designed to assess the D^{4}IC; then followed by description of the feature structure and lastly some benchmark results are presented.

7.1 Database Design and Realization

The Intelligibility Enhancement (IE) Hypothesis provides flexibility and convenience to create potentially unlimited amount of ground truth, that is, pairs of degraded utterances with assumed the relative intelligibility of $I_{S_A} \geq I_{S_B}$. This section describes how the data sets for development and evaluation of the D^{4}IC are obtained.

7.1.1 Generate Data Sets by Applying Intelligibility Enhancement (IE) Hypothesis

This section describes the procedures involved to generate pairs of signals with known intelligibility relationship. The process can be summarised into four steps:

Step 1: Define Degradation Pool: This is a selection of degradation processes of interest. Ideally these are degradations that are likely to be encountered during normal use of D^{4}IC.

Step 2: Generate IE lines: A set of clean signals are progressively processed with degradations semi-randomly chosen from the pool (randomness explained in Appendix A.2). This results in a series of signals where signals at each stage of the processing are degraded to different levels. The series of signals is called an IE line and a set of such IE lines are generated. The input to the first process of an IE line are clean signals, subsequently output signals of one process become input signals of the next process. After each process the signals are assumed to be equally or less intelligible than before. The typical number of processes per line is 10 to 15, which is sufficient for the output signals of the final process to be rendered unintelligible. Figure 7.1 shows two examples of IE lines with 10 nodes per line where a node represents a process chosen from the degradation pool.

Step 3: Informal Listening Test: Due to the s-curve property of human intelligibility response where both ends of the curve are notionally of constant intelligibility namely 100% and 0%, signals from the first few nodes of a IE line are likely to be perfectly intelligible while signals from the last few nodes, totally unintelligible. Hence the useful region in practice (where intelligibility is an issue) is in between the node where the signals start to lose intelligibility and the nodes before signals have lost all intelligibility. An informal SRT (speech reception threshold) test is used for this purpose. For every IE line the SRT test identifies the key nodes where intelligibility is at the limit of 100% and just above 0%. In addition the key node for 50% is also estimated. In practice these 3 intelligibility thresholds (i.e., (i) 100%, (ii) 50% and (iii) 0%) are relatively easy to identify. According to the listening tests setup described in Chapter 3, these 3 thresholds would correspond respectively to (i) 4, (ii) 2 and (iii) 0 correctly recognised digits per utterance of 4 digits. It takes roughly 1 minute to perform SRT listening test for each IE line to find the 3 key nodes. More key nodes at more refined intelligibility intervals could be identified depending on the application condition of interest. Figure 7.2 shows the same IE lines in Figure 7.1 after being divided into 4 sections of different intelligibility range (shown by different shading) using the three key nodes mentioned. Similar to a s-curve, the IE line has 3 regions:

- wholly intelligible region (above key node (i) where intelligibility = 100%),
- dynamic intelligibility region (between keynode (i) and (iii) where 100% $\leq$ intelligibility $\leq$ 0%)
- unintelligible region (below key node (iii) where intelligibility = 0%).

Signals at unintelligible region are discarded.

Step 4: Identify Signals Pairs: The final step is to identify pairs of signals, S_A and S_B where S_A is more intelligible than S_B (the precise difference remains unknown). Pairings are performed using assumptions (i), (ii), (iii) stated below:

(i) within the same IE line, in the Dynamic Intelligibility Region (see Figure 7.2), the output signal of any node below the current node is less intelligible.

(ii) within and across any IE line, the output signals of any node in the Wholly Intelligible Region is more intelligible than the output signals of any node in the Dynamic Intelligibility Region.

(iii) Within or across any IE line, within the Dynamic Intelligibility Region, output signals of any node from the upper dynamic regions (100% < intelligibility < 50%) are assumed to be more intelligible than output signals of any node in the lower dynamic regions (50% ≤ intelligibility < 0%).

Examples of pairings are shown in Figure 7.2 where an arrow connecting two nodes implies that output signals of the first node are more intelligible than those of the second node. Examples of pairs identified using assumption (i), (ii) and (iii) mentioned above are shown respectively in the figure; let m be the number of IE lines needed to generate a data set. The size of m depends on the size of the pool (i.e., the number of different degradations), the bigger the pool the bigger m is in order to simulate enough occurrences of all degradations. Typically 20 IE lines would give >2000 pairs. N is the number of nodes in each line while n the number of remaining nodes after the SRT test, therefore $n \leq N$. Two example pairs from Figure 7.2 are:

- IE Line 1-node 1: Car more intelligible than
  IE line 2-node 5: Exhibition+G723+LD_CELP+Cityrain+Highway
- IE line 1- node 1: Car more intelligible than
  IE line 1-node 3: Car+Gaussian+GSM

7.1.2 Database Design

The speech signals into every IE line is the set of 566 clean 4-digits utterances which are subsets of the standard Aurora2 database. They are also the same clean reference signals used in Part I (see Section 3.2.1 for a description of the signals). As mentioned previously the benefit of using digits includes the use of the Aurora2 database with its large range of speakers and versions. Another important advantage is being able to use the Aurora2 framework which is optimised for ASR in noisy environments, the scores from which are used as features for the D^4IC.

The degradation conditions are chosen to reflect work done in previous chapters, going across the intelligibility range of 0% to 100% and a broad range of degradation types as tabulated in Table 7.1. Briefly, there are 19 environmental noises, 7 coding algorithms, 12 non-linear spectral subtraction (NLSS) configurations and 10 common speech processings, totalling to 48 degradations in the complete pool. None of the processes is deemed likely to improve intelligibility.

Four data sets are generated, namely one training set, two developmental test sets and one evaluation test set. The data sets are abbreviated as Train, DevI, DevII and Eval. All four data sets

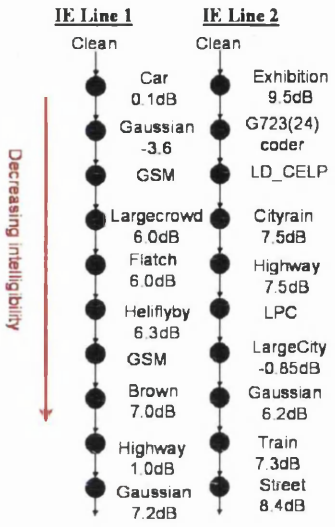


Figure 7.1: Two examples of IE lines with 10 processes (black nodes) per line. Processes are randomly chosen from the degradation pool. The input to the first node is a set of clean signals; output signals of each node become the input signals to the next node. 10 to 15 is typical number of nodes per line which is sufficient for output signals of the last node to be unintelligible.

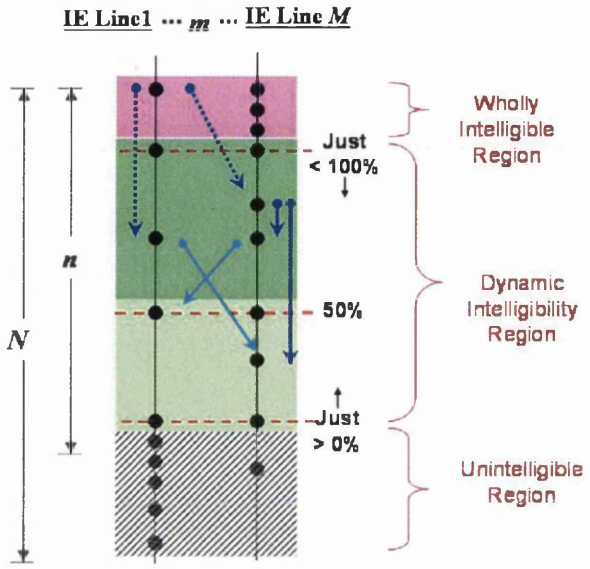


Figure 7.2: The SRT test is used to divide each IE line into sections of different intelligibility range. Signal pairs are identified using assumption (i), (ii) and (iii) stated on page 95 and 96, example pairings shown by blue, blue dotted and light blue arrows respectively. Signals from the Unintelligible Region are discarded (grey shaded). n is the number of node per line and is typically in the range of 10 to 15; m is the number of IE line needed to generate a data set such as the training set, typically 20 IE lines would produce >2000 pairs.

19 environmental noises	airport, babble, car, exhibition, restaurant, street, subway, train, aircraft, cityrain, flatch, gaussian, heliflyby, highway, ibmcoolfan, largecity, largecrowd, suncoolfan (at random SNRs in the range of 20dB down to -5dB).
7 coding algorithms	GSM, LD-CELP, LPC, MELP, G721, G723-24, G723-40
13 NLSS processes	various combinations of noise over estimations and noise floor (noise over estimation: 3.0 to 6.0; noise floor: 0.001 to 0.5)
10 common procedures	resampling, lowpass, highpass, vibro, flanger, pitch, phaser, chorus, fading, echo (at various rates).

Table 7.1: Degradations in the original pool.

are created in the same way but independently, i.e., from different sets of IE lines. A detailed list of degradations involved in generation of each data set can be found in Appendix A.2. The purposes of each data set and the degradations involved are as follows:

(i) Train: To train the classifier. A sub-pool of 23 degradations are used for its generation. The degradations include 15 environmental noises, 5 coding algorithms and 3 NLSS processes.

(ii) DevI: The same degradation pool for generation of Train is used here. However, signals pairs are deduced from independent sets of IE lines hence there is no overlap in Train and DevI due to different permutations of degradations when the IE lines are formed. However, since there is no unseen degradation good accuracy is expected. This test set aims to give confidence for the classifier setup.

(iii) DevII: The same degradation pool for generation of Train plus 13 unseen degradations are used here. The unseen degradations are 2 environmental noises, 1 coding algorithm, 5 non-linear spectral subtraction (NLSS) configurations and 5 common speech processings. This test set is designed to be sufficiently challenging so that improved accuracies help identify good features.

(iv) Eval: The same degradation pool for generation of Train plus 16 unseen degradations are used here. The unseen degradations are 4 environmental noises, 2 coding algorithm, 5 NLSS configurations and 5 common speech processings. 3 of the unseen degradations are in DevII's pool. This test set is to evaluate the robustness of features identified during development stage.

During development, goodness of a feature set is judged mainly on its performance for DevII rather than DevI. Meanwhile Eval set aims to predict the performance during normal usage of the D^{4}IC.

Table 7.2 shows the four data sets in terms of their number of IE lines and resultant pairs of nodes. All pairs have the known intelligibility relationship of $I_{S_A} \geq I_{S_B}$ where I_{S_A} is the intelligibility of first signal and I_{S_B} that of the second signal. In conventional classification of two-class problem, the classes are referred to as 'in-class' and 'out-of-class'. Here we use the class labels of 'YES' and 'NO' corresponding to the question 'Is the first signal more intelligible than the second?'. Initially all pairs correspond to YES class, therefore half of total number of pairs are inverted and labelled 'NO'. The 'YES' pairs and 'NO' pairs are wholly balanced and complementary, but not identical. Note that since the input to each IE line is a set of 566 clean utterances, the output of each node are also 566 utterances. These data sets are used for experimental work described in remaining chapters.

Data set	No. of IE lines	No. of nodes pairs
Train	35	5000
DevI	20	2000
DevII	20	2000
Eval	20	2000

Table 7.2: Number of pairs of nodes in each data set and the number of IE lines from which they are generated. Each node is 566 4-digits utterances. Half the pairs are labelled 'YES', another half labelled 'NO', corresponding to scenarios where the first signal is more intelligible than second and vice versa.

7.1.3 Generic and Specific Classifier

Critical to performance of the classifier is representativeness of the training data. In practice, for specific applications such as a codec design or optimisation of parameters during development of a new system, or selection of the best system among competing systems; the operational conditions are likely to be constrained to specific degradation types (for e.g., all kinds of codecs) and intelligibility range (for e.g., in the region of 75%). Therefore it is natural to constrain the classifier to work in this range too by biasing its training to the underlying condition. In that case we could adjust the SRT test to identify key nodes reflecting the desired intelligibility range, for example, 65% and 85% rather than 0% and 100%. In the same way, the range of degradation types might be constrained by selecting only relevant degradations for the degradation pool used in training. For instance, if the application relates to car noise environment, then components in the pool should be relevant, namely examples of different car noises.

Clearly, the more representive the training data, and the narrower and the more specific the operational ranges, the more accurate the classifier is likely to be. This idea of classifier tailored for specific application is discussed as future work. Here a general purpose classifier is implemented in order to gain confidence with building such a system and to have an understanding of selecting good features that give good classification accuracy.

7.2 A Preview of the Data Sets and D^{4}IC Features

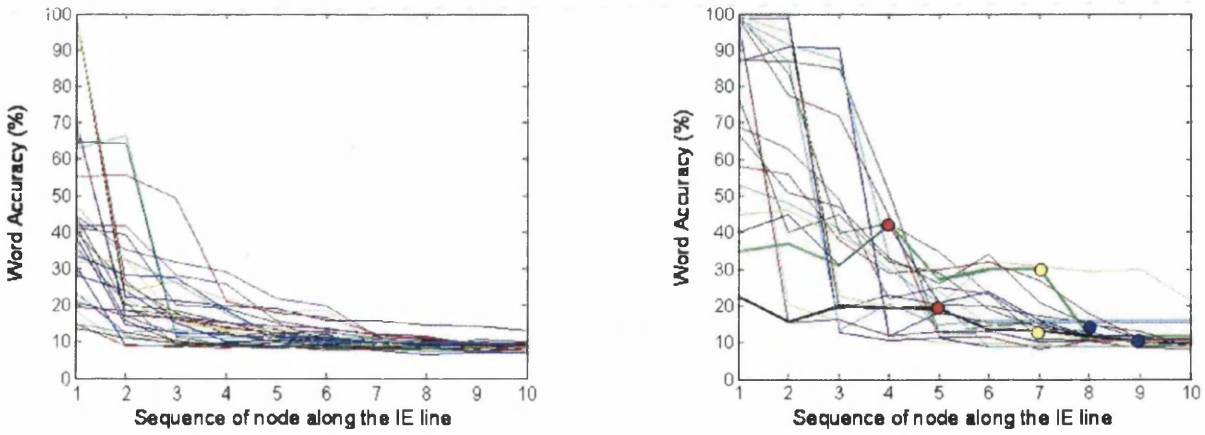
Chapter 6 proposes using measures considered in Part I as source of features for the D^{4}IC. These previous investigations suggest that WordAcc from the clean-trained ASR is potentially useful as intelligibility assessor compared to other measures considered. As a preliminary investigation of the data sets and D^{4}IC features, we look further at ASR WordAcc scores as potential features. Note that the D^{4}IC is not involved at this stage. Two data sets namely Train and DevII are chosen for illustration here. Figure 7.3 plots ASR WordAcc obtained for these two data sets as a function of increasing degradation along the x-axis (increasing degradation is due to the increasing number of processes (nodes) that the signals pass through).

The IE Hypothesis states that the profiles should decline monotonically as the number of processes that signals pass through increases (though rate of decline is unknown without listening test), hence if ASR WordAcc is to emulate human performance then its profiles should also decline monotonically. Clearly if the profiles are monotonic this should lead to potentially good features and in turn, lead to a good intelligibility assessor. However, notice that while both figures show overall decreasing trends from left to right, the profiles obtained for DevII are noticeably less monotonic as shown in Figure 7.3(b). This less-monotonic tendency could be due to the fact that there are more variety of degradations involved in the generation of this data set, and the fact that speech enhancement processes are among

the degradations involved. Recall from Chapter 5 that WordAcc of clean-trained ASR correlates less well with degradation introduced by the NLSS processes.

To further illustrate the limitation of ASR in this area, two arbitrary profiles in Figure 7.3(b) are marked with red, yellow and blue dots which are key-points of those lines that correspond to 100%, 50% and 0% intelligibility, as identified by informal listening test. As shown in Figure 7.3(b), the red dot is at node 4 of the green profile and at node 5 of the purple profile, meaning that the 4th process of the green IE line and 5th process of the purple IE line produce signals with intelligibility that begins to fall from 100%. These two red dots should correspond to similar ASR WordAcc since humans deem both to be around 100% intelligibility. However notice that while the first dot scores around 42% ASR WordAcc, the red dot for the purple line scores less than its half at approximately 19% ASR WordAcc.

In producing these statistics we gain insight into the difficulty of the data sets; in other words, how challenging they are for the intelligibility assessor. DevII is meant to be challenging so that when the other factors such as classifier, training data and test data remain unchanged, improved classification accuracy could help identify the good features. Besides confirming the difficulty level, Figure 7.3 points out how much ASR scores depart from human perception of intelligibility, especially when presented with a challenging test set. This hints at possible impediment faced by the subsequent classifier. Nonetheless, with larger range and variety of features plus extensive training, the classifier might achieve better performance and this is investigated in Chapter 8.



(a) (b)

Figure 7.3: ASR WordAcc against node count for (a) Train set and (b) DevII test set. Scores are given by a clean-trained ASR. Profiles in (b) are less monotonic than (a). Red, yellow and blue dots in (b) refer to key nodes of those two IE lines where signals correspond to 100%, 50% and 0% intelligibility according to informal SRT test. Dots of the same colour should correspond to similar WordAcc scores but figure shows discrepancy. Notice that one red dot corresponds to 42% WordAcc and the other only 19%.

7.3 Classifier

Various classification approaches are available, such as support vector machine (SVM), Gaussian mixture models (GMM) and neural networks (NN), there are also simpler ones such as the K-nearest neighbour (K-NN). Each has their strengths and limitations. The choice of classifier depends on the characteristic of the data set including the amount of training and the spatial variability of the effective average distance between data samples. It also depends on nature of the problem or the question posed, for example in this case the question is 'is A more intelligible than B?'. Choosing a suitable classifier assists in classification performance. For example, there are common beliefs that discriminative classifiers such as SVMs tend to be more accurate than generative ones such as GMM [].

The SVM is introduced by Vapnik [112]. It is essentially a binary classifier searching for an optimal decision boundary in two classes of data. The boundary is a hyperplane which maximises the margin between in-class and out-of-class data. Given the 2-class nature of the classification task attempted here, a discriminative model such as SVM is deemed appropriate. Thus throughout the work presented here the D^{4}IC uses an SVM classifier.

7.4 Feature Structure

Features-based experiments include utterance level, word level and frame level analysis, reflecting the position of score integration along the time courses. For the utterance level, scores are integrated across the full 4-digit utterance prior to becoming a feature element to the D^{4}IC; For the word level, individual digits in each utterance are considered, these are scores deduced from ASR word recognition hence is referred to as word level features; and finally the frame level features are raw features coming from front-end processes such as cepstra-based features used in the ASR.

In summary, the dimension of the raw features (prior to pre-processing and prior to training the D^{4}IC) is $m \times n \times p$ where m is the vector order, n is the number of vector across single utterance reflecting the level of integration across the time axis in creating the feature vector, i.e., n=1 when a feature vector is derived for each utterance; n=4 when derived for each digit across the 4-digit utterance and lastly n=N when derived for each frame where N is the number of 25ms frames across each of the 566 utterances. Lastly, p is the number of 4-digit utterances degraded by the same processes, i.e., produced by the same node.

Several example options of feature dimension are illustrated in Figure 7.4 where the corresponding dimension is given below each condition in the form of $m \times n \times p$ where m is the y, n the x and p the z dimension. Example (a) and (b) represent full integration across time and the 566 utterances, the difference is that condition (a) has just 1 feature component per vector, i.e., m=1; while condition (b) has more where m is more than 1 and up to 39. Example of m=1 is when each of the measures

(for e.g., PESQ, CSNR, etc) is considered individually as is done in original benchmark experiment in Section 7.4, in which case when all nine quality measures are fused then m would be 9. In example (c) integration is across the time course but not across the 566 different utterances. Example (d) is as (c) but with integration over each digit in the 4-digit utterances hence 4 vectors are derived for any single utterance. Lastly, example (e) is without such score integration and provides the largest feature element into the D^{4}IC. Largest feature dimension is $39\times N\times 566$ where 39 cepstra-based features are extracted from each frame from each utterance, N is the number of frames in each utterance which ranges from 126 to 392 across the 566 utterances (equivalent to signals of length 1.28s to 3.94s) with average at 192 frames. Clearly then when such big dimension is obtained some form of pre-processing is needed to concise the features in a meaningful way. One strategy is the GMM supervector approach used in speaker verification where the vector of stacked means serves as features for the SVM. This will be explained in more details in Chapter 8.

The integration over 566 utterances was driven by the ASR situation and of course, by the wish to include ASR-based features into the D^{4}IC. This is because a large number of test utterances are needed to obtain meaningful statistics from ASR. For instance, if standard ASR statistic such as WordAcc is used as feature, a set of 566 test utterances gives WordAcc in the range of 0 to 100% whereas recognition of single digit string would only produce either 0% or 100% corresponding to successful and unsuccessful recognition. Clearly if standard ASR statistic is not employed then this integration over large number of utterances needs not apply. Nonetheless, some experiments without ASR-based features apply the same integration for ease of comparison. In the benchmark experiment presented in next section, dimension shown in example(a) in Figure 7.4 is used.

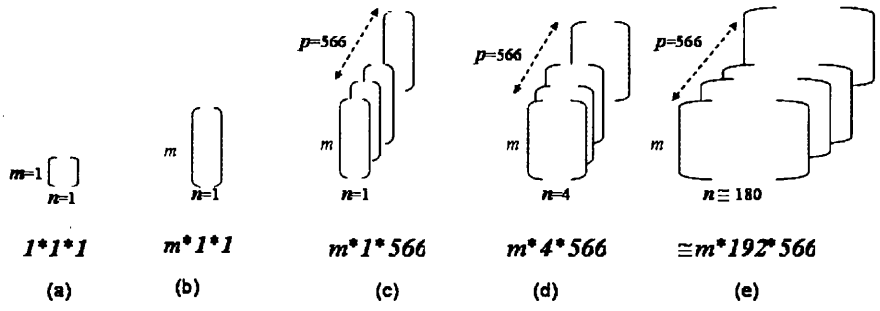


Figure 7.4: Examples of dimensions of features for D^{4}IC, dimension is given in the form of $m(vertical)\times n(horizontal)\times p(diagonal)$ below each example where m is vector order; n the number of vector derived from each utterance and p the number 4-digit utterances processed the same way. m ranges from 1 to 39, n is either 1, 4 or N where N is the number of frames in each utterance which ranges from 126 to 394; lastly, p is either 1 or 566 depending on whether scores are integrated across the 566 utterances processed by the same condition. See text for details.

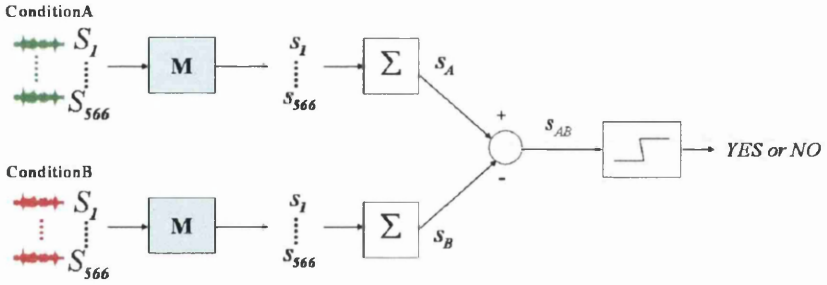


Figure 7.5: Part I measures are used to assess the new test sets. M is a measure such as PESQ; s_A and s_B are average scores given by a chosen measure to the 566 utterances under each condition. These two scores are compared where a positive s_{AB} implies that ConditionA signals are more intelligible than ConditionB signals, and vice-versa.

7.5 Benchmark Experiment

7.5.1 Pre-benchmark Tests

Here the 2 development test sets are assessed using measures considered in Part I namely 9 quality measures and 5 standard statistics coming from a clean-trained ASR. Figure 7.5 illustrates the assessment approach where ConditionA and ConditionB are the two sets of signals under comparison, the sets are processed in two different ways and each set consists of 566 4-digit utterances; M is a chosen measure such as PESQ which outputs an intelligibility estimation for each signal. The 566 scores for each condition are averaged as s_A and s_B respectively, which are subsequently compared where positive difference implies that ConditionA signals are more intelligible than ConditionB signals, and negative difference implies otherwise. The deduced relationship is matched with the labels provided by test sets DevI and DevII. Every correct match scores a *correct* and accuracy is defined as total *correct* over total number of tests. Note that D^{4}IC is not involved in this test. In order not to confuse this accuracy with classification accuracy achieved by D^{4}IC in latter sections, this accuracy is referred to as percentage correct instead.

Table 7.3 shows percentage correct obtained with the Part I measures for the 2 development test sets. Results show that DevII always gives lower percentage correct compared to DevI. This is expected since DevII consists of more variety of degradations and more cases involving degradations coming from enhancement algorithms. Among all quality measures WSS is the most accurate for both test sets which agrees with the observation made in Chapter 4. However, in contrast to the relatively poor performance of SNR-based measures in Chapter 4, their results are not exceptionally poor here. This is perhaps due to the way the database is formed, where pairs of signals under test usually involve one being degraded by more environmental noises. In a way the comparison becomes comparing the amount of noise present rather than comparing the different types of noise. Lastly, the performance of all ASR outputs with the exception of Subst and Ins are generally better than the quality measures.

(a)

Measure	Percentage Correct	
	DevI	DevII
PESQ	81.6	71.6
MNB	70.9	69.4
MBSD	74.7	61.4
WSS	85.1	78.5
IS	74.3	62.3
LAR	76.2	65.9
LLR	74.5	63.8
SNR	78.3	63.6
SEGSNR	78.9	66.9

(b)

Measure	Percentage Correct	
	DevI	DevII
Word Accuracy	86.3	79.1
Correct	89.5	76.1
Deletion	80.8	76.8
Substitution	68.8	57.4
Insertion	69.2	59.5

Table 7.3: Percentage correct obtained with Part I measures for the new test sets.

In overall WordAcc and Corr give the highest percentage correct which agrees observation in Chapter 5.

7.5.2 Benchmark D^{4}IC Experiment

The best performing measures for DevI and DevII according to Table 7.3 are Corr and WordAcc respectively with percentage correct at 89.5 and 79.1. Here Corr and WordAcc are used as feature for the D^{4}IC in the benchmark experiment. Experiment results would serve as baseline performance of the D^{4}IC.

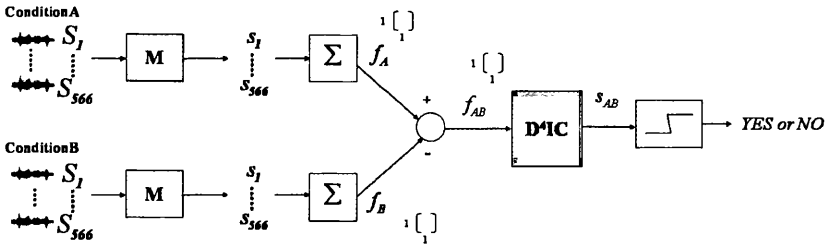


Figure 7.6: Setup for D^{4}IC benchmark experiment. Same as Figure 7.5 except that s_{AB} is used as feature for the D^{4}IC and decision is made based on the D^{4}IC output.

The benchmark experiment approach is similar to that illustrated in Figure 7.5 except that s_{AB} is not used directly for decision making, but serves as features to the D^{4}IC instead, as denoted by f_{AB} in Figure 7.6. M is a feature generator which, in this benchmark experiment, is a clean-trained ASR. ASR Corr is used as feature when testing with DevI, meanwhile WordAcc is used when testing with DevII. An ASR Corr or a WordAcc is produced for each set of 566 signals for each condition, resulting in f_A and f_B. The feature difference, f_{AB} is of dimension is $1\times1\times1$ which means that the scores are

integrated across the 4 digits as well as the 566 utterances. The D^{4}IC predicts the class (YES class or NO class corresponding to the question: is ConditionA signals more intelligible than ConditionB signals?) that f_{AB} belongs, the deduced relationship is then matched with the labels provided by the data set. A *correct* is scored if the deduced relationship matches with the label given and vice-versa. The accuracy is defined by Equation 7.1. Table 7.4 shows baseline performance of the D^{4}IC for test sets DevI and DevII.

$$Accuracy = \frac{\text{total } correct}{\text{total pairs under test}} \tag{7.1}$$

	DevI	DevII
Baseline	90.5	79.9

Table 7.4: Baseline results of D^{4}IC for DevI and DevII.

Chapter 8

Feature Assessments

This chapter is dedicated to the assessment of features that are potentially discriminative of relative intelligibility. There are two distinct characteristics about the test signals used in this classification task in this research: (i) the pair of signals under comparison originate from 1 signal which is processed in two different ways, hence the comparing signals are identical in terms of speaker, content, length and number of frames; the difference is the intelligibility level of each signal; (ii) signal contents including the exact 4 digits spoken in each utterance and the digit that each time frame corresponds to, are known. The first characteristic means that the pair of signals are directly comparable thus allowing signals difference to be investigated at all levels: from the highest, utterance level to the lowest, frame or even sample level. Meanwhile, the second characteristic enables the use of ASR-based scores as features which could be particularly beneficial due to the close link between word recognition and intelligibility.

Features from both high and low levels are considered. In the context here higher level features refer to features that are integrated over frames to make up words or the whole utterance, for example, scores coming from quality measures and ASR system. Meanwhile low-level features could be short term spectral or cepstral estimates. Hierarchically the features can be thought of as coming from utterance (by quality measures), word (by ASR system) and frame level respectively. Features from higher levels may inherit robustness from individual measures (quality measures or an ASR system), hence may lead to good performance especially when fused together. On the other hand, potential intelligibility-related information may have been lost during the processing imposed by these measures, in which case, the use of low level, holistic features may prove more useful. To the author's best knowledge no data-driven system that directly predicts intelligibility or relative intelligibility has been reported. Therefore it is hard to refer to or review on features used by others in this context.

This chapter presents a systematic assessments of features for the proposed differential classifier. Section 8.1 focuses on frame-based, low-level features; while Section 8.2 on features deduced from various stages of the word recognition process in an ASR system and section 8.3 on utterance-level features coming from the quality measures. In each section the features under investigation are described followed by experiments performed on series of features sets. In all sections feature performances are reported in terms of classification accuracy defined in Equation 7.1, which is the total *correct* over total pairs under test; where *correct* is scored when the classifier's estimation matches the label given by the database. Experimental approach is shown in Figure 8.1 where ConditionA and ConditionB are the two sets of signals under comparison (from a pair of nodes); FE is the feature extraction (FE) unit

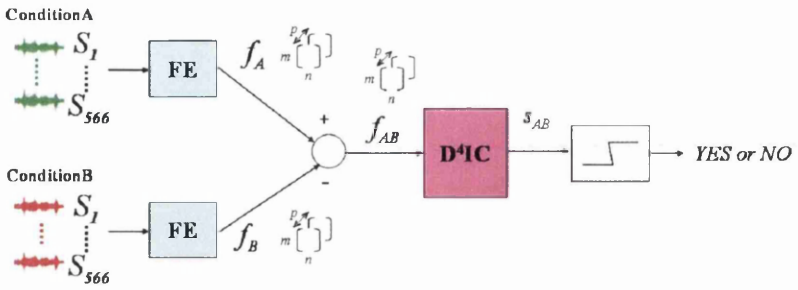


Figure 8.1: Experimental approach for feature assessments.

which contains processes that generate features for the D^{4}IC. The size of the feature vector, $m*n*p$ varies depending on the process involved in the FE. It is the content of this FE that this chapter is investigating so that the resultant feature difference vector, f_{AB} is robust for classification of differential intelligibility.

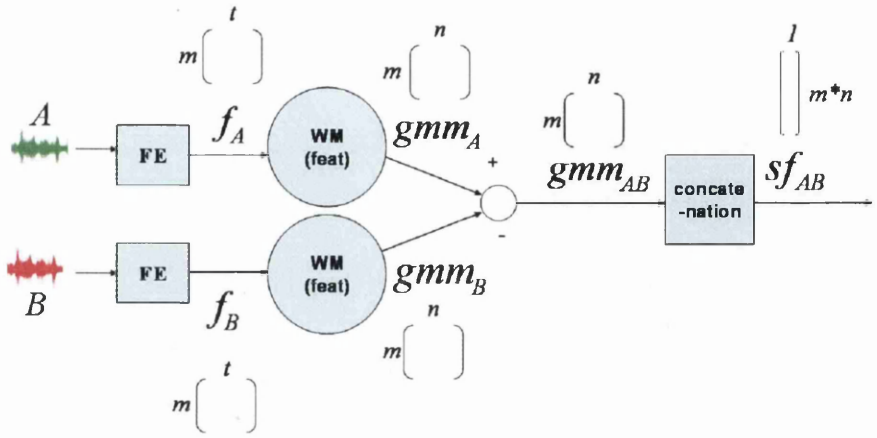
8.1 Frame Level Features

Frame-based, low level features can be considered as holistic features. Advantages of holistic features are that very little or no data is disregarded. There is little possible information reduction as it is not known what information is truly important for classifying relative intelligibility. Instead the classifier is counted on to determine the characteristic features. Holistic features have gained popularity in the field of face and handwritten word recognition reported in [113–116]. A recent investigation on features for footstep recognition also suggests the superiority of holistic features over geometric ones [117].

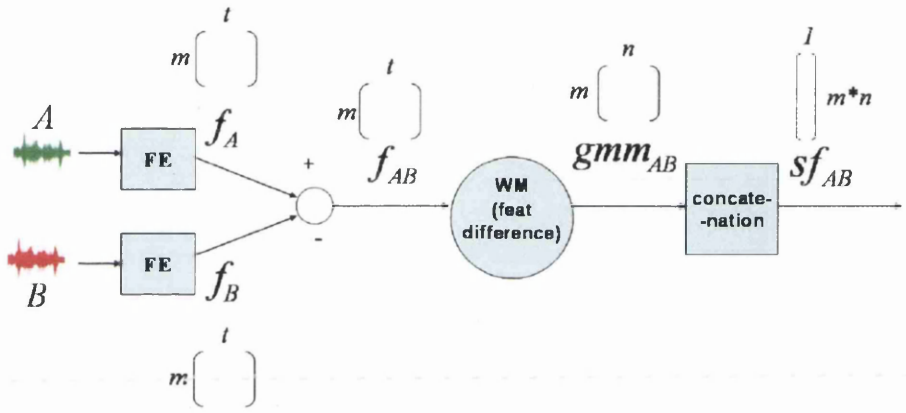
Two sets of frame level features are considered here. The first set is the same as that employed by the HTK reference recogniser used throughout this thesis. Full details can be found in [26]. Briefly, they are 12 cepstral coefficients (without the zeroth coefficient) and the logarithmic frame energy plus the corresponding delta and acceleration coefficients, totalling to 39 components per vector, hence the feature dimension is $39\times N\times 566$ where N is the number of frames in each utterance which ranges from 126 to 392 across the 566 utterances (average at 192 frames). The second set of features are weighted distances between the slopes of the reference (clean) and degraded spectra in each critical band as computed in WSS [76]. The feature vector for each frame is computed according to Equation 4.4 and 4.5 (in Section 4.1.3 of Chapter 4) except that integration of slope distances across the critical bands is not performed. The choice of this feature set is due to the outstanding correlation given by WSS in Part I. Note that whilst Klatt's original measure uses 36 critical-band filters, the WSS algorithm used in this thesis considers a bank of only 25 filters spanning the 4kHz bandwidth, hence 24 spectral slopes are computed per vector per frame (feature dimension = $24\times N\times 566$). The first set is referred to as ASRceps and the second as WSSspec.

Comparison are performed on a node-by-node basis so 566 output signals from each node are concatenated into one long utterance where total number of frames equals 108841. One problem with using frame-level features is explosion of feature dimension leading to a practical computational constraint. A method is needed to compress the features meaningfully. In speaker recognition, Gaussian mixture models (GMM) with universal background model (UBM) or world model (WM) have become the standard method where a speaker model is constructed by maximum a-posteriori (MAP) adaptation of the means of the UBM [118, 119]. Recent advances propose the idea of stacking the means of the adapted components to form a GMM mean supervector. The supervectors later act as features for a second-stage classifier. In other words the GMM is no longer the classifier but serves as a feature processing unit to give a mapped fixed length output. In particular the combination of this GMM supervector concept with a SVM (i.e., GMM supervector linear (GSL) kernel) has proved to give superior speaker verification performance compared to a standard GMM approach [119, 120]. This supervector approach is applied here for data compression purposes and to facilitate the use of an SVM as the classifier throughout.

Two approaches are taken here as shown in Figure 8.2 where f_A and f_B are matrix of frame-based features of signal A and B respectively (matrix size = $m \times t$ where m is the vector order which is either 39 or 24; and t is the total number of frames for the concatenated 566 utterances, i.e. 108841); and sf_{AB} is the supervector which serves as feature for the SVM classifier. In both figures WM refers to the world model, *gmm* refers to the adapted model and FE refers to the feature extraction unit where ASRceps and WSSspec are computed. The first approach is shown in Figure 8.2(a) where MAP adaptation is performed on f_A and f_B separately, producing an adapted model for each. Differencing is then carried out on the adapted models gmm_A and gmm_B resulting in gmm_{AB}. gmm_{AB} is of dimension $m \times n$ where n the number of mixture component per GMM. The means of this differenced model is stacked into a supervector sf_{AB} which serves as features for the SVM classifier. The second approach is shown in Figure 8.2(b) where the difference between the raw features f_A and f_B is computed prior to adaptation. Adaptation is then performed on f_{AB} producing gmm_{AB}. Obviously in this case the world model is also trained on differences between features of signals in pairs. As far as comparative intelligibility (i.e., which is more intelligible?) is concerned, the earlier differencing in the second approach might be beneficial. The number of Gaussian components, n considered ranges from 2 to 512. In total 4 feature sets are investigated: ASRcepsI and WSSspecI obtained using the first approach; ASRcepsII and WSSspecII obtained using the second approach.



(a) ApproachI



(b) ApproachII

Figure 8.2: Two different supervector approaches employed to compress the frame-level features: figure (a) illustrates ApproachI where world model is trained on frame-level features, adapdation performed separately on f_A and f_B and then differencing is performed on the adapted GMMs; figure (b) illustrates ApproachII where difference between f_A and f_B is computed prior to adaptation, the world model is also trained on differences of raw frame-level features. With both approaches, the means of the gmm_{AB} are concatenated into a long vector sf_{AB} serving as features for the SVM classifier. In both figures, brackets show dimensions of the features at each stage, t is the number of frames in the signal, m the feature order per vector per frame, n the number of Gaussian components.

8.1.1 Results and Discussion

Figure 8.3 shows accuracy obtained for DevI and DevII as the number of GMM components, n increases from 2 to 512. DevI results are shown by solid profiles and DevII by dotted profiles. As expected DevI accuracy are higher than that of DevII's. All profiles show steady increment with increasing n however such trend slows down and majority of the profiles become stagnant at around n=64 to 128. Table 8.1 shows highest accuracy obtained with ASRceps and WSSspec for the two test sets over the range

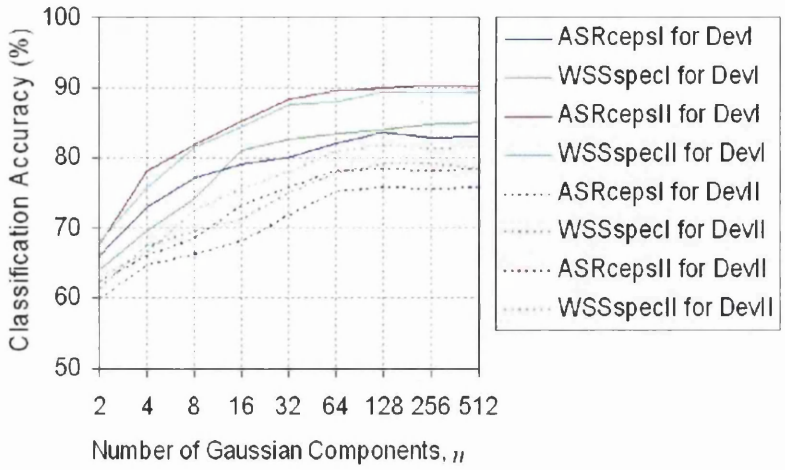


Figure 8.3: Classification accuracy versus the number of Gaussian components, n. Increasing accuracy are obtained as n increases however most saturates at around $n = 64$ to 128.

Feature	Accuracy	
	DevI	DevII
ASRcepsI	83.5	75.7
ASRcepsII	90.3	78.3
WSSspecI	84.8	79.1
WSSspecII	89.4	81.6

Table 8.1: Classification accuracy obtained using ASRcepsI, ASRcepsII, WSSspecI and WSSspecII. Results shown are the best obtained for DevI and DevII respectively over the range of Gaussian components, n considered.

of GMM components from 2 to 512. When compared to baseline results in Section 7.5.2, the biggest improvement is obtained by WSSspecII for DevII at 81.6 compared to the baseline at 79.9 (see Table 7.4 in Section 7.5.2). Besides, WSSspec seems to be slightly more robust than ASRceps when handling unseen degradations of DevII. All approachI results are worse than the baseline results and early differencing in approachII proves to be beneficial with almost all its results being higher than the baseline. Best accuracy are given by ASRcepsII at 90.3% for DevI and by WSSspecII at 81.6% for DevII.

In overall, though commonly used in other classification tasks, these short-term spectral-based and cepstral-based features do not seem to give particularly promising performance in this task. The search for robust features that could better discriminate comparative / relative intelligibility continues.

8.2 ASR-based Word-Level Features

This section aims to investigate the potential of various ASR-based scores including both its standard outputs as well as those internal scores obtained by harnessing various parameters at different stages of the word recognition process. Specifically, Section 8.2.1 looks at potential of the 5 standard outputs (namely Word Accuracy, Percentage Correct, Deletion, Substitution and Insertion) as features for the classifier; Section 8.2.2 investigates scores from recognition of short-term speech segments at frame level prior to forming the whole words; Section 8.2.3 exploits a parameter called N-Best hypothesis which can be seen as attempts (or guesses) made by human while trying to comprehend a signal; presumably the more intelligible a signal is, the sooner it could be recognised (i.e., at earlier rather than latter attempts).

8.2.1 Standard ASR Scores

Five standard ASR scores as investigated as features for the classification task. Definition for these standard ASR scores are as defined in Section 5.3. Each of the 5 features are first considered individually (feature dimension = 1×1×1) then combined (feature dimension = 5×1×1) . Table 8.2 shows performance of the features in terms of classification accuracy. Results show similar trends to those shown in Table 7.4 with Word Accuracy, Percentage Correct, and Deletion scoring significantly higher than Insertion and Substitution. All values here are slightly higher than in Table 7.4 showing confidence in the setup of the classification system. Using all 5 ASR standard outputs together gives 90.1 accuracy for DevI and 79.3 for DevII. These results show no improvement over those that already achieved with WordAcc alone for DevI at 90.5% and Corr alone for DevII at 79.9% for DevII. This is perhaps due to correlation between nature of these features or lack of good features in the combination. Though 90.5% accuracy for DevI is encouraging but almost all DevII results are at least 10% lower than those for DevI. This discrepancy shows that the features considered here are not sufficiently robust to generalise towards DevII which contains unseen degradations. The next section attempts to deduce more information from the recognition process by looking into lower-level scores produced behind these standard outputs.

8.2.2 Frame-level Recognition

In the ASR setup, each word is modelled by an 18-state HMMs (16 active states plus a begin and an end state). During training, features from short-term signals are allocated to different state models depending on the time frame they are in. During recognition, features extracted from each frame of a given test signal are matched against a 'network' of HMM states which can be imagined as a 3D matrix with choice of words on one axis, states belonging to the corresponding word model on the another axis and time scale on the third. Apart from the begin-state and the end-state which are

Feature	Accuracy	
	DevI	DevII
Word Accuracy	87.1	79.9
percentage Correct	90.5	76.1
Deletion	83.0	78.2
Insertion	71.5	60.6
Substitution	70.24	58.3
All	90.1	79.3

Table 8.2: Classification accuracies obtained when 5 standard ASR outputs are used as features. The 'All' configuration on last row means that all 5 features are used together.

non-emitting, every match incurs a log probability for observing that frame at the time and a log transition probability for moving from model state of the previous frame to current model state. In principle each new observation frame can be associated with any model state of any word model (under temporal constraints, state transitions to the left is not permitted). The path that the frames take to eventually conclude to a word is governed by the Viterbi algorithm, which finds the best path (i.e., sequence of state models) through the matrix to end up with highest possible joint log probability.

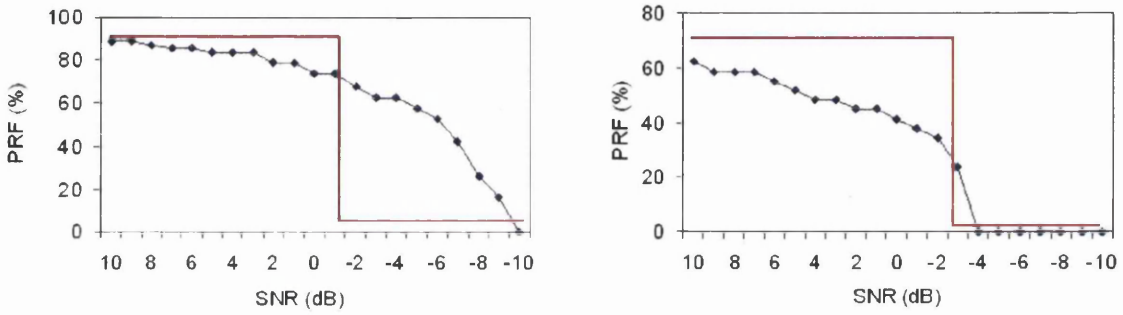
During this matching process a speech frame that is severely distorted could incur higher probability with a model state belonging to the wrong word model. The percentage of frames (out of total frames of a given word) that are correctly matched to model states of the target word model is thought to be strongly related to the amount of recognisable segments in the degraded signal, hence could be linked to intelligibility. If the word is eventually successfully recognised, this feature could indicate the level of confidence in the recognition; likewise if the word failed to be recognised, the feature potentially implies how close it is to successful recognition. This soft decision could be more informative than the binary hard decision given by the standard outputs where a whole word is either recognised or not recognised. The feature is referred to as percentage of recognised frames (PRF) and is defined as

$$PRF = \frac{1}{n}\sum_{i=1}^{n} s|_{\bar{M}_f = M_f} \tag{8.1}$$

where i refers to the ith frame and n the total number of frames of the given signal. $\bar{M}_f$ is the word model the state model of which matches ith frame as decided by the Viterbi algorithm, while M_f refers to the target word model according to transcription of the clean signal. s is a binary score of value 1 when $\bar{M}_f = M_f$ and of value 0 otherwise.

Two illustrative examples are shown in Figure 8.4 where PRF of an airport noise degraded digit '9' and a babble noise degraded digit '1' are plotted against SNR. Matched frames refer to frames that are associated (by the viterbi algorithm) to states of the word model for 'nine' and 'one' respectively. As shown in Figure 8.4(a) matched frames decreases steadily from around 90% at 10dB to 0% at -10dB. The abrupt staircase-like profile shows the hard decision of whether the word is eventually recognised

where high edge refers to successful recognition and low edge the opposite. Notice that though word recognition fails from -1dB downwards in Figure 8.4(a), there is only slight difference between PRF at 0dB (successful recognition) and at -1dB (unsuccessful recognition). In another example shown in Figure 8.4(b), correct recognition of digit '1' is obtained at SNR -3dB onwards. Below -3dB there is zero PRF and zero word recognition. Though the hard and soft decisions seem to coincide well in this example, the PRF is able to show the degree of confidence behind the successful recognitions. Again it is thought that this soft decision might give extra information on signal intelligibility.



(a) (b)

Figure 8.4: An illustration of gradual decrease of percentage of recognised frames (PRF) as opposed to hard decision of 1 or 0 given for recognition of whole word. In both figures the gradual profile is PRF, staircase-like profile is the hard decision where high edge means successful and low edge means unsuccessful word recognition. Notice that PRF decreases gradually as SNR decreases while word recognition score is abrupt.

Procedure

To obtain recognition status of each frame, the trace option of HVite (recognition tool of HTK) command is set to '-T 4'. This outputs a log of frames with their best tokens, which is a list of word models associated with each frame as determined by the Viterbi algorithm. The target word model of each frame can be found by testing the ASR system with clean signals. The two transcriptions (one from testing with degraded test signals and one from testing with corresponding clean signals) are matched and PRF computed accordingly. Recall that signals used are strings of four digits of the structure: [silence, digit, digit, digit, digit, silence] with optional short pause between the digits. The option is whether to include or exclude statistics from the silence periods (both the beginning and ending silence and short pauses between the digits) from calculation of the PRF. Both options are investigated.

During recognition the HVite uses a network that describes the allowable word sequences. The word network used in the Aurora2 framework and so far in this thesis specifies optional silence in beginning and end of the utterance and allows any amount of connected digits in the middle with

optional short pause between each digit. A new word network is specified where there is fixed silence both in beginning and end of the signal with fixed 4 digits in the middle and optional short pause between the digits. The task grammar [1] for this setting is as shown below where 'sil' refers to silence and 'sp' refers to short pause:

$digit = one | two | three | four | five | six | seven | eight | nine | oh | zero;
(sil $digit [sp] $digit [sp] $digit [sp] $digit sil)

This new word network is referred to as the restricted word net and the original one as the full word net. The use of both word networks are investigated. PRF is computed for each of the 566 utterances and the resultant feature dimension is 1×1×566.

Results and Discussion

Feature	Accuracy	
	DevI	DevII
PRF with silence & full word net	86.4	74.9
PRF with silence & restricted word net	88.5	77.4
PRF with no silence & full word net	90.7	80.3
PRF with no silence & restricted word net	91.4	82.5

Table 8.3: Classification accuracies obtained using frame-level recognition scores as features.

Table 8.3 reports classification accuracy obtained for percentage of recognised frames (PRF) with full word net (i.e., as in Aurora2 framework) and PRF with restricted word net (i.e., restricted to recognition of 4 digits with fixed silence in beginning and end; and optional short pause between digits), both with and without statistic from silent period in the calculation. Several observations can be made: firstly notice that exclusion of scores from silent period from calculation improves accuracies. This is probably because a clean-trained ASR having no knowledge of noise is likely to match a silent frame wrongly even when only the slightest noise is present, therefore unable to provide meaningful information about intelligibility. Second observation is that the use of restricted word net seems to be useful as shown by improved accuracies when comparing 1^{st} row with 2^{nd} row, 3^{rd} row with 4^{th} row of the table. This is not least due to the extra information given hence allowing the recogniser to perform better. Besides, the restriction imposed by the new word network ensures better alignment between the two transcriptions hence better use of PRF. As a whole, PRF with no silence and with restricted network gives best result for both test sets at 91.4% and 82.5% respectively. Third observation is that despite strong expectation for bigger improvement, this PRF-based features give only marginal improvement over the accuracies achieved by standard outputs in Table 8.2 where hard decision is used.

[1] The task grammar defines constraint on what the recogniser can expect as input. Vertical bars denote alternatives and square brackets denote optional items. The HParse tool [121] is used to create a word network from the grammar.

With the highest accuracies presented in Table 8.2 being 90.5% and 79.9%, it seems that harnessing the internal scores does not improve the accuracy much more than the standard output has to offer.

8.2.3 N-Best Hypotheses

As mentioned earlier the recognition process involves finding the path that gives the highest log probability through the network of model states. The recognition tool of HTK, HVite (refer [121] for details) uses a *token passing* algorithm for this purpose where a token represents a partial path through the network that spans from time 0 to time t. As the token propagates through the network, at each time step, the token can be copied to all possible connecting states and the log probability incremented by the corresponding transitions. At the end of time t all but the token with the highest probability are discarded. The multiple token passing option allows more than just one best token to be saved. The route that each token takes can be recorded and this gives information about alternative possible paths through the network. This option to record the token route histories and outputting the paths of N highest log probability is referred to as the N-Best hypotheses, which can be seen as analogous to the repeated attempts made by humans while trying to comprehend a message, especially a degraded one. Conceptually it can also be seen as the cognitive or thought process in the human auditory system while deciphering a speech signal. The logical assumption is that a perfectly intelligible signal would be recognised at first attempt (i.e., top hypothesis), subsequently the less intelligible a signal is, the more attempts are needed before it is recognised (i.e., at latter hypotheses). This section investigates fusion of the N best hypotheses as features for the D^4IC.

The trace option of HVite command (recognition tool of HTK) is simply set to '-n p N' for the ASR system to output N best hypothesis using p tokens in each state. p is fixed at 10 and the number of N investigated ranges from 1 to 10. One thing to take note is that the N-Best hypothesis setting tends to result in large number of insertions in the hypotheses. The insertions are normally made up of digit 'oh' or 'sp' (i.e., short pause) simply because they are the shortest words in the dictionary. The different hypotheses therefore become largely similar apart from the insertions of 'oh' and 'sp' at different places. For example, the 5-Best hypotheses for two 5dB airport-noise degraded signals, '3O79' (three-oh-seven-nine) and '6815' (six-eight-one-five) are as shown in Table 8.4 where 'O' refers to 'oh' and 's' refers to short pause. In the example given in Table 8.4(a), 's' and the leading digit '9' are insertions; meanwhile in Table 8.4(b), 's', the leading 'O' and the ending '9' are insertions. Notice that if 'O' and 's' are ignored then all 5 hypotheses are the same thus the N-Best hypothesis fails to provide any extra information.

A new word network which specifies fixed 4 digits between a beginning and an ending silence is created. The recogniser is now 'forced' to output a 4-digit transcription for each test utterance. The new 5-Best hypotheses for the 5dB test signals, '3O79' and '6815' are shown in the third columns of Table 8.5. The 5-Best hypotheses of the same utterances degraded by airport noise at 7.5dB and 2.5dB

(a) Actual transcription: 3O79

N^{th} Hypothesis	Transcription
1^{st}	93O79
2^{nd}	93O79s
3^{rd}	93O7s9
4^{th}	93Os79
5^{th}	93sO79

(b) Actual transcription: 6815

N^{th} Hypothesis	Transcription
1^{st}	OZ8159
2^{nd}	OZ8159s
3^{rd}	OZ815s9
4^{th}	OZ81s59
5^{th}	OZ8s159

Table 84: Transcriptions obtained at the N^{th} hypothesis for two airport-noise degraded signals at 5dB. Actual signal is '3O79' for Table (a) and '6815' for Table (b).

(a) Actual transcription: 3O79

N^{th} Hypothesis	Transcription		
	2.5dB	5.0dB	7.5dB
1^{st}	O3O9	O3O9	3O79
2^{nd}	93O9	93O9	9379
3^{rd}	OO79	3O79	O379
4^{th}	OOO9	O397	93O9
5^{th}	9O79	9379	O3O9

(b) Actual transcription: 6815

N^{th} Hypothesis	Transcription		
	2.5dB	5.0dB	7.5dB
1^{st}	O819	O819	9815
2^{nd}	OO19	9819	1815
3^{rd}	9819	O815	5815
4^{th}	9O19	9815	O819
5^{th}	O81O	O159	9819

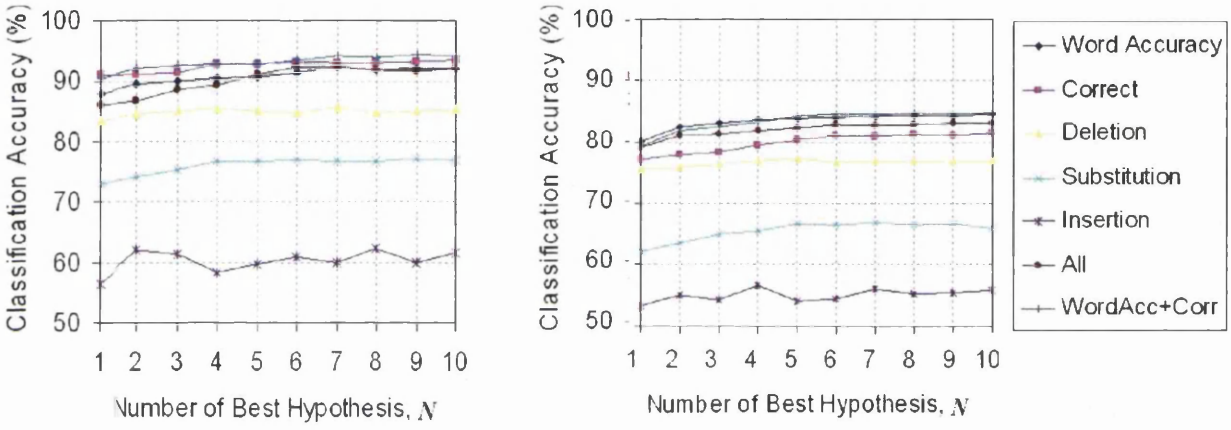
Table 85: Transcriptions obtained at the N^{th} hypothesis for two airport-noise degraded signals at 2.5dB, 5dB and 7.5dB. Actual signal is '3O79' for Table (a) and '6815' for Table (b). The best hypotheses at each SNR are underlined.

are also shown for comparison. Correct recognition of the best hypothesis at each SNR are underlined. As shown in Table 8.5(a), at 7.5dB correct recognition of all 4 digits of '3O79' are obtained at 1^{st} hypothesis. As SNR decreases to 5dB correct recognition only occurs at the 3^{rd} hypothesis. At 2.5dB, though the best recognition occurs at the 1^{st}, 2^{nd}, 3^{rd} as well as the 5^{th} hypothesis (3O9 at 1^{st} and 2^{nd}; O79 at 3^{rd} and 5^{th}), the recognition rate is only 3 (out of 4 digits). Table 8.5(b) shows similar trends where the highest recognition rate occurs at earlier hypotheses when the SNR is high and at latter hypotheses when the SNR is low. As shown the highest recognition rate at 7.5dB occurs at the 1^{st}, 2^{nd} and 3^{rd} hypotheses meanwhile at 5dB the same recognition rate occurs at the latter hypotheses of the 3^{rd} and the 4^{th}. Besides, as expected the recognition rate is better at higher SNR: 3 (out of 4 digits) at 7.5dB and 5dB, but only 2 at 2.5dB. Notice that if only single best hypothesis is considered, then the 2.5dB and 5dB signals would be deemed equally intelligible in both examples.

Five standard ASR outputs (i.e., Word Accuracy, Percentage Correct, Deletion, Substitution and Insertion) are computed for each set of hypothesis. Hence the setting of N-Best hypotheses would yield a feature vector of $5 \times N$ components for each pair of node. A range of N from 1 to 10 are investigated. The feature is referred to as N-Best, for example, 1-Best, 2-Best, and so on.

Results and discussion

Figure 8.5 shows classification accuracies against N-Best hypotheses for both test sets. Results of 1-Best are different from those presented in Table 8.2 because recognition here is restricted to 4 digits. As expected all DevI accuracies are higher than those of DevII. Notice that the order of the profiles given by various ASR outputs are the same with that indicated in Table 8.2 except that insertion gives lower accuracies than Subst here whereas the opposite is observed in Table 8.2. This could be due to restriction of 4-digits scoring which directly affects insertion scores. As in Table 8.2, in overall insertion and substitution give lower while word accuracy and correct give higher accuracies.



(a) DevI (b) DevII

Figure 8.5: Classification accuracies for (a) DevI and (b) DevII plotted against number of hypothesis considered.

Notice that with the exception of Ins, all other indicators give increasing accuracies as N increases. However, this increment saturates at various point of N. Among these 4 indicators the increase for Del and Subst are least obvious, both saturating at around 4-Best hypothesis. Note that WordAcc and Corr increase steadily till around 6 to 8-Best for both test sets. As in Table 8.2, the fusion of all 5 outputs seem to bring negligible improvements. The best accuracies are achieved by the combination of WordAcc and Corr. 94.3% is obtained for DevI using 7-Best and 84.8% is obtained for DevII using 8-Best, which are improvements of 3.9% and 4.9% respectively over the best result reported by single-hypothesis in Table 8.2.

On the whole, the N-Best hypotheses features motivated by its possible link to human's repeated attempts while trying to comprehend a message, proves to be beneficial especially when used in conjuction with standard ASR outputs of WordAcc and Corr. However, the improvement obtained is rather limited and performance seems to have reach its maximum at around 84.8% accuracy for DevII.

8.3 Quality-based Utterance-Level Features

Following Chapter 4 which observes that no single measure is entirely useless, the obvious progression is to combine the scores given by the various quality measures. The optimistic possibility of such fusion is that having gathered the inherit robustness from respective measures, the classifier would perform better than any one measure alone. The features from these quality measures are considered as from utterance-level because each measure computes one single score as indication of distortion/quality for one given utterance. The process to extract utterance-level features is simply by using the output scores of the quality measures as features into the D^{4}IC. The quality measures effectively function as feature generators in this context.

The experiment first investigates the quality measures individually one by one. The feature extraction component comprises of a single quality measure such as PESQ and the resultant feature vector f_{AB} is a single-order vector (feature dimension = 1×1×1). Next, all nine quality measures are used and scores are concatenated into a 9-order feature vector (feature dimension = 9×1×1). Table 8.6 shows classification accuracies obtained for both development test sets. The configuration 'All' on the last row refers to combined features where feature vector f_{AB} is a 9-order vector where features are contributed by all nine quality measures.

Feature	Accuracy	
	DevI	DevII
PESQ	80.5	71.3
MNB	69.2	70.3
MBSD	74.5	61.3
WSS	84.5	78.3
IS	76.0	68.0
LAR	76.1	69.1
LLR	76.1	70.0
SNR	78.1	62.6
SEGSNR	79.1	67.1
All	86.1	78.1

Table 8.6: Classification accuracies obtained when quality scores are used as features for the D^{4}IC.

Results for individual measures agree with baseline result shown in Table 7.3 in Chapter 7 showing again confidence in the classifier built. WSS and PESQ remain the best performing among all measures. Considering features from all quality measures achieves a slight improvement from 84.77% of WSS to 86.09% for DevI but no improvement is seen for DevII. In overall the features considered under this section are not sufficiently robust for classification of relative intelligibility.

8.4 Concluding Remarks

In this chapter various potential features for the D^{4}IC are investigated. In ascending order features considered range from frame-level spectral and cepstra-based features to word-level ASR-based features, to utterance-level features given by the quality measures.

Firstly, common spectral and cepstra-based features that are widely used in other speech classification tasks do not seem to be sufficiently discriminative of differential intelligibility. Highest accuracies obtained using the frame-level features are 90.3% for DevI and 81.6% for DevII by ASRcepsII and WSSspecII respectively.

Secondly, scores stemming from the quality measures do not seem to be particularly useful as D^{4}IC features. This perhaps is not suprising since the poor correlations between human intelligibility and objective quality scores have already been reported in Part I. Highest accuracies are obtained using WSS scores as features, with accuracies at 84.5% for DevI and 78.3% for DevII. The fusion of all 9 measures only marginally improves the accuracy for DevI to 86.1%.

On the other hand, both ASR-based feature sets namely the PRF and N-Best give improvement over the baseline accuracies. Briefly, the PRF relates to frame-level recognition in the word recognition process which is thought to link directly to the amount of recognisable segments in a signal; meanwhile the N-Best is thought to represent human's repeated attempts at comprehending a signal where signal intelligibility is related to how soon (earlier hypothesis) or late (later hypothesis) the signal is recognised. The N-Best feature set in particular scores 94.3% and 84.8% accuracy for DevI and DevII respectively (equivalent to 3.8% and 4.9% improvement over baseline).

In overall the ASR-based feature sets give higher accuracies since afterall there is a close link between word recognition and intelligibility. The next chapter introduces a novel feature set based on ASR.

Chapter 9

Anchor Models

Findings in Chapter 8 suggest that ASR-based features are potentially robust for classification of differential intelligibility. Improvements of classification accuracies are obtained through harnessing of parameters relating to the word recognition process where signal intelligibility is represented by various forms of ASR success. Modest improvements are achieved. This section proposes an alternative ASR-based feature set that represents the intelligibility of a given signal in relative terms to intelligibility of signals degraded by other (specifically chosen) degradations. This can be seen as similar to the anchor modelling technique used in speaker identification/verification where a speaker is represented, not in an absolute manner, but in relative manner to a set of chosen speaker models referred to as the anchor models. The anchoring feature set introduced in this chapter have brought the accuracy for test sets DevI and DevII to 96.7% and 91.2% respectively.

This chapter begins by introducing the concept of anchor models and proposing the use of differently trained ASR systems as the anchor models. Section 9.2 describes how the proposed feature set could potentially tackle two problems associated with objective intelligibility assessment as mentioned in Chapter 2. Followed is Section 9.3 which discusses a distinctive benefit that this feature set has for classification of differential intelligibility. Lastly, experimental procedures and results are presented in Section 9.4.

9.1 Introduction to Anchor Models

The anchor modelling technique is used in speaker tracking and indexing [122, 123], applied later to the field of speaker identification/verification by Yang et al [124]. Well trained speaker models are used as anchor or reference markers to represent a multi-dimensional space called the speaker reference space (SRS) where each dimension is a specific speaker. An utterance is projected into the space to get an observation for each anchor model. Eventually the utterance is represented by a vector of scores generated by a cohort of anchor models. One motivation for using anchoring features in speaker identification/verification is that there is limited speaker specific information but potentially unlimited information about other speakers which could supply external speaker-discriminative information to the system.

The proposed feature set intends to represent signal intelligibility in relative term to intelligibility of signals degraded by other chosen degradations. In other words, if we consider the intelligibility of

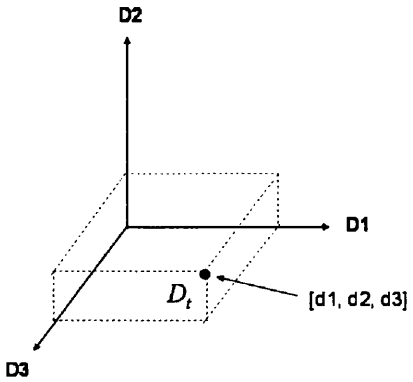


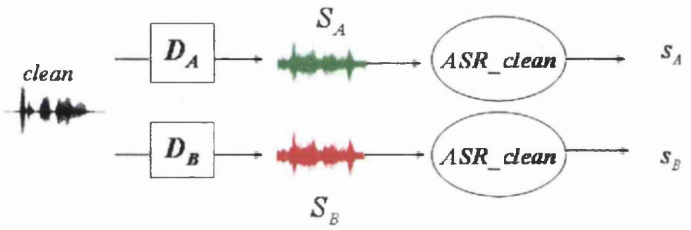
Figure 9.1: An illustration of the concept of anchoring features where each dimension of the feature space refers to degrading capability of a chosen anchoring degradation. The space is referred to as degradation reference space (DRS). Let D_t be the degradation under test (i.e. the degradation that the test signal has undergone) and D1, D2 and D3 the chosen anchoring degradation. Degrading capability of D_t (hence intelligibility of the test signal) is characterised by its location in the DRS.

a signal as being directly related to 'degrading capability' of the degradation it undergoes, then the feature set to be introduced can also be thought of as characterising this degradation in relative term to a range of other reference degradations. Hence in the context here, a degradation reference space (DRS) rather than a speaker reference space (SRS) is formed. Each dimension is directly associated with a specific degradation and a collection of N anchor scores defines a coordinate for a given utterance in the DRS. The concept is illustrated in Figure 9.1 where each dimension of the DRS relates to degrading capability of a chosen anchoring degradation. Let D_t be the degradation that a given signal undergoes and let the anchors be degradation D1, D2 and D3. Degrading capability of a D_t is characterised by its location in the DRS namely the coordinate [d1, d2, d3] (i.e., its distance to the axis of D1, D2 and D3). As shown in this particular illustration the coordinate of D_t has a small d2 value but larger d1 and d3 value, hence degrading capability of D_t is deemed most similar to that of D2 and less similar to those of D1 and D3 .

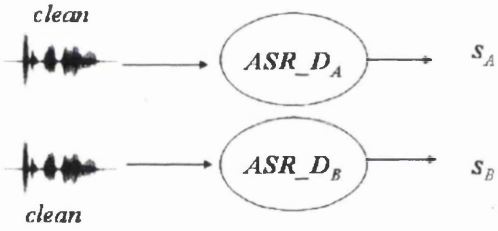
9.1.1 ASR Systems as Anchor Models

The idea for anchoring features is partly inspired by some observations during ASR experiments in Chapter 5. In section 5.4 clean-trained ASR was used to estimate signal intelligibility. Degraded signals were used as test signals and the clean-trained ASR as the model which the test signals match against. This scenario is shown in Figure 9.2(a) where *ASR_clean* is an ASR system trained on clean signals; S_A and S_B are test signals degraded by degradation D_A and D_B respectively; and s_A and s_B are ASR scores given for signals S_A and S_B respectively. When this scenario is reversed, i.e., clean signals being the test signals and the degraded signals the training data for the ASR systems, it is found that the ranking of ASR scores in both scenarios coincides. This is illustrated in Figure 9.2(b)

where ASR_D_A and ASR_D_B are two ASR systems trained on D_A-degraded signals and D_B-degraded signals respectively, similarly s_A and s_B are ASR scores produced when clean signals are tested against the ASR systems. It is observed that the ranking of s_A and s_B in both figures are the same (i.e., if $s_A > s_B$ in the first scenario, it is true in the second scenario too). This seems to suggest that, as much as score s_A in Figure 9.2(a) is used to indicate intelligibility difference between S_A the test data and clean signals the training data (or in other words, the difference of degrading capability between D_A and clean), the score s_A in Figure 9.2(b) is able to give the same indication. The same applies to score s_B in both figures.



(a)



(b)

Figure 9.2: Figure (a) shows experimental setup in Chapter 5 where degraded test signals are tested against a clean-trained ASR system, s_A and s_B are scores indicating intelligibility of S_A and S_B. When the scenario is reversed as shown in figure (b) where clean signals are the test signals and ASR systems are trained on degraded signals, the ranking of s_A and s_B is the same as that obtained in scenario shown in Figure (a).

As a progression to that finding, instead of representing a signal with scores coming from a single ASR such as a clean-trained ASR, the signal might be better characterised by scores coming from a range of differently trained ASR systems. This is illustrated in Figure 9.3 where the test signal S_t could be represented by a vector of scores coming from ASR_clean, ASR_D_A, ASR_D_B and so on. If each score indicates intelligibility difference between signal S_t and the training data of that ASR system (or in other words, the difference in degrading capability between D_t and the degradations that the training data underwent), a vector of such scores from a cohort of differently trained ASR systems could conceptually give a better indication of where D_t is in a space of degradations where D_A, D_B and clean are one of the many dimensions, in turn better characterising intelligibility of the signal S_t.

Some real illustrative examples are given here using data from test set $DS1_{add}$ which considers the 8 Aurora2 environmental noises, namely airport, babble, car, exhibition, restaurant, street, subway,

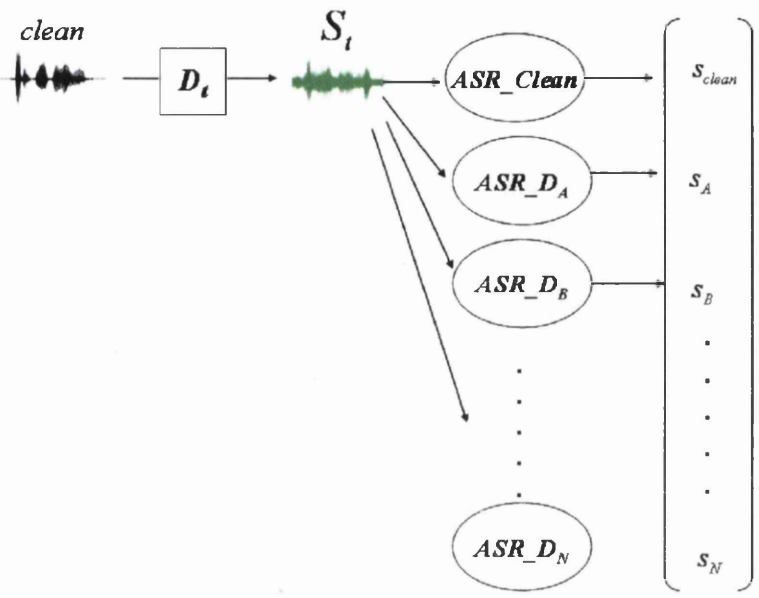


Figure 9.3: Intelligibility of the test signal S_t can be characterised by a vector of ASR scores where the ASR systems are trained on signals degraded by different chosen degradations. The vector can also be thought of as charactering degrading capability of D_t in a multi-dimensional space where clean, D_A, D_B, :. and D_N are one of the dimensions.

Abbreviation	Meaning
X signals	Signals degraded by degradation X (for example, car signals refer to signals degraded by car noise.)
ASR_X	an ASR system trained on signals degraded by degradation X (for example, ASR_car refers to an ASR system which is trained on signals degraded by car noise.)

Table 9.1: Abbreviations used in chapter.

and train (details can be found in Section 3.2 and Appendix A.1). The examples are to demonstrate the potential of ASR scores in indicating intelligibility difference between the test and training data of the ASR system, hence the potential usefulness of anchoring ASR scores. The examples are presented in 2D graphs rather than in a feature space of N-dimensions (where N is the number of anchoring degradations) for ease of observation. For ease of explanation, the abbreviations shown in Table 9.1 are used:

In the first example, exhibition signals are tested against six ASR systems trained on subway, street, car, train, airport babble signals respectively. Profiles shown in Figure 9.4(a) are word accuracies given by the cohort of ASR systems when tested with the exhibition signals over a range of SNRs. Different profiles are produced by differently trained ASR systems as shown in the graph legend, for example, the lowest profile in Figure 9.4(a) is given by ASR_babble (ASR system trained on signals degraded by

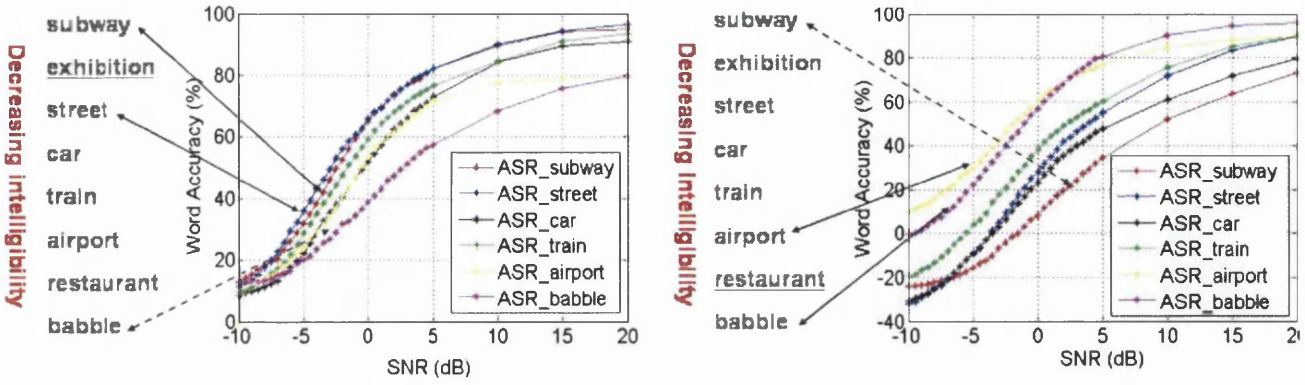
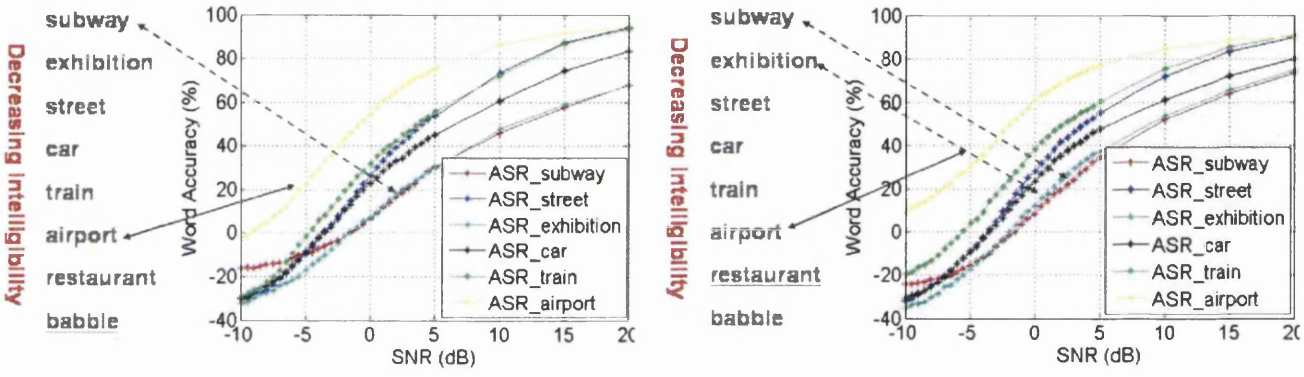


Figure 9.4: Illustrative examples of ASR scores indicating intelligibility difference between its training and test data. In both figures intelligibility ranking of the different degradations are shown on left in descending order from the top. Figure (a) uses exhibition signals as test signals, top profiles are given by ASR trained on subway and street signals; both degradations are next to exhibition on the ranking (pointed by solid arrows), lowest profile by ASR trained on babble as it is farthest from exhibition on the ranking (dotted arrow). Figure (b) uses restaurant signals as test signals, now top profiles are given by ASR trained on airport and babble signals while lowest is given by ASR trained on subway signals; this is because restaurant is nearest to airport and babble while farthest to subway according to the ranking.

babble noise). The intelligibility ranking (in descending order from the top) as perceived by human is placed on the left side of the profiles for easy referencing. Since exhibition noise is at the upper range of the intelligibility ranking, higher word accuracy is expected from ASRs trained on (signals degraded by) degradations at the upper range of the ranking too; likewise lower word accuracy is expected from ASRs trained on degradations at the lower range. As shown in Figure 9.4(a) the experimental outcome agrees with this expectation: notice that the top two profiles are produced by ASR system trained on subway signals and street signals respectively, both degradations are from the upper range of the ranking and are actually next to exhibition on the ranking as shown by the solid arrows. Notice also that the profile produced by ASR_babble is at the lowest among all profiles as shown by the dotted arrow. This is thought to reflect the big difference between degrading capability of exhibition noise and babble noise. Figure 9.4(b) shows profiles obtained using restaurant signals as test signals. As shown the top two profiles are now produced by ASR trained on babble signals and airport signals respectively, both degradations of which are from the lower range of the ranking as indicated by the solid arrows. This is expected since restaurant noise is also from the lower range. Notice also that the lowest profile is now from ASR trained on subway signals due to big difference between degrading capability of restaurant noise and subway noise as they are on opposite ends of the ranking.

More examples are given in Figure 9.5 with a different set of ASR systems, a selection of which are shown in the graph legend. Figure 9.5(a) shows results obtained with babble test signals and Figure 9.5(b) with restaurant test signals. Both graphs show that the test signals associate best with

ASR trained with airport signals. Again this agrees with expectation since airport noise is the next closest to both babble and restaurant noise according to the ranking. Notice also that the profiles Figure 9.5(a) are generally lower than profiles in Figure 9.5(b) implying that babble noise is more degrading than restaurant noise. Though the examples given above are not statistically conclusive, they highlight the potential effectiveness of the anchoring features in characterising signal intelligibility (or degrading capability). The next section discusses how anchoring ASR scores could potentially improve classification by addressing directly two difficulties associated with intelligibility assessment.



(a) (b)

Figure 9.5: Same as Figure 9.4 but a different cohort of ASR systems are used as shown in figure legend. Babble signals are used as test signals in Figure (a), and restaurant signals in Figure (b). The two noises are next to each other on the ranking hence both score highest with ASR trained on airport signals which is the nearest to the two noises. However, notice that the top profile in Figure (b) is slightly higher than that in Figure (a) since restaurant signals is more intelligible than babble signals at the given SNRs.

9.2 Tackling two difficulties of intelligibility assessment

In Chapter 2, three difficulties/challenges anticipated in objective intelligibility assessment are discussed. Briefly, the difficulties are :

(i) Small dynamic score range: scores given are constrained to small range due to operation under high degradation, for example, signals degraded to 5dB car noise corresponds to 100% human intelligibility but only 25% of PESQ's full score.

(ii) Confusing degradations: machine-based scores can easily be improved by processings such as enhancement algorithms. This, however, normally does not mean improvement of speech intelligibility. The term 'confusing' is used to describe that the fact the machines are 'confused' by these artificial signal enhancement that generally cause intelligibility to fall yet boost up scores.

(iii) Lack of reliable ground truth: reliable ground truth is difficult to establish but is essential for development and evaluation of intelligibility measures. This difficulty is addressed by introduction of IE Hypothesis and the subsequent data generation scheme in Chapter 6 and 7.

The anchoring features could potentially address difficulty item (i) and (ii) in the following manner:

(i) Small dynamic score range: this difficulty is eased when the ASR systems (acting as anchor models and feature generators) are trained on degraded signals (chosen degradations acting as anchors). By doing so the dynamic range of ASR scores obtained are increased as compared to those obtained previously from either quality measures or a clean-trained ASR system. Each feature can now be investigated at more meaningful dynamic range, for instance, car signals at 5dB and -5dB yield 17% and 11% respectively with a clean-trained ASR system but yield 65% and 23% with an ASR system trained on car signals. Each dimension of the degradation reference space (DRS) or feature space can be imagined as having been increased in resolution

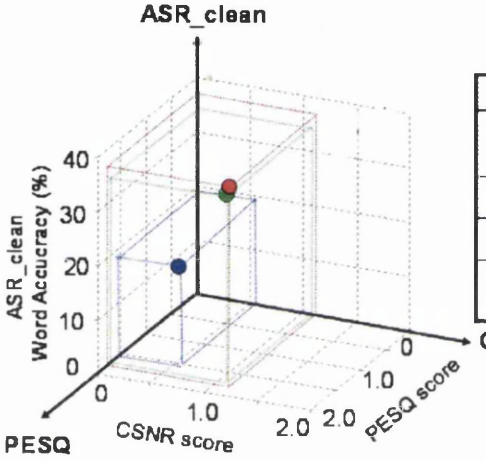
(i) Confusing degradations: this difficulty is eased when the 'confusing' degradations can potentially be better characterised using the anchoring features. Theoretically the usefulness of anchoring features shall increase as the number as well as the variety of characteristics of the anchoring degradations increase meaningfully. In fact, if among the anchoring degradations exist those that are similar to those 'confusing' degradations, then the corresponding anchor models could detect the presence of these 'confusing' degradations, hence avoiding the confusion since the classifier has learnt about such pattern during training.

Two illustrative examples are given in Figure 9.6 and 9.7. Three set of differently degraded test signals used in the illustration are:

- A: car signals at 5dB;
- B: car signals at 0dB;
- C: car signals at 0dB processed by NLSS
(noise over estimation=0.001, noise floor=3.0),referred in short as the 0dB NLSS signals.

The intelligibility relationship of these signals as perceived by human are A >B >C according to listening tests reported in Chapter 3. Figure 9.6 shows the feature space obtained (not normalised) when the FE (feature extraction unit of the classifier) consists of PESQ, CSNR and a clean-train ASR as the feature generators. Features of signals A, B and C described earlier are represented by red, blue and green dot respectively in the feature space. Signal intelligibility are characterised by scores given by these three measures, which are also the coordinates for their locations in the space. The scores are tabulated next to the figure alongside with human scores for these signals. First of all notice that while human scores (5th column) decreases from A to B to C with signals A having 100% intelligibility, signals B approximately 90% and signals C approximately 75% intelligibility; all 3 objective measures

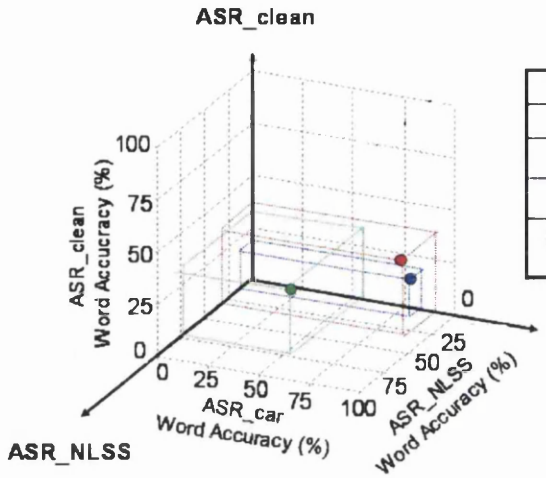
deemed signals C as having higher intelligibility than signals B due to the enhancement process which improves scores artificially without improving intelligibility. As shown in the feature space the red dot and green dot are overlapping. This suggests that features considered here are not able to discriminate 'real intelligibility improvement' caused by increment of SNR (from 0dB to 5dB when comparing signals A with B) from 'artificial intelligibility improvement' introduced by the NLSS process.



	Features			HUMAN (%)
	ASR_clean word accuracy (%)	PESQ score	CSNR score	
5dB car signals	38.13	1.81	1.10	100
0dB car signals	18.95	1.51	0.53	approx. 90
0dB car signals + NLSS	34.44	1.79	1.12	approx. 75

Figure 9.6: Figure illustrates feature space obtained if PESQ, CSNR and ASR_clean function as feature generator for the classifier. The 3 coloured dots represents locations of 3 differently processed signals in the space, namely 5dB car signals, 0dB car signals and 0dB car signals plus NLSS process. All measures deem NLSS signals as more intelligible than 0dB car signals and as equally intelligible with 5dB car signals, which does not reflect human. Red and green dot occupy similar region despite difference in intelligibility. This is due to features that are less discriminative and fail to distinguish artificial enhancement introduced by NLSS.

Figure 9.7 shows feature space obtained when the FE component now consists of three ASR systems trained on clean signals, car signals and NLSS signals respectively, abbreviated as ASR_clean, ASR_car and ASR_NLSS. The feature space can be considered as a degradation reference space (DRS) where clean, car noise and NLSS degradation serve as anchors. First of all, notice that by ASR_car gives much higher score to signal A at 93.1% compared to 38.1% by ASR_clean showing that the score range has increased. Next, note that ASR_car gives lower score for signals C at 53.1% compared to 86.8% for signals B which is of the same SNR but without NLSS, this is because characteristic of the NLSS-ed signals departs greatly from that of signals degraded by just car noise. There is no 'confusion' between signals A and C; and the trend of ASR_car scores for all 3 signals agrees with that of human. Similarly, ASR_NLSS gives highest score to signals C as expected due to better match of train and test. The feature space shows that the dots are now spaced out. Whilst example in Figure 9.6 show that the features are not able to tell the difference between artificial improvement given by NLSS and real improvement obtained with increased SNR, here this difference is detected by both ASR_car and ASR_NLSS. Although ASR_clean gives similar scores for signals A and the C, the other two anchors are able to 'fix' this mistake. Therefore, confusion caused by confusing degradations such



	Features (ASR Word Accuracy (%))			HUMAN (%)
	ASR_clean	ASR_car	ASR_NLSS	
5dB car signals ●	38.13	93.12	30.76	100
0dB car signals ●	18.95	86.79	10.03	approx. 90
0dB car signals + NLSS ●	34.44	53.07	73.32	approx. 75

Figure 9.7: Same as Figure 9.6 except that the feature generators used are ASR trained on car signals, clean signals and NLSS-ed car signals respectively. Different degradations are distinctively detected by the ASRs: ASR_car gives high score for car signals while ASR_NLSS gives high score for NLSS-ed signals. The red and green dot are no longer overlapped as was in Figure 9.6.

as NLSS could effectively be removed as the classifier is 'taught' during training that feature pattern (i.e., coordinate) like that of the green dot corresponds to lower intelligibility than that of the red dot. Different degradations should now be better disctintively characterised.

For recap, the anchoring features provide distinct benefits which effectively tackle 2 out of the 3 problems anticipated with intelligibility assessment. The first benefit deals with the problem of constrained dynamic range due to operation under high degradation. The second benefit tackles the tendency of the classifier to fail when dealing with difficult or 'confusing' degradations such that speech enhancement which artifically improve intelligibility. Philosophically the first benefit can be thought of as increasing resolution of the feature space, this second benefit can be thought of as increasing its dimension.

9.3 Robust Features for Classification of Differential Intellegibility

Apart from the two distinct benefits mentioned earlier, good performance is expected from the anchoring features due to the fact that relative intelligibility between signals under comparison is already 'built' into features of each signal. For illustration, let S_A and S_B in Figure 9.8 be two signals under comparison where S_A is degraded by D_A and S_B by D_B. By having an ASR anchor model trained on D_B-degraded signals and acting as a feature generator in the FE (Feature Extraction) component, the ASR scores obtained, s_{AB} in particular already contains information about relative intelligibility between S_A and S_B (or, in other words, the mismatch or the difference in degrading capability of D_A (in test signal) and D_B (in training signal)). Similarly, feature vector for signal S_B, f_B would also

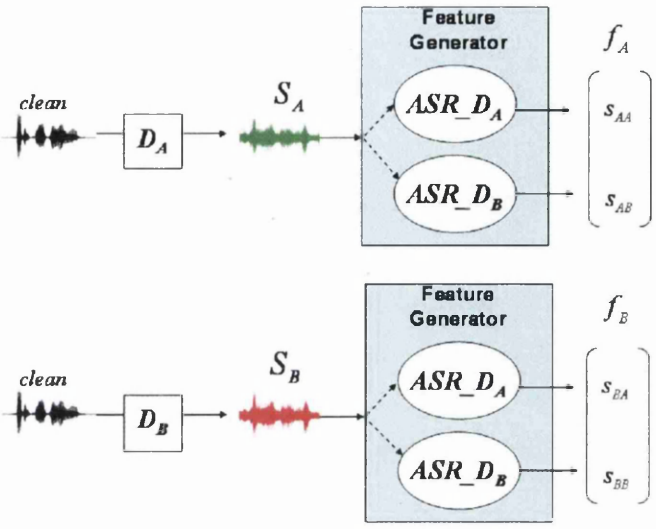


Figure 9.8: An illustration of how anchoring features bring the minus sign one step earlier. f_A and f_B has independent information about relative intelligibility between S_A and S_B even prior to the subtraction process.

have the same information when tested against an ASR trained on D_A degraded signals.

This 'cross-referencing' process could be beneficial since f_A and f_B independently on its own already possess information about relative/differential intelligibility with each other even before the subtraction process of f_A - f_B. Whereas with other features investigated previously, this differential information is only available after the obtaining the feature vector f_{AB} (i.e., result of f_A - f_B). Conceptually this anchoring approach can be thought of as bringing the differencing one step earlier.

Chapter 7 briefly mentioned the possibility of building a degradation-specific classifier where only relevant degradations are involved in generation of the IE lines and subsequently used for training. Now the idea can be furthered by using not only data, but also features that are specific to degradations of interest. This is done by using the degradations under test themselves as anchors, obviously this can only be realised when those degradations are available during assessment.

9.4 Experiment

9.4.1 Procedures

In this experiment, an anchor model is an ASR system trained on signals degraded by a chosen degradation. Having had a cohort of anchor models (i.e., a range of differently trained ASR systems), the features representing a given utterance are then a vector of ASR scores produced by the respective ASR systems. The ASR scores could be standard scores such as word accuracy or frame level scores

like the percentage recognised frame (PRF) considered in Section 8.2.1. Let an anchor model, λ trained on signals with degradation a be λ_a; and let the model output be $\lambda_a(i)$. A feature vector,f can be denoted by:

$$f = \left\{ \begin{array}{l} \lambda_a(1) \\ \lambda_a(2) \\ \vdots \\ \lambda_a(I) \\ \vdots \\ \\ \vdots \\ \lambda_z(1) \\ \lambda_z(2) \\ \vdots \\ \lambda_z(I) \end{array} \right\}. \tag{9.1}$$

where λ_a to λ_z are the cohort of anchor models and I is the number of scores from each model. For instance, $I = 5$ if the five ASR standard scores are used namely Word Accuracy, Percentage Correct, Deletion, Substitution and Insertion.

Ideally anchors should be chosen carefully in order to characterise the signal's location in DRS effectively. In the experiments carried out anchoring degradations used are all the degradations involved in the making of the training set of D^4IC namely Train. Recall from Section 7.2 that three types of degradations are involved in the making of Train, namely 15 background noises, 5 coding systems and the 3 NLSS configurations. Therefore in total there are 23 anchor models, one for each of the 23 degradations considered.

To build ASR anchor models for background noises, the 566 clean signals are degraded by chosen background noise at SNR 20dB, 15dB, 10dB and 5dB to -6dB at 1dB interval, totalling to 8490 training utterances for that particular ASR model. Meanwhile, the Aurora2 multi-train set is used to train ASR anchor models with coding degradations. The multi-train set consists of 8440 signals, one-fifth of which are clean signals, the other four-fifth are degraded by 4 different degradations (subway, babble, car, exhibition) at 4 chosen SNRs (20, 15, 10 and 5dB) respectively. To train ASR anchor models for this experiment, each of the one-fifth subset are divided into 3 smaller sets which are en-decoded once, twice and thrice respectively by a chosen speech coding system. These en-decoded signals train an ASR system which acts as an anchor model. The same procedure is applied to build anchor models with other coding systems. Lastly, to build ASR anchor models with NLSS degradations, the multi-train set divided in the same way are processed by the NLSS algorithm with two different configurations.

ASR scores from each of the anchor model are concatenated to form a long feature vector as

Measure	DevI	DevII
Set I: All 5 standard outputs (115)	96.0	89.4
Set II: WordAcc and Corr (46)	96.7	91.2
Set III: WordAcc and Corr with 5-BEST hypothesis (230)	97.0	91.6

Table 9.2: Classification accuracy obtained using anchor feature sets.

illustrated by Equation 9.1. The choice of ASR output from each anchor model could be standard outputs or other ASR-based outputs such as frame-based recognition scores like those investigated under Section 8.2.1. Three different sets of outputs are investigated here: (i) all 5 standard outputs from each ASR anchor model; and (ii) WordAcc and Corr from each ASR anchor model; (iii) 5-BEST hypotheses in terms of WordAcc and Corr from each ASR anchor model. They result in feature dimension of (i) 115×1×1; (ii) 46∗1∗1 and (iii) 230×1×1 components respectively.

9.4.2 Results and Discussion

Table 9.2 shows classification accuracy for the anchoring features. As shown The potential of anchoring feature is well-demonstrated here with accuracy as high as 96.7% and 91.2% for DevI and DevII respectively. Worth noting is the improvement achieved for DevII when compared with the 79.9% accuracy reported in Table 8.2 where only outputs from a clean-trained ASR system are considered.

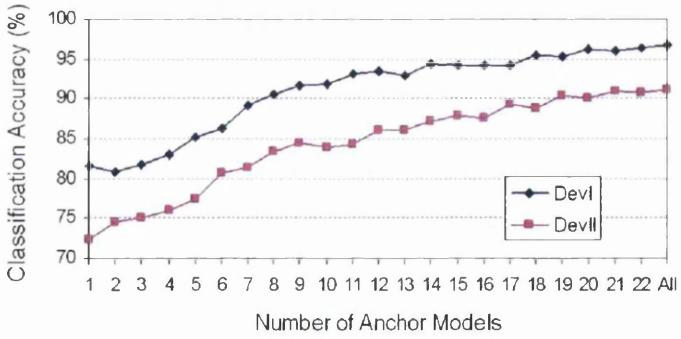


Figure 9.9: Classification accuracy plotted against the number of anchors.

The N-best option has boost accuracy from 90.5% to 94.3% for DevI and 79.9% to 84.8% for DevII in the experiments reported in section 8.2.3. Despite anticipation for the same improvement, the N-Best option in this experiment has not proved to outperform the single-best output option significantly. On the other hand, improvement is observed as the number of anchors increases. This is shown in Figure 9.9 where accuracies obtained from anchor features set III is plotted against the number of anchors used. The anchors to be added on as the number of anchors increases along the x-axis of the graph is randomly chosen from the 23 anchors considered. Notice that the accuracy for DevI has

increased from 80.5% with one anchor model (first anchor model is trained on airport signals) to 96.7% with 23 anchor models; while accuracy for DevII increases from 72.3% with one anchor model to 91.2% with 23 anchor models. Increment of accuracy for DevI seems to have slowed down as the number of anchors reach 14 while increment for DevII slows down at about 18 anchor models. Obviously this trend might change slightly if the choice of anchor that is added on along the x-axis is different, however, improvement is almost certain as the number of anchors increases. The fact that N-Best option fails to bring the desired improvement suggests that the improvement is gained from the act of anchoring, hence is due more to quality and quantity of anchors rather than by gleaning more features from each anchor. Next chapter assesses the usefulness of this feature set with the evaluation test set (i.e., Eval) and 6 human-evaluated test sets from Part I, namely DS1 to DS6.

Chapter 10

Evaluation of D^{4}IC

In Chapter 8 and 9 a number of different feature sets are examined in the context of the D^{4}IC. Here, those features deemed to be good are further examined. These are:

(i) The weighted distances between the slopes of the reference and degraded spectra in each critical band as computed in WSS [76]. Features are obtained using Approach II where feature differencing is performed prior to GMM adaptation. See Section 8.1.

(ii) Five ASR standard statistics from a clean-trained ASR, namely Word Accuracy, Percentage Correct, Deletion, Substitution and Insertion. See Section 8.2.1

(iii) Percentage of recognised frame (PRF) deduced from frame-based ASR where the transcription for clean and degraded signals are matched frame-by-frame. Statistics from silence period are excluded from the calculation and fixed 4-digit recognition is applied. See Section 8.2.2.

(iv) The ASR 10-Best hypothesis obtained by recording alternative routes through the network of HMM states during the word recognition process. Word Accuracy and a percentage Correct scores are deduced from each hypothesis. See Section 8.2.3.

(v) Nine quality measures assessed in Part I, namely CSNR, SNR, IS, LAR, LLR, WSS, MNB, MBSD and PESQ. See Section 8.3

(vi) ASR scores obtained from 23 anchor models; Word Accuracy and a percentage Correct scores are deduced from each anchor. See Chaper 9.

These are summarised in Table 10.1 with their abbreviations.

Section in thesis	Abbreviation	Feature Set
8.1	WSSspecII	Frame-based WSS using ApproachII
8.2.1	5ASRstd	5 ASR standard scores
8.2.2	ASRPRF	PRF with no silence & restricted word net
8.2.3	ASR10BEST	10-Best hypothesis using WordAcc & Corr
8.3	9QM	9 quality measures
9	ASRanchor	Anchoring ASR scores using WordAcc & Corr

Table 10.1: Feature sets used for the evaluation of the D^{4}IC.

The classifier is first evaluated with the evaluation test set, namely Eval (see Section 7.2 or Appendix A.2 for details on Eval) and then with the 6 human-evaluated test sets used in Part I namely DS1 to DS6. A comparison with the performance of measures considered in Part I is also made.

10.1 Performance of the D4IC with the Eval test set

Feature	DevI	DevII	Eval
WSSspecII	89.4	81.6	79.3
5ASRstd	90.1	79.3	78.4
ASRPRF	91.4	82.5	76.9
ASR10BEST	94.3	84.8	80.1
9QM	86.1	78.1	74.2
ASRanchor	96.7	91.2	88.4

Table 10.2: Classification accuracy of D4IC with Eval using chosen feature sets. Accuracies for DevI and DevII are shown for comparison.

The Eval set considers additive noises, coding schemes, NLSS configurations and other common speech processings. 16 unseen degradations are in the make-up of the Eval (i.e., the degradation pool for the Eval set has an extra 16 components compared to the pool for Train) on top of the 23 seen degradations. This compares with developmental test set I (DevI) which has no unseen degradations and development test set II (DevII) which has a set of 13 unseen degradations (partially different from the Eval's unseen set). Table 10.2 shows classification accuracy obtained for the Eval set using the various features sets mentioned in Table 10.1. The accuracy obtained for DevI and DevII are shown for comparison. As expected the accuracy for Eval are always significantly lower than those for DevI since there is no unseen degradation in DevI. Slight declines are observed when comparing accuracy obtained for DevII. The least decrement is shown by 5ASRstd where the accuracy for Eval at 78.4% is only 0.9% less than that for DevII at 79.3%. Meanwhile, the biggest decrement is shown by ASR10BEST with 4.7% difference. In overall the trend of performance among the feature sets remains, with ASRanchor giving the highest accuracy at 88.4% and 9QM giving the lowest at 74.2%. This gives confidence in the potential of the ASRanchor feature set, which consistently gives the highest accuracy for all 3 test sets.

10.2 Performance of the D4IC with Human-evaluated Test Sets

The D4IC is evaluated with the 6 human-evaluated test sets for direct comparison with the measures assessed in Part I. Note that the ground truth used here are not the synthetic ground truth generated based on the Intelligibility Enhancement (IE) Hypothesis, but human ground truth obtained from the

listening tests described in Chapter 3. The same ground truth are used to assess the quality measures and ASR in Chapter 4 and 5. Brief descriptions of the 6 test sets are shown in Table 10.3 while details can be found in Section 3.2 or Appendix A.1.

Type	Test Set	Descriptions
Additive	$DS1_{add}$	additive noises of diverse characteristics including both speech-like and more stationary ones.
	$DS2_{add}$	additive noises, most fairly stationary.
Coding	$DS3_{cod}$	car noise and tandeming of single coding scheme
	$DS4_{cod}$	various $DS1_{add}$ noises and tandeming of mixed coding schemes
Enhancement	$DS5_{enh}$	car noise and different speech enhancement processes
	$DS6_{enh}$	various $DS1_{add}$ noises and different configurations of the NLSS process

Table 10.3: Brief descriptions of the 6 human-evaluated test sets namely $DS1_{add}$ to $DS6_{enh}$.

Classification accuracy for the human-evaluated test sets are made directly comparable with $Kendall_2$ correlation computed in Part I. Computation of $Kendall_2$ correlation can be found in Section 4.2.1. Briefly, it is the average of Kendall correlations obtained for 3 sets of human and objective scores at 3 chosen SNRs. The chosen SNRs correspond approximately to 75%, 62.5% and 50% intelligibility as perceived by human for that particular test set. Test pairs are constructed by taking all possible pairing combinations at each chosen SNR. The total number of pairs for each test set is $3\times{}^{N}C_2$ where 3 corresponds to the 3 fixed SNRs, N is the number of degradations considered in a that particular test set, and C refers to *combination* (i.e., a form of permutation where order of elements does not matter). As an example, test set $DS1_{add}$ considers 8 types of environmental noises, hence N=8 and $3\times{}^{N}C_2 = 672$. The ground truth for each pair is deduced from human scores obtained in Chapter 3. Classification accuracy is defined as total *correct* over total pairs under test where a *correct* is scored when the classifier's output matches the human ground truth.

Table 10.4 shows classification accuracy obtained for test sets $DS1_{add}$ to $DS6_{enh}$. Shown in Table 10.5 are $Kendall_2$ correlations obtained for the same test sets using PESQ, WSS and Word Accuracy from a clean-trained ASR (WordAcc_clean). These 3 measures are chosen for comparison since WSS and WordAcc_clean are the best-performing intelligibility assessors from Part I, both with overall

Feature	$DS1_{add}$	$DS2_{add}$	$DS3_{cod}$	$DS4_{cod}$	$DS5_{enh}$	$DS6_{enh}$	Average
WSSspecII	82.7	69.8	86.5	70.3	66.9	67.5	74.0
55ASRstd	89.1	82.3	88.7	69.6	59.7	57.2	74.4
ASRPRF	86.1	80.5	88.4	70.8	61.3	62.6	75.0
ASR10BEST	89.9	81.6	89.3	70.6	62.0	61.4	75.8
9QM	61.3	89.6	86.8	66.4	58.3	60.7	70.5
ASRanchor	90.3	86.6	89.5	72.3	75.8	79.7	82.4

Table 10.4: Classification accuracy of D^4IC with the human-evaluated test sets using chosen feature sets.

Part I Measure	$DS1_{add}$	$DS2_{add}$	$DS3_{cod}$	$DS4_{cod}$	$DS5_{enh}$	$DS6_{enh}$	Average
PESQ	16.2	90.7	82.8	53.6	54.2	43.9	56.9
WSS	76.4	55.5	88.1	55.2	73.7	64.2	68.9
WordAcc_clean	85.1	66.7	87.2	62.3	57.8	46.4	67.6

Table 10.5: $Kendall_2$ correlation obtained by PESQ, WSS and WordAcc_clean for 6 human-evaluated test sets.

$Kendall_1$ correlation at 0.72; meanwhile PESQ is chosen for being the state-of-the-art quality assessor hence serves as an indication of the general applicability of quality measures in the human-evaluated test sets namely context of intelligibility assessment. Kendall correlation is directly comparable with the classification accuracy since both are defined as the number of correct match over total pairs under test. The $Kendall_2$ correlations shown in Table 10.5 have been converted to the percentage range of 0% and 100% for direct comparison with the classification accuracy in Table 10.4.

Table 10.4 shows that $DS1_{add}$ and $DS3_{cod}$ generally obtain higher accuracy than the rest of the test sets. This scenario is consistent for all feature sets apart from the 9 quality measures (9QM), which gives low accuracy for $DS1_{add}$ at 61.3%. This seems to agree with the findings in Part I where almost all quality measures especially the modern perceptual-based ones, fail to correlate when noises of diverse characteristics such as speech-like noises are considered, as shown by the poor score of PESQ at 16.2%.

Notice that high accuracy is obtained for $DS1_{add}$ whenever ASR-based features are used. This again, perhaps is due to the fact that ASR scores, particularly Word Accuracy, Substitution and Insertion have shown to be potentially good at identifying impairment caused by the speech-like noises, as shown by accuracy of WordAcc_clean at 85.1% in Table 10.5. The use of the D^4IC has shown to be beneficial with this particular test set, as shown by the improvement achieved by ASRanchor at 90.3% over 85.1% obtained with the simple WordAcc_clean in Part I. Also worth noting is the performance of WSSspecII which gives 82.7% accuracy for $DS1_{add}$ compared 76.4% by WSS which integrates distortion scores across the critical bands in each frame as well as along the time course. This coincides with the notion that adding more relevant data to the classifier potentially leads to better performance.

In Part I, the correlations obtained for $DS2_{add}$ are very different to those for $DS1_{add}$. For example, WSS correlates well for $DS1_{add}$ but poorly for $DS2_{add}$; PESQ correlates well for $DS2_{add}$ but poorly for $DS1_{add}$. This phenomenon has been discussed in Part I and is thought to be due to the differences between the nature of degradations considered in the two test sets. With the exception of WSSspecII and 9QM, here such distinctive difference in the accuracy for the two test sets are not always observed. All the ASR-based feature sets namely 5ASRstd, ASRPRF, ASR10BEST and ASRanchor give comparable accuracy for both test sets, for example, ASRanchor gives 91.3% for $DS1_{add}$ and 88.6% for $DS2_{add}$, the performance difference of which is marginal.

Though both $DS3_{cod}$ and $DS4_{cod}$ consider degradation introduced by coding schemes, notice that

classification accuracy obtained for $DS4_{cod}$ are always significantly lower than those of $DS3_{cod}$. Apart from the possibility that degradations in $DS4_{cod}$ are more challenging, it is possible that the ground truth for this test set might be less reliable due to poor combinations of degradation settings. This is observed from the bar plot shown in Section 3.4 where the human scores for the different degradations in $DS4_{cod}$ are only marginally different. As shown the human scores for this test set only range from approximately 48% to 65%, whereas the human scores for other test sets occupy a much wider range, for example, 63% to >90% for $DS2_{add}$. This warrants further investigation. Nonetheless, the highest accuracy achieved by ASRanchor at 72.3% is still better than any achieved in Part I.

Lastly, relatively poor accuracy are obtained for the enhancement test sets, namely $DS5_{enh}$ and $DS6_{enh}$. The lowest accuracy is obtained when 9QM is employed, at 58.3% and 60.7% respectively. This perhaps is not surprising since it is known that degradations/processes considered here improve machine scores without improving intelligibility. Similar low accuracy are obtained when the 5ASRstd feature set is employed, due to presumably the same reason. Apart from ASRanchor, other ASR-based feature sets show negligible improvement. The D^{4}IC with ASRanchor also gives the highest accuracy for $DS5_{enh}$ and $DS6_{enh}$ at 75.8% and 79.7% respectively, which compares favourably with the accuracy obtained in Part I by WordAcc_clean at 57.8% and 46.4% respectively, again demonstrating the potential of the D^{4}IC.

Notice that both WSS and WordAcc_clean give higher accuracy for $DS5_{enh}$ than for $DS6_{enh}$, however, the reverse is true for ASRanchor which gives higher accuracy for $DS6_{enh}$ where the NLSS configurations are considered. The fact that the NLSS process is in the pool for the generation of the D^{4}IC's training set (i.e., data set Train, see Section 7.1.2) could possibly contribute to this higher accuracy of $DS6_{enh}$ compared to $DS5_{enh}$.

In overall, the D^{4}IC with ASRanchor feature set proves to be a potentially good intelligibility assessor with average accuracy at 82.4%. Although the accuracy obtained for an individual test set may not be higher than the best obtained in Part I as shown in Table 10.5, however, the average accuracy across the 6 test sets is significantly higher at 82.4% compared with 68.9% and 67.6% achieved by WSS and WordAcc_clean respectively. Furthermore, it is envisaged that the D^{4}IC accuracy can be improved by targetting its training and anchoring feature set towards the known information of the signals under test. This is suggested by the improved accuracy obtained for $DS6_{enh}$, which is thought to be due to the inclusion of NLSS process in the degradation pool during generation of the D^{4}IC's training set (i.e., data set Train, see Section 7.1.2), and the fact that NLSS process also forms some of the anchors in the cohort of anchoring ASRs used in features extraction.

Part III

Conclusions and Future Work

Chapter 11

Conclusions, Final Thoughts and Future Work

Speech intelligibility is becoming an increasingly important issue. Three factors could be contributing to this growth: (i) the basic and growing requirement of domestic telephony including the 'anytime, anywhere' expectation [4] of mobile phone users; (ii) business transactions and 'black box' flight conditions, the recordings of which are pointless if unintelligible; (iii) interceptions for security and anti-terrorism. In the last two cases intelligibility is of upmost importance, almost regardless of the presence of other quality attributes such as naturalness or ease-of-listening. The awareness of its importance is further illustrated by the Interspeech Conference in 2009 having the theme of "Speech and Intelligence", and a study group (study period: 2005-2008) formed by ITU-T to extend PESQ, the state-of-the-art quality measure, for intelligibility assessment [125].

The cornerstone of the research presented in this thesis is the objective assessment of intelligibility. The primary objective is to deduce comparative intelligibility between two or more signals, the question being 'which is the more intelligible?'. Three difficulties associated with the task are identified. The first difficulty is caused by the inevitability of operating under high degradation conditions, resulting in a constrained dynamic score range from most if not all objective measures. This constrained range makes measurement or comparison of intelligibility more difficult. The second difficulty relates to the paradox that speech processings such as enhancement algorithms can relatively easily improve machine-based scores, such as ASR word accuracy, yet very rarely improve intelligibility. Such processings could confound an intelligibility measure. Lastly, the third difficulty is to obtain a substantial amount of reliable ground truth which is needed for the development and evaluation of objective intelligibility measures. The thesis begins by recognising these problems and aims to tackle them in the research.

The investigatory chapters of this thesis is divided into 2 parts where Part I investigates the usefulness of existing measures including quality measures and ASR; while Part II proposes a direct data-driven approach. This chapter summarises the work presented highlighting key findings, followed by some suggestions for future works.

11.1 Part I: Existing Measures

Given the wealth of objective measures available for quality assessment and the widely-accepted concept of intelligibility being an attribute of overall quality [20, 27], Chapter 4 investigates the potential of

quality measures in the context of intelligibility assessment. Nine prominent quality measures are considered including 2 time-domain measures: classical signal-to-noise ratio (CSNR) and segmental SNR (SegSNR), 4 spectral-domain measures: Itakura-Saito (IS), log area ratio (LAR), log likelihood ratio (LLR), and weighted spectral slope (WSS) [76]) and 3 perceptual-domain measures: measuring normalising blocks (MNB) [70, 81], modified Bark spectral distortion (MBSD) [78], and perceptual evaluation of speech quality (PESQ) [8]). Six test sets are prepared where objective scores given by the measures are compared with human scores on identical test sets. The test sets consider 3 main categories of degradations namely environmental noises, speech coding schemes and enhancement processes.

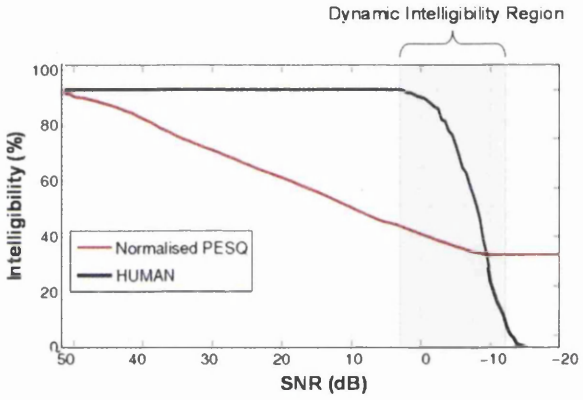


Figure 11.1: Reproduced from Figure 4.6 except that PESQ profile is the average of PESQ scores for 566 utterances. Profiles are humans and PESQ scores for car noise degraded signals. PESQ scores mapped from the original scale of -0.5:4.5 to 0:100%. Humans deem signals as 100% intelligible until intelligibility threshold of about 3dB, while PESQ score decreases steadily with decreasing SNR; shaded region is the critical region of dynamic intelligibility where assessment is often needed, PESQ profile is close to saturation in this region indicating lack of sensitivity.

It is shown that scores given by the quality measures are in a constrained range due to the different operational ranges for quality and intelligibility. Figure 11.1 (and Section 4.3 in Chapter 4) shows the experimental results for PESQ and human scores for car noise degraded signals; as illustrated the PESQ profile is close to its saturation point at the region of dynamic intelligibility where assessment is an issue. This trend of levelling (seemingly-flat response) could indicate lack of sensitivity of the measure in this critical region. Indeed, experimental results show that all quality measures, with the notable exception of perhaps WSS, correlate poorly with human intelligibility. Details can be found in Section 4.3.

One pleasant surprise is WSS which gives an average Kendall_1 correlation of 0.72 and is the only quality measure capable of identifying speech-like noises as more impairing than the stationary noises when test set DS1_{add} is considered. Among the 9 measures assessed, WSS is also the only one to give reasonably high correlations at 0.75 and 0.68 respectively for the 2 enhancement test sets, whereas PESQ gives only 0.54 and 0.40 respectively. This finding supports the observation made by Manohar and Rao in [93] that WSS outperforms PESQ in terms of intelligibility correlation when considering speech enhancement in factory noise. This places WSS, a relatively traditional measure, in a new light as a potential estimator of intelligibility. These findings are reported in Section 4.3 in Chapter 4.

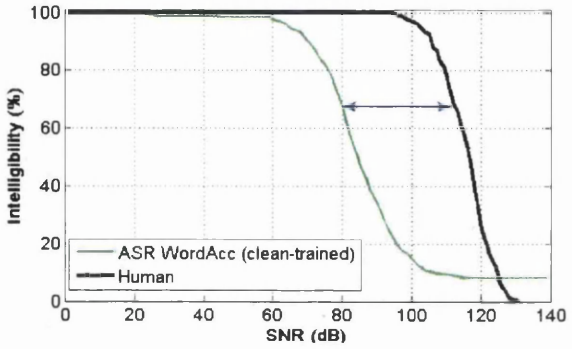


Figure 11.2: Reproduced from Figure 5.2. Profiles are ASR word accuracy and human scores for the same set car noise degraded signals for profiles in Figure 11.1. Figure illustrates that ASR word accuracy exhibit an s-curve trend similar to that of human intelligibility response, though with an offset (indicated by arrow).

Due to the close link between word recognition and intelligibility [21, 23, 24, 126], Chapter 5 investigates the potential of automatic speech recognition (ASR). It is interesting to observe that the ASR word accuracy exhibits an s-curve trend similar to that of humans as illustrated in Figure 11.2. Notice that such a trend constrasts the trend of PESQ profile in Figure 11.1. In this thesis the Aurora2 standard HMM-based digit-string recogniser is used to give confidence for the ASR setup and justification for the evaluation of its potential. Five word recognition statistics are considered, namely Word Accuracy, Percentage Correct, Deletion, Substitution and Insertion, where each is treated as a potential intelligibility measure.

Three ASR variants are investigated, namely clean training, multi-condition (mixed) training and application of missing data techniques. One major contribution here is the observation that each recognition statistic is linked to human speech recognition in specific ways and their respective correlation performance depends on the nature/characteristic of the degradations considered. For example, Substitution and Insertion could identify with the presence of speech-like noises in $DS1_{add}$, while Deletion identifies with the stationary noises in $DS2_{add}$; speech coding ($DS3_{cod}$ and $DS4_{cod}$) causes spectral rather than masking distortion hence Substitution rather than Deletion proves to be more informative; similarly, enhancement processes may impose excessive noise attenuation which changes the pronunciation of words and introduces clicking noise, hence again Substitution and Insertion rather than Deletion are more relevant. Such links enable different statistics to be chosen for different applications. This presents an advantage over the quality measures where their good and bad aspects of performances are difficult to interpret.

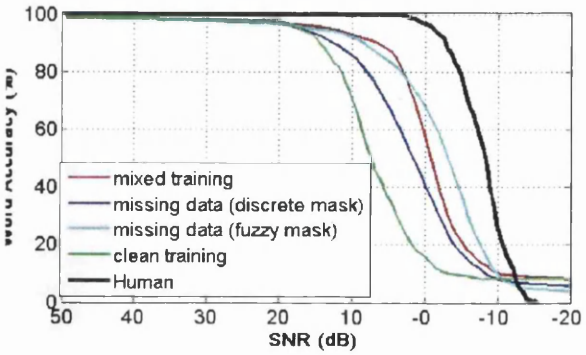


Figure 11.3: Reproduced from Figure 5.10. Same as Figure 11.2 but with 3 extra profiles produced by missing data ASR systems and mixed-trained ASR system. All 3 new profiles show obvious improvement of recognition rate over the clean-trained profile and can be seen as approaching human performance.

Among all recognition statistics considered Word Accuracy from the clean-trained system proves to be the best-performing with average Kendall$_1$ correlation of 0.72 across the 6 test sets; and 0.81 if the enhancement test sets are not included. One observation is that majority of the recognition statistics correlate poorly in the context of degradation caused by enhancement processes. This is presumably because ASR is commonly used as a yardstick for development and optimisation of enhancement systems, hence those systems are designed to improve ASR scores without necessarily improving intelligibility.

Another observation is that whilst mixed training and missing data techniques greatly improve the recognition rate, with ASR performance approaching that of humans in the given context of digit recognition as illustrated in Figure 11.3, disappointingly the intelligibility correlation has not benefited from this (absolute) improvement in ASR performance. Poor correlation of statistics from the mixed-trained system is perhaps expected since recognition improvement is biased towards degradation conditions seen during training. Improved correlation, however, is expected from statistics of the missing data system since the technique claims to mimic humans' ability to perform auditory scene analysis hence should better reflect humans' judgement. However, rather unexpectedly the correlations obtained with the missing data ASR system are also poor. The average Kendall$_1$ correlation obtained with the Word Accuracy of the missing data system (fuzzy mask) is only 0.52 compared to 0.72 with the primitive clean-trained ASR. Obviously training an ASR system to recognise words does not equate to training it to measure intelligibility, and improved absolute scores do not necessaily lead to improved correlation.

As a final thought for Part I, ASR technology is bound to advance and, given time, it is totally possible that the gap between ASR and human speech recognition (HSR) will close, both in terms of recognition rate and intelligibility judgement that reflects humans'. Afterall, ASR was once deemed impossible in 1969 by J.R.Pierce [127] from Bell Labs who described speech recognition as similar to 'going to the moon'. Now, not only has man visited the moon (in the same year as the quotation, i.e., 1969!), ASR technology has matured and is widely applied in our daily life. However, current ASR technology is still a distance away from human performance [?, 17, 18], both in terms of recognition rate and intelligibility assessment. The latter situation is made worse by the presence of degradations introduced by enhancement processes. In other word, while human word recognition is intelligibility, ASR word recognition is not, at least not yet at the present moment. As a mean of going beyond the limitation of ASR, or as an extension of word modelling, Part II of this research proposes to model intelligibility directly.

11.2 Part II: Direct, Data-driven, Differential Intelligibility Classifier (D^{4}IC)

Part II seeks to model intelligibility or intelligibility-related information so that intelligibility can be estimated directly rather than indirectly through quality or ASR scores. The main constraint for such

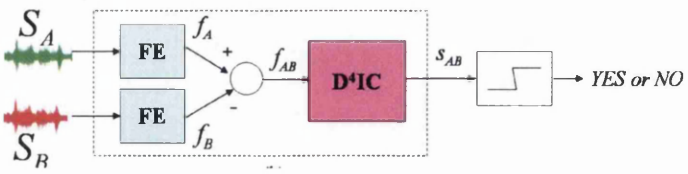


Figure 11.4: An illustration of the D^{4}IC arrangement: S_A and S_B are two signals under comparison; FE is the feature extraction unit; f_A and f_B are feature vectors of the two comparing signals where f_B is subtracted from f_A resulting in a vector of feature difference, f_{AB}. The classifier is trained on feature difference and the output score s_{AB} relates directly to signal difference hence intelligibility difference.

statistical modelling is the source of training data, namely substantial amounts of signals with information regarding intelligibility. This constraint perhaps explains why, while the data-driven classification approach proves successful in many other speech processing tasks such as speaker verification and ASR or even very recently, quality assessment [25, 85–89, 128], to date it has not been attempted for intelligibility assessment. A major contribution claimed here is overcoming this constraint of training the intelligibility classifier. A strategy is proposed to generate large amounts (potentially unlimited) of signal pairs where intelligibility of one is assumed to be greater than the other. This is made possible by what we call the Intelligibility Enhancement (IE) Hypothesis which states that under normal, daily encountered circumstances such as background noise, speech coding and majority if not all of the enhancement processes, intelligibility of degraded speech is most unlikely to be improved [1–3, 36]. The supporting evidences for the hypothesis is presented in Section 6.2. This strategy together with the pairing scheme introduced in Section 7.1 provides convenience and flexibility to simulate an operating environment by choosing desired degradations for the degradation pool, from which training data for the classifier are generated.

As mentioned the data generated has known relative intelligibility (which of a pair is the more intelligible?) though the absolute values of each signal and absolute difference between the two are unknown. This leads to implementation of a differential/relative classifier hence the name data-driven, direct, differential intelligibility classifier (D^{4}IC). The D^{4}IC takes as input two signals of the same origin but processed diffferently and outputs a score indicating their relative intelligibility. The D^{4}IC is believed to be superior over previous measures for three reasons: (i) being data-driven, direct intelligibility information is imparted into the classifier; (ii) output score is directly related to differential intelligibility, rather that indirectly through quality or ASR word recognition; (iii) signal differencing is performed at the early stage of feature level where features into the classifier is f_A-f_B, the classifier is thus sensitive to signal difference and subsequently intelligibility difference. Figure 11.4 shows the structure of the D^{4}IC where S_A and S_B are two signals under comparison; and FE is the feature extraction unit that generates f_A and f_B as feature vectors for the two comparing signals. Notice that the feature extraction (FE) process is followed immediately by differencing of f_A and f_B so that the D^{4}IC receives signals that relates directly to intelligibility differences.

For the development and evaluation of the classifier, 4 data sets are generated namely 1 training set, 2 development test sets and 1 evaluation test set, referred to as Train, DevI, DevII and Eval respectively. The make-ups and purpose of each data set is described in Section 7.1.2 as well as Appendix A.2. Briefly, Train and DevI generated from the same degradation pool hence there is no unseen degradation and DevI aims to give confidence to the classifier setup; meanwhile DevII has 13 unseen (and 23 seen degradations) and is a more challenging test set aiming to identify robust features for the classifier; lastly Eval has 16 unseen degradations aiming to evaluate the classifier using potential feature sets identified during development. The baseline accuracy for DevI and DevII are 90.5% and 79.9% are respectively.

Another contribution of this thesis is the search for good features for the classifier. Features considered are scores coming from the quality measures and ASR. In terms of categories, the features can be deduced from frame, word or utterance level of the signal. Firstly, 2 sets of low-level features namely the 39-component cepstra-based feature used in the Aurora2 recogniser and the frame-based weighted spectral slopes computed using WSS are considered. Experimental results show that the cepstra-based features which are commonly used in other speech-related classification tasks (ASR, speaker verification) do not seem to be particularly promising for the task here with accuracy lower than the baseline. The frame-based WSS, however, shows slight improvement over the baseline giving 81.6% for DevII.

On the other hand, bigger improvements of classification accuracy are obtained with 2 feature sets derived from the word recognition process in ASR, namely percentage of recognised frames (PRF) and N-Best hypothesis (N-BEST). PRF is derived from frame level recognition where the percentage of frames correctly identified with the target word model is used as an intelligibility indicator. This feature set intends to measure the number of recognisable segments, with the hypothesis that the higher the percentage, the more intelligible is the speech. The second feature set, N-BEST, is computed by recording all alternative routes of the token within the network of HMM states. These alternative routes might reflect the number of attempts a human listener would take to comprehend a signal, with the hypothesis that a degraded signal would need more attempts (recognised at latter rather than earlier attempt) than a cleaner one. PRF and N-BEST give 82.5% and 84.8% accuracy respectively for DevII, equivalent to 2.6% and 4.9% improvement over the classifier baseline result. This work is reported in Section 8.2. These improvements, though small, re-affirm the link between ASR and intelligibility especially when compared to 78.1% accuracy obtained with the fusion of 9 quality measures reported in Section 8.3. On a side note, this may imply that as ASR technology advances, conceptually so does the classifier which employs these features.

Chapter 9 introduces a novel feature based on the anchor models concept used in speaker verification and indexing [122, 123, 129]. Here the anchors come from chosen degradations rather than chosen speakers, giving rise to the concept of a degradation reference space rather than speaker reference space. Given that the intelligibility of a signal is a direct function of the degrading capability of the

	D_{add1}	D_{add2}	D_{cod1}	D_{cod2}	D_{enh1}	D_{enh2}	Average
PESQ	16.2	90.7	82.8	53.6	54.2	43.9	56.9
WSS	76.4	55.5	88.1	55.2	73.7	64.2	68.9
WordAcc_clean	85.1	66.7	87.2	62.3	57.8	46.4	67.6
D^4I with Anchor model feature set	90.3	86.6	89.5	72.3	75.8	79.7	82.4

Table 11.1: Comparison of Accuracy obtained by D^4IC with respect to WSS, PESQ and WordAccc_clean for the 6 human-evaluated test sets.

degradation it underwent, here signal intelligibility is characterised by the similarity between that degradation and a cohort of chosen anchoring degradations. ASR systems are used as the anchor models where each ASR is trained on signals degraded by a chosen degradation. Output scores of each anchoring ASR (i.e., Word Accuracy, Percentage Correct, Deletion, etc) are simply stacked up to form a feature vector. This feature set is believed to be more discriminative of intelligibility since it contains information about representative degradations. Classification accuracy is shown in rise as the number of anchors employed increases as shown in Figure 9.9 in Section 9.4. Another major benefit of this feature set is the potential to identify those degradations that have shown themselves to be difficult (confusing), for example, those introduced by the enhancement processes. Classification accuracy achieved by this feature set at 96.7% and 91.2% respectively for DevI and DevII are the highest among all feature sets considered. Furthermore, this feature set provides the flexibility to build an application-specific classifier by selecting only relevant degradations to be the anchors.

Finally, Chapter 10 evaluates the D^4IC using the evaluation test sets (i.e., Eval) and the human-evaluated test sets used in Part I. The anchoring feature set proposed in Chapter 9 is again shown to perform best among other feature sets with 88.4% accuracy when tested with the Eval set. Table 11.1 compares the D^4IC performance with selected exisiting measures which are assessed in Part I. Note that the accuracies reported by these measures are converted from $Kendall_2$ correlations (Accuracy = $Kendall_2 \times 100\%$) . Three measures are selected for comparison namely WSS, Word Accuracy from a clean-trained ASR (WordAcc_clean) and PESQ. The first two are chosen since the measures prove to be the best performing as suggested by findings in Part I. Meanwhile PESQ is also included since it is the state-of-the-art quality measure. As shown the D^4IC gives highest overall accuracy at 82.4% which compares favourably with 68.9% by WSS and 67.6% by WordAcc_clean; whereas PESQ gives the lowest overall accuracy at 56.9%.

This research claims to be the pioneer in applying data-driven classification approach to intelligibility assessment. Obviously this can only realisable upon identifying the strategy to generate large amounts of training data needed for such statistical modelling. The D^4IC introduced is shown to correlate better than existing measures considered here especially the quality measures. One major improvement over other measures is its potential to overcome inconsistencies introduced by processes that improve machine-based scores artificially without necessarily improving intelligibility. Though it can be argued that the classifier at its current accuracy is a distance away 'perfect' performance such as

the 0.98 correlation reported by PESQ the state-of-the-art for quality testing; however, this pioneering technology offers enormous room as well as flexibility for further research and optimisation.

11.3 Future Work

This research has implemented a general-purpose D^4IC where the degradations considered encompass environmental noises, coding schemes and enhancement processes. Clearly the narrower and the more specific the operational range is, the more accurate the classifier will become. Further lines of investigation could address application specific scenarios where the degradations likely to be encountered are known, in which case, the training data and features employed can be targetted for the specific application. Representative training data can be achieved by choosing only desired or relevant degradations for the pool during the data generation process. Meanwhile the classifier features are likely to be more robust if relevant degradations are chosen as anchors.

Bibliography

[1] S. F. Boll. Speech enhancement in the 1980s: noise suppression with pattern matching. In S. Furui and M. M. Sondhi, editors, *Advances in speech signal processing.* Marcel-Dekker, 1991.

[2] Y. Hu and P. C. Loizou. A comparative intelligibility study of speech enhancement algorithms. *ICASSP*, 4(4):561–564, 2007.

[3] Y. Hu and P. C. Loizou. A comparative intelligibility study of speech enhancement algorithms. In *Proc. ICASSP*, volume 4, pages 561–564, 2007.

[4] R. Martin. Speech enhancement based on minimum mean-square error estimation and super-gaussian priors. *IEEE Trans. on SAP*, 13:845–856, 2005.

[5] S. R. Quackenbush, T. P. Barnwell, III, and M. A. Clement, editors. *Objective Measures of Speech Quality.* Prentice Hall, Eaglewood Cliffs, 1988.

[6] W. Yang, M. Benbouchta, and R. Yantorno. A modified bark spectral distortion measure as an objective speech quality measure. *IEEE ICASSP*, pages 541–544, 1998.

[7] S. Voran. Estimation of perceived speech quality using measuring normalizing blocks. In *IEEE Speech Coding Workshop*, pages 83–84, 1997.

[8] ITU-T P.862. Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. *Int. Telecommunication Union, Geneva, Switzerland*, 2001.

[9] ITU-T P.563. Single ended method for objective speech quality assessment in narrowband telephony applications. *Int. Telecommunication Union, Geneva, Switzerland*, 2004.

[10] N. R. French and J. C. Steinberg. Factors governing the intelligibility of speech sounds. *J. Acoust. Soc. Am*, 19:90–119, 1947.

[11] H. J. M. Steeneken and T. Houtgast. The modulation transfer function in room acoustics as a predictor of speech intelligibility. *Acustica 28*, pages 66–73, 1973.

[12] B. J. van Gils and S. J. van Wijngaarden. Objective measurement of the speech transmission quality of vocoders by means of the speech transmission index. *RTO-MP-HFM-123, NATO Research and Technology Organisation*, page paper12, 2005.

[13] S. J. van Wijngaarden and J. A. Verhave. Recent advances in STI measuring techniques. In *Proc. of the Institute of Acoustic*, volume 28, 2006.

[14] S. E. Mercy and M. J. Aitchison. Subjective speech intelligibility measurements and 3d audio implementation. In *Proc. of the Institute of Acoustic*, volume 28, 2006.

[15] K. Worrall, R. Fellows, J. Causer, and L. Craigie. Intelligibility testing at HMGCC. In *Proceedings of the Institute of Acoustics*, pages 12–22, 2006.

[16] T. South. Speech intelligibility and respirator use. In *Proc. of the Institute of Acoustic*, volume 28, 2006.

[17] R. Lippmann. Speech recognition by machines and human. *Speech Communications*, 22(1):1–15, 1997.

[18] O. Scharenborg. Reaching over the gap: A review of efforts to link human and automatic speech recognition research. *Speech Communications-Special Issue on Bridging the Gap Between Human and Automatic Speech Processing*, 49:336–347, 2007.

[19] P. Mapp. Error mechanisms in speech intelligibility measurements. In *Proc. of the Institute of Acoustic*, volume 28, 2006.

[20] F. L. Chong, I. McLoughlin, and K. Pawlikowski. A methodology for improving PESQ accuracy for chinese speech. *IEEE TENCONS*, pages 1–6, 2005.

[21] H. Fletcher. *Speech and Hearing.* D. Van Nostrand Company, 1929.

[22] C. M. Chernick, S. Leigh, K. L. Mills, and R. Toense. Testing the ability of speech recognizers to measure the effectiveness of encoding algorithms for digital speech transmission. *IEEE Int. Military Comm. Conf (MILCOM)*, 1999.

[23] W. Jiang and H. Schulzrinne. Speech recognition performance as an effective perceived quality predictor. In *IEEE Int. Workshop on Quality of Service*, pages 269–275, 2002.

[24] W. T. Hicks, B. Y. Smolenski, and R. E. Yantorno. Testing the intelligibility of corrupted speech with an automated speech recognition system, 2003.

[25] T. H. Falks, W. Y. Chan, and P. Kabal. An improved GMM-based voice quality predictor. *9th European Conf. on Speech Comm. and Tech.*, 2005.

[26] H. G. Hirsch and D. Pearce. The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. *ISCA ITRW ASR2000 'Automatic Speech Recognition: Challenges for the next Millenium'*, 2000.

[27] J. E. Preminger and D. J. Van Tasell. Quantifying the relation between speech quality and speech intelligibility. *Journal of Speech and Hearing Research*, 38:714–725, 1995.

[28] J. R. Deller, J. G. Proakis, and J. H. L. Hansen. *Discrete-time processing of speech signals.* Macmillan, 1993.

[29] W. J. Ebel and J. Picone. Human speech recognition performance on the 1994 CSR spoke 10 corpus. In *Proc. Spoken Language Systems Technology WOrkshop*, pages 53–59, 1995.

[30] A. Varga and H. J. M. Steeneken. Assessment for automatic speech recognition: II NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communications*, 12(3):247–251, 1993.

[31] Y. Hu and P. C. Loizou. A perceptually motivated approach for speech enhancement. *IEEE Trans. on Speech and Audio Processing*, 11(5):457–460, 2003.

[32] N. W. D. Evans, J. S. Mason, and M. J. Roach. Noise compensation using spectrogram morphological filtering. In *Proc. 4th IASTED International Conference on Signal and Image Processing*, pages 157–161, 2002.

[33] N. W. D. Evans and J. S. Mason. Computationally efficient noise compensation for robust automatic speech recognition assessed under the AURORA 2/3 framework. In *Proc. ICSLP*, volume 1, pages 485–488, 2002.

[34] F. Romero Rodriguez, W. M. Liu, N. W. D. Evans, and J. S. D. Mason. Morphological filtering of speech spectrograms in the context of additive noise. In *Proc. Eurospeech*, 2003.

[35] N. W. D. Evans. Spectral subtraction for speech enhancement and automatic speech recognition. *PhD Thesis, University of Wales Swansea*, 2003.

[36] J. S. Lim. Speech enhancement. In *Proc. ICASSP*, pages 3135–3142, 1986.

[37] R. H. Frasier, S. Samsam, L. D. Braida, and A. V. Oppenheim. Enhancement of speech by adaptive filtering. *ICASSP*, pages 251–253, 1976.

[38] MESAQIN Laboratories. Measurement of speech and audio transmission quality in telecommunication networks (MESAQIN). *http://www.mesaqin.com*, 31 Dec 2007.

[39] ITU recommendation P.800. *Methods for Subjective Determination of Transmission quality.* ITU, 1996.

[40] Y. Teng and R F. Kubichek. Speech intelligibility evaluation of low bit rate speech codecs. In *Digital Sig. Proc. Workshop, 12th - Sig. Proc. Education Workshop, 4th*, pages 251–256, 2006.

[41] A. S. House, C. E. Williams, M. H. L. hecker, and K. D. Kryter. Articulation testing methods: Consonantal differentiation with a closed-response set. *J. Acous. Soc. Am*, 37(1):158–166, 1965.

[42] G. Fairbanks. Test of phonemic differention: The rhyme test. *J. Acous. Soc. Am*, 30(7):596–600, 1958.

[43] S. L. Greenspan, R. W. Bennett, and A. K. Syrdal. An evaluation of the diagnostic rhyme test. *International Journal of Speech Technology*, 2:201–214, 1998.

[44] J. P. H. Van Santen. Perceptual experiments for diagnostic testing of text-to-speech systems. *Computer Speech and Language*, 7:49–100, 1993.

[45] J. M. Steeneken. The measurement of speech intelligibility. In *Proceedings IoA 2001*, volume 23(8), 2001.

[46] ANSI S3.2-1989. *American National Standard for Measuring the Intelligibility of Speech over Communication Systems.* American National Standards Institute (ANSI), 1989.

[47] J. Logan, B. Greene, and D. Pisoni. Segmental intelligibility of synthetic speech produced by rule. *Journal of the Acoustical Society of America, JASA*, 86 (2):566–581., 1989.

[48] C. De logu, A. Paolini, P. Ridolfi, and K. Vagges. Intelligibility of speech produced by texto-to-speech systems in good and telephonic condtions. *Acta Acoustica 3*, pages 89–96, 1995.

[49] A. W. Rix. Comparison between subjective listening quality and P.862 PESQ score. *White Paper*, pages –, 2003.

[50] M. E. Hawley. *Speech intelligibility and speaker recognition (ed.). Stroudsburg, PA: Dowden,.* Hutchinson & Ross, Inc., 1977.

[51] ITU recommendation P.56. *Objective measurement of active speech level.* ITU, 1993.

[52] M Mouly and M-B Pautet. *GSM system for mobile communications.* Palaiseau, 1992.

[53] ITU Recommendation G.728. *Coding of Speech at 16 kbit/s using Low-delay Code Excited Linear Prediction.* ITU, 1992.

[54] L. N. Supplee, R. P. Cohn, J. S. Collura, and A. V. McCree. MELP: The new federal standard at 2400 bps. In *Proc. ICASSP*, volume 2, pages 1591–1594, 1997.

[55] CCITT. *Red Book, Recommendation G721.* Tome III-3, October 1984.

[56] S. F. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. on ASSP*, 27(2):113–120, 1979.

[57] M. Berouti, R. Schwartz, and J. Makhoul. Enhancement of speech corrupted by acoustic noise. In *Proc. ICASSP*, pages 208–211, 1979.

[58] S. Kamath and P. Loizou. A multi-band spectral subtraction method for enhancing speech corrupted by colored noise. In *Proc. ICASSP*, 2002.

[59] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE Trans. on ASSP*, pages 1109–1122, 1984.

[60] P. Scalart and J. Filho. Speech enhancement based on a priori signal to noise estimation, in proc. *IEEE Int. Conf. Acoust. Speech Signal Processing, Atlanta, GA*, pages 629–632, 1996.

[61] M. Kepesi and T.V. Pham. Noise cancellation frontends for automatic meeting transcription. *Euronoise'06, Tampere, Finland*, 2006.

[62] T.V. Pham and G. Kubin. WPD-based noise suppression using nonlinearly weighted threshold quantile estimation and optimal wavelet shrinking. *Interspeech'05, Lisboa, Portugal*, 2005.

[63] E. Rank and G. Kubin. Lattice LP filtering for noise reduction in speech signals. In *international Conference on Spoken Language Processing (Interspeech - ICSLP)*, 2006.

[64] T.V. Pham and E. Rank. New algorithms for noise reduction and speech augmentation. *http://www.snow-project.org/*, 2007.

[65] W. M. Liu, J. S. Mason, N. W. D. Evans, and K. A. Jellyman. An assessment of automatic speech recognition as speech intelligibility estimation in the context of additive noise. In *Proc. ICSLP*, 2006.

[66] W. D. Voier. Interdependencies among measures of speech intelligility and speech quality. In *Proc. ICASSP*, pages 703–705, 1980.

[67] J. C. R. Licklider. Effects of amplitude distortion upon the intelligibility of speech. *Journal of the Acoustical Society of America*, 18(2):429–434, 1946.

[68] J. G. Beerends, S. V. Wijngaarden, and R. V. Buuren. Extension of ITU-T recommendation P.862 PESQ towards measuring speech intelligibility with vocoders, 2005.

[69] R. Kaga and K. Kondo. Estimation of japanese speech intelligibility using objective speech quality evaluation method PESQ. *Processdings of the 5th Technical Meeting of the Information Processing Society of Japan Tohoku Chapter*, A4-1, 2006.

[70] S. Voran. Objective estimation of perceived speech quality, part I: Development of the measuring normalizing block technique. *IEEE Transactions on Speech and Audio Processing*, 1999.

[71] N. Kitawaki, M. Honda, and K. Itoh. Speech quality assessment methods for speech coding systems. *IEEE Commun. Magazine*, 22(10):26–33, 1984.

[72] N. Kitawaki and H. Nagabuchi. Objective quality evaluation for low-bit-rate speech coding systems,. *IEEE J. on Sel. Asreas in Comm*, pages 242–248, 1988.

[73] R.E. Crochiere and J.E. Tribolet. An interpretation of the log likelihood ratio as a measure of waveform coder performance. *IEEE Trans. Acoust, Speech and Signal Processing*, ASSP-28, No.3, 1980.

[74] B. Juang. On the hidden markov model and dynamic time warping for speech recognition - a unified view. *ATT Bell Laboratories Technical Journal*, pages 1213–1243, 1984.

[75] D. H. Klatt. A digital filter bank for spectral matching. In *Proc. ICASSP*, pages 573–576, 1976.

[76] D. H. Klatt. Prediction of perceived phonetic distance from critical band spectra: a first step. In *Proc. ICASSP*, pages 1278–1281, 1982.

[77] H. Fletcher and W A Munson. Loudness, its definition, measurement and calculation. *J.Acoust. Soc*, 5:82–108, 1933.

[78] M-H. Yang and N. Ahuja. Detecting human faces in color images. In *Proc. ICIP*, volume 1, pages 127–130, 1998.

[79] W. Yang, M. Benbouchta, and R. Yantorno. Performance of the modified bark spectral distortion measure as an objective speech quality measure. In *Proc. ICASSP*, pages 541–544, 1998.

[80] W. Yang. Enhanced modified bark spectral distortion (EMBSD): An objective speech quality measure based on audible and cognition model. *PhD. thesis, Temple University*, 1999.

[81] S. Voran. Objective estimation of perceived speech quality, part II: Evaluation of the measuring normalizing block technique. *IEEE Transactions on Speech and Audio Processing*, 1999.

[82] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra. Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs. *ICASSP*, pages 749–752, 2001.

[83] A.W. Rix. A new PESQ-LQ scale to assist comparison between P.862 PESQ score and subjective MOS. *ITU-T delayed contribution COM12-D086*, 2002.

[84] A.W. Rix. Analysis of P.862 PESQ scores using PESQ-LQ and alternative logistic mapping. *ITU-T delayed contribution COM12-D124*, 2003.

[85] T. H. Falks and W. Y. Chan. Objective speech quality assessment using gaussian mixture models. In *22nd Biennial Symposium on Comm*, 2004.

[86] T. H. Falks, Q. Xu, and W. Y. Chan. Non-intrusive GMM-based speech quality measurement. *ICASSP*, 2005.

[87] T. H. Falks and W. Y. Chan. A sequential feature selection algorithm for GMM-based speech quality estimation. *13th European Signal Proc. Conf.*, 2005.

[88] T. H. Falks and W. Y. Chan. Non-intrusive speech quality assessment using gaussian mixture models. *IEEE Signal Proc. Letters*, 13(2):108–111, 2006.

[89] T. H. Falks and W. Y. Chan. Enhanced non-intrusive speech quality measurement using degradation models. *ICASSP*, 2006.

[90] J. G. Beerends, E. Larsen, N. Iyer, and J. M. V. Vugt. Measurement of speech intelligibility based on the PESQ approach. *Measurement of Speech and Audio Quality in Networks (MESAQIN)*, 2004.

[91] T. Yamada, M. Kumakura, and N. Kitawaki. Word intelligibility estimation of noise-reduced speech. In *Proc. ICSLP*, pages 169–172, September 2006.

[92] N. Kitawaki and T. Yamada. Subjective and objective quality assessment for noise reduced speech. In *ETSI Workshop on Speech and Noise in Wideband Communication*, pages 1–4, May 2007.

[93] K. Manohar and P. Rao. Speech enhancement in nonstationary noise environments using noise properties. *Speech Communication*, 48(1):96–109, 2006.

[94] F. J. Fraga, C. A Ynoguti, and A. G. Chiovato. Further investigations on the relationship between objective measures of speech quality and speech recognition rates in noisy environments. In *Proc. ICSLP*, pages 185–188, 2006.

[95] H. Sun, L. Shue, and J. Chen. Investigations into the relationship between measurable speech quality and speech recognition rate for telephony speech. In *Proc. ICASSP*, pages 865–868, 2004.

[96] T. P. Barnwell and W. D Voiers. An analysis of objective measures for user acceptance of voice communicaton systems. *Final Report, DCA100-78-C-0003*, September 1979.

[97] J. M. Tribolet, P Noll, B. J. McDermott, and R. E. Crochiere. A study of complexity and quality of speech waveform coders. In *Proc. ICASSP*, pages 586–590, 1978.

[98] D. Pearce and H. G. Hirsch. The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *Proc. ICSLP*, volume 4, pages 29–32, 2000.

[99] B. Juang and L. R. Rabiner. A probabilistic distance measure for hidden markov models. *ATT Bell Laboratories Technical Journal*, pages 391–408, 1985.

[100] J. K. Baker. The dragon system - an overview. *IEEE Trans. on ASSP*, 23:24–29, 1975.

[101] C. M. Chernick, S. Leigh, K. L. Mills, and R. Toense. Can speech recognizers measure the effectiveness of encoding algorithms for digital speech transmission? *National Technical Information Service, PB99-127979*, 1999.

[102] ETSI Standard Document. Speech processing, transmission and quality aspects (STQ): Distributed speech recognition; front-end feature extraction algorithm; compression algorithm". *ETSI ES 201 108 v1.1.1*, 2000.

[103] C. Hory and N. Martin. Spectrogram segmentation by means of statistical features for non-stationary signal interpretation. *IEEE Trans. on Signal Processing*, 50:2915–2925, 2002.

[104] C. Hory and N. Martin. Mixture densities formulation of a spectrogram segmentation task. In *Proc. of EUSIPCO*, pages 427–430, 2002.

[105] W. M. Liu, V. J. Rivas Bastante, F. Romero Rodridguez, N. W. D. Evans, and J. S. Mason. Morphological filtering of spectrograms for automatic speech recognition. *Proc IASTED VIIP*, 2004.

[106] J. P. Barker, M. Cooke, and P. Green. Robust ASR based on clean speech models: An evaluation of missing data techniques for connected digit recognition in noise. In *Proc. Eurospeech*, volume 1, pages 213–216, 2001.

[107] A. Vizinho, P. Green, M. Cooke, and L. Josifovski. Missing data theory, spectral subtraction and signal-to-noise estimation for robust ASR: an integration study. In *Proc. Eurospeech*, volume 5, pages 2407–2410, 1999.

[108] Bregman. *Auditory Scene Analysis.* MIT Press, 1990.

[109] A. Martin and M. Przybocki. The NIST speaker recognition evaluation series, national institute of standards and technology's website. *http://www.nist.gov/speech/tests.spk.*

[110] A database for speech intelligibility testing using japanese word lists (fwo3), NTT advanced technology corporation, tokyo, japan. *http://www.ntt-at.com*, 2003.

[111] J. Makhoul, R. Schwartz, and A. El-Jaroudi. Classification capabilities of two-layer neural nets. In *Proc. ICASSP*, volume 1, pages 635–638, May 1989.

[112] V. N. Vapnik. *Statistical Learning Theory.* Wiley, New York, 1998.

[113] A.Pentland and M.Turk. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, pages 71–86, 1991.

[114] W. Zhao, A. Krishanaswamy, and R. Chellappa. Discriminant analysis of principal components for face recognition. In *Proc. International Conf on Automatic Face and Gesture Recognition*, pages 336–341, 1998.

[115] W. Zhao, R. Chellappa, A. Rosenfeld, and P. Phillips. Face recognition: A literature survey. *UMD CfAR Technical Report CAR-TR-948*, 2000.

[116] S. Madhvanath and V. Govindaraju. The role of holistic paradigms in handwritten word recognition. In *Proc. Trans. on Pattern Analysis and Machine Intelligence*, 2001.

[117] R. Vera Rodriguez, R. P. Lewis, N. W. D. Evans, and J. S. D. Mason. Feature optimization on geometric and holistic approaches for a footstep biometric verification system. In *Proc. 4th Summer School for Advanced Studies on Biometrics for Secure Authentication: New Technologies and Embedded Systems*, 2007.

[118] Douglas A. Reynolds. Speaker identification and verification using gaussian mixture speaker models. *Speech Communication*, 17(2):91–108, August 1995.

[119] W. M. Campbell, D. E. Sturim, and D. A. Reynolds. Support vector machines using GMM supervectors for speaker verification. In *Proc. Interspeech*, pages 3117–3120, 2005.

[120] B.G.B. Fauve, D. Matrouf, N. Scheffer, J.-F. Bonastre, and J.S.D. Mason. State-of-the-art performance in text-independent speaker verification through open-source software. *Audio, Speech and Language Processing, IEEE Transactions on*, Vol. 15, No. 7:1960–1968, September 2007.

[121] S. J. Young and P. C. Woodland. *HTK: Hidden Markov model toolkit V1.4 User manual.* Cambridge University Engineering Department, Speech Group, 1992.

[122] D. E. Sturim. Tracking and characterization of talkers using a speech processing system with a microphone array as input. *PhD. thesis, Brown University*, 1999.

[123] D. E. Sturim, D. A. Reynolds, E. Singer, and W. M. Campbell. Speaker indexing in large audio databases using anchor models. In *Proc. ICASSP*, volume 1, pages 429–432, 2001.

[124] Y. C. Yang, M. Yang, and Z. H. Wu. A rank based metric of anchor models for speaker verification. *IEEE Int. COnf on Multimedia and Expo*, pages 1097–1100, 2006.

[125] J. G. Beerends. Extending p.862 PESQ for assessing speech intelligibility. *White contribution COM 12-C2 to ITU-T Study*, Group 12, October 2004.

[126] J. M. McQueen. *The Handbook of Cognition.* Sage Publications, London, 2004.

[127] J. R. Pierce. Whither speech recognition? *Journal of the Acoustical Society of America*, 46(4), 1969.

[128] T. H. Falks, W. Y. Chan, and P. Kabal. Feature mining for GMM-based speech quality measurement. *38th Asilomar Conf. on Signals, Systems, and Computers*, 2004.

[129] X. Yang, J.B. Millar, and I. Macleod. On the sources of inter- and intra-speaker variability in the acoustic dynamics of speech. In *Proc. ICSLP*, volume 3, page 1792, 1996.

Part IV

Appendices

Appendix A

Databases

Two databases are used in this research:

(1) The first database is introduced in Section 3.2 and consists of 6 data sets, often referred to as test sets DS1 to DS6. This database has ground truth of intelligibility coming from the humans and is mainly used in Part I for investigation on the quality measures and ASRs.

(2) The second database is introduced in Section 7.2. The database is created for the development and evaluation of the direct, data-driven, differential intelligibility classifier (D^4IC). This database has synthetic ground truth deduced using the Intelligibility Enhancement (IE) Hypothesis proposed in Section 6.3. There are 4 data sets, namely 1 training set, 2 development test sets and 1 evaluation test set, referred to as Train, DevI, DevII and Eval respectively.

The raw signals for both databases come from the ETSI-Aurora2 digit-string corpus [26] which is derived from the TIDigits database. There are 11 words in the vocabulary, namely the digits one - nine, 'oh' and zero. Signals are taken from Test Set A defined in the Aurora2 framework which consists of 4004 digit strings collected from 104 adult speakers (52 male and 52 female). 'clean' signals are the original 20kHz signals from the TIDigits downsampled to 8kHz and filtered with the G.712 characteristic [26]. All the 4-digit strings in the Aurora2 Test Set A are extracted, totalling 566 utterances. Examples of signals are '139O' (one-three-nine-oh), '9486' (nine-four-eight-six), etc.The set of 566 four-digit clean signals becomes the origin, from which all databases are generated by processing the clean signals through various degradations.

A.1 Test Sets DS1 to DS6

Among the 6 test sets, 2 sets consider additive (environmental) noises, 2 sets consider degradation introduced by coding schemes and another 2 sets consider degradation introduced by enhancement processes. Brief descriptions of the test sets are presented in Table A.1.

Type	Test Set	Descriptions
Additive	$DS1_{add}$	additive noises of diverse characteristics including both speech-like and more stationary noises.
	$DS2_{add}$	additive noises, mostly fairly stationary.
Coding	$DS3_{cod}$	car noise and tandemings of single coding schemes
	$DS4_{cod}$	various $DS1_{add}$ noises and tandeming of mixed coding schemes
Enhancement	$DS5_{enh}$	car noise and different speech enhancement processes
	$DS6_{enh}$	various $DS1_{add}$ noises and different configurations of NLSS

Table A.1: Brief descriptions of the 6 test sets.

Details of each test set including sources of degradations are listed below:

(1) $DS1_{add}$:

- Degradation Type: Additive or environmental noises
- Degradation: 1) airport, 2) babble (crowd of people), 3) car, 4) exhibition hall, 5) restaurant, 6) street, 7) subway (suburban train) and 8) train station.
- Abbreviation: 1) airport, 2) babble, 3) car, 4) exhibition, 5) restaurant, 6) street, 7) subway, 8) train
- Description: Environmental noises of diverse characteristics including speech-like, stationary, impulsive, periodic.
- SNR range: 5dB to -10dB at -0.5dB intervals
- Number of degradations: 8
- Source of degradations: ETSI Aurora2 framework [26] [98]
- Noise adding software: ITU-T Rec., P.56 [51]

(2) $DS2_{add}$:

- Degradation Type: Additive or environmental noises
- Degradation: 1) aircraft cockpit, 2) city rain, 3) flat communication channel, 4) automobile highway, 5) helicopter fly-by, 6) large city, 7) large crowd, 8) IBM cooling fan, 9) SUN cooling fan, 10) white Gaussian noise.
- Abbreviation: 1) aircraft 2) cityrain, 3) flatch, 4) highway, 5) heliflyby, 6) largecity, 7) largecrowd, 8) ibmcoolfan, 9) suncoolfan, 10) Gaussian.

- Description: Environmental noises which are fairly stationary.
- SNR range: 5dB to -10dB at -0.5dB intervals
- Number of degradations: 10
- Source of degradations: Center for Spoken Language Understanding (CSLU) [?]
- Noise adding software: ITU-T Rec., P.56 [51]

(3) DS3$_{cod}$:

- Degradation Type: Coding schemes
- Degradation & abbreviation: 1) GSM, 2) 2GSM, 3) 4GSM, 4) 6GSM, 5) LPC, 6) 2LPC, 7) 3LPC, 8) 4LPC, 9) MELP, 10) 2MELP, 11) 4MELP, 12) G723_24, 13) G723_40, 14) G721_32, 15) LDCELP

  (abbreviation in the form of x.CODEC where x is the number of tandeming, for eg, 2GSM means that signals are firstly degraded by car noise, then endecoded by GSM twice.)
- Description: Car noise and tandemings of single codec. (all signals degraded by car noise then endecoded by selected coding scheme for 1 or more times)
- SNR range: 10dB to -10dB at -1dB intervals
- Number of degradations: 15
- Source of coding schemes:
  - GSM (13kbps) [52] downloaded from http://kbs.cs.tu-berlin.de/ jutta/toast.html;
  - LPC-10e (2.4kbps) [54] downloaded from ftp://ftp.super.org/pub/speech/lpc-1.0.tar.gz;
  - Federal standard MELP (2.4kbps) downloaded from http://www.data-compression.com/melp1.2.tar.gz;
  - low-delay-CELP (16kbps) [53] downloaded from http://svr-ftp.eng.cam.ac.uk/pub/comp.speech/coding/ldcelp-2.0.tar.gz;
  - the 3 CCITT ADPCM coders, namely G721 4-bit (32kbps), G723 3-bit (24kbps) and G723 5-bit (40kbps) downloaded from ftp:ftp.cwi.nl/pub/audio/ccitt-adpcm.tar.gz

(4) DS4$_{cod}$:

- Degradation Type: Coding schemes
- Degradation & abbreviation:

  1) Train.GSM.MELP.LPC, 2) Train.G721.G723(40).LPC,
  3) Train.G721.LPC.CELP, 4) Train.MELP.GSM.G721,
  5) Train.CELP.LPC.GSM, 6) Train.CELP.GSM.GSM,
  7) Train.LPC.MELP.CELP, 8) Train.LPC.GSM.MELP,
  9) Rest.LPC.G723(24).CELP, 10) Rest.G723(24).MELP.GSM,

11) Rest.CELP.GSM.MELP, 12) Rest.MELP.LPC.GSM,
13) Rest.MELP.GSM.LPC, 14) Rest.GSM.MELP.MELP ,
15) Rest.LPC.CELP.GSM
(abbreviation in the form of Noise.CODEC1.CODEC2.CODEC3,
for eg, Train.GSM.MELP.LPC refers to train noise degraded signals
endecoded firstly by GSM, followed by MELP then by LPC.)

- Description: Train noise or street noise from $DS1_{add}$ with tandemings of mixed codec. (all signals degraded by either train or street noise then endecoded by 3 selected coding schemes)
- SNR range: 10dB to -5dB at -0.5dB intervals
- Number of degradations: 15
- Source of coding schemes: same as $DS3_{cod}$

(5) $DS5_{enh}$:

- Degradation Type:: Enhancement processes
- Degradation:
  1) spectral substraction (SS) from Boll [56],
  2) SS from Boll with quantile filtering [56],
  3) SS from Berouti [57],
  4) SS from Kamath [58],
  5) minimum mean-square error short-time spectral amplitude [59].
  6) Wiener filter [60]
  7) SS with modified minimum statistics [61]
  8) perceptual wavelet filter (PWF) optimized for car and factory noise [62],
  9) PWF optimized for hearing aids [62]
  10) PWF optimized for noise recognition [62]
  11) PWF-nr with continuous buffer [62]
  12) lattice filter optimized for hearing aids [63]
  13) lattice filter optimized for ASR in car and factory noise [64]
  14) car noise
- Abbreviation: 1) SSBoll, 1) SSBoll79, 3) SSBerouti79, 4) SSKamath02, 5) MMSE, 6) Wiener, 7) SSmms, 8) PWF-cf, 9) PWF-ha, 10) PWF-nr, 11) PWF-nr2, 12) LF-ha, 13) LF-asr 14)car
- Description: car noise degraded signals processed by various enhancement algorithms.
- SNR range: 0dB to -10dB at -0.5dB intervals
- Number of degradations: 14

- Source of enhancement processes: signals are added with car noise in Swansea, enhancement processing carried out by Dr Tuan V. Pham from the Signal Processing & Speech Communication Laboratory at Graz University of Technology in Austria [62,64].

(6) $\mathbf{DS6}_{enh}$:

- Degradation Type:: Enhancement processes
- Degradation & abbreviation:

  1) Street.NLSS.a1.5-b0.03, 2) Babble.NLSS.a4-b0.1,
  3) Subway.NLSS.a3-b0.01, 4) Babble.NLSS.a2-b0.32,
  5) Street.NLSS.a2.5-b0.15, 6) Train.NLSS.a3-b0.05,
  7) Car.NLSS.a5-b0.01, 8) Car.NLSS.a4.5-b0.5,
  9) Airport.NLSS.a4-b0.02, 10) Airport.NLSS.a1.5-b0.03,
  11) Car.NLSS.a3-b0.05, 12) Exhibition.NLSS.a4-b0.1,
  13) Subway.NLSS.a2-b0.3, 14) Airport
  15) Car, 16) Train
  (abbreviation in the form of Noise.NLSS.ax-by, where ax-by is NLSS configuration;
  x is the noise over estimate and y is the noise floor.
  For eg, Street.NLSS.a1.5-b0.03 refers to signals degraded by street noise then processed by NLSS with noise over estimate set to 1.5 and noise floor set to 0.03).
- Description: Various noise from $\mathrm{DS1}_{add}$ processed by NLSS at a chosen configuration.
- SNR range: 5dB to -10dB at -0.5dB intervals
- Number of degradations: 15
- Source of enhancement process: Evans [35]

A.2 D^{4}IC Data Sets

Four data sets are created for the development and evaluation of the D^{4}IC. All data sets are generated using the data generation procedure proposed in Section 7.1 where IE lines are formed by progressively processing a set of signals with degradation randomly chosen from a pool, after which pairs of signals with assumed relative intelligibility are identified from within the same line or across IE lines. The 4 data sets consists of 1 training set, 2 developmentall test sets and 1 evaluation test set, referred to as Train, DevI, DevII and Eval respectively. The purpose and brief description of each data set is listed here:

(i) Train: To train the classifier. A sub-pool of 23 degradations are used for its generation. The degradations include 15 environmental noises, 5 coding algorithms and 3 NLSS processes.

(ii) DevI: The same pool for generation of Train is used here. However, signals pairs are deduced from independent sets of IE lines hence there is no overlap in Train and DevI due to different permutations of degradation when the lines are formed. However, since there is no unseen degradation good accuracy is expected. This test set aims to give confidence for the classifier setup.

(iii) DevII: The same pool for generation of Train plus 13 unseen degradations are used here. The unseen degradations are 2 environmental noises, 1 coding algorithm, 5 NLSS processes and 5 common speech processings. This test set is designed to be sufficiently challenging so that improved accuracies help identify good features.

(iv) Eval: The same pool for generation of Train plus 16 unseen degradations are used here. The unseen degradations are 4 environmental noises, 2 coding algorithm, 5 NLSS processes and 5 common speech processings. 3 of the unseen degradations are in DevII's pool. This test set is to evaluate the robustness of features identified during development stage.

A.2.1 Degradation Pools

The content of degradation pools used for the making of the 4 data sets are shown in Table A.2 to A.4. The percentage value shown in bracket represents the proportion for which degradations from that category can be chosen while forming the IE lines. For example, 35 IE lines with 10 processes per line are generated for the Train set hence in total there are 350 processes. According to Table A.2, environmental noises constitutes 50% of the total processes, i.e., 350*0.5 = 175 processes are additive environmental noise. Note that though processes are randomly chosen from the pool, the first process of an IE line is always from the category of additive noise. The same pool used for the making of the Train set is used for DevI hence there is no unseen degradations in this test set. However, unseen degradations exist in DevII and Eval. These degradations are in bold font in Table A.2 and A.3.

Degradation Pool for Train and DevI	
15 environmental noises(50%)	airport, babble, car, exhibition, subway, train, cityrain, flatch gaussian, heliflyby, highway, ibmcoolfan,largecity, largecrowd (at random SNRs in the range of 20dB down to -5dB).
5 coding schemes(40%)	GSM, LD-CELP, LPC, G723-24, G723-40
3 NLSS processes(10%)	a1.5-b0.03, a3-b0.01, a4-b0.02 where a is noise over estimate and b is noise floor

Table A.2: Degradation pool for the making of Train and DevI.

Degradation Pool for DevII	
17 environmental noises(30%)	airport, babble, car, exhibition, subway, train, cityrain, flatch gaussian, heliflyby, highway, ibmcoolfan,largecity, largecrowd **restaurant,suncoolfan** (at random SNRs in the range of 20dB down to -5dB).
6 coding algorithms (25%)	GSM, LD-CELP, LPC, G723-24, G723-40, **G721**
5 NLSS processes(25%)	**a4-b0.1**, **a2-b0.3**, **a2.5-b0.15**, **a6-b0.01** where a is noise over estimations and b is noise floor
5 common procedures(20%)	**resampling**, **lowpass**, **vibro**, **pitch**, **phaser**, (at various rates).

Table A.3: Degradation pool for the making of DevII.

Degradation Pool for Eval	
17 environmental noises (30%)	airport, babble, car, exhibition, subway, train, cityrain, flatch gaussian, heliflyby, highway, ibmcoolfan,largecity, largecrowd **restaurant,suncoolfan,street,cityrain** (at random SNRs in the range of 20dB down to -5dB).
6 coding algorithms (25%)	GSM, LD-CELP, LPC, G723-24, G723-40, **G721**,**MELP**
5 NLSS processes(25%)	**a1-b0.1**, **a2-b0.2**, **a5-b0.02**, **a4.5-b0.03**, **a6-b0.5** where a is noise over estimations and b is noise floor
5 common procedures (20%)	**highpass**, **chorus**, **flanger**, **fading**, **echo**, (at various rates).

Table A.4: Degradation pool for the making of Eval test.

Appendix B

Human Scores Profiles

Profiles of human scores for test set DS1 to DS6 are shown here:

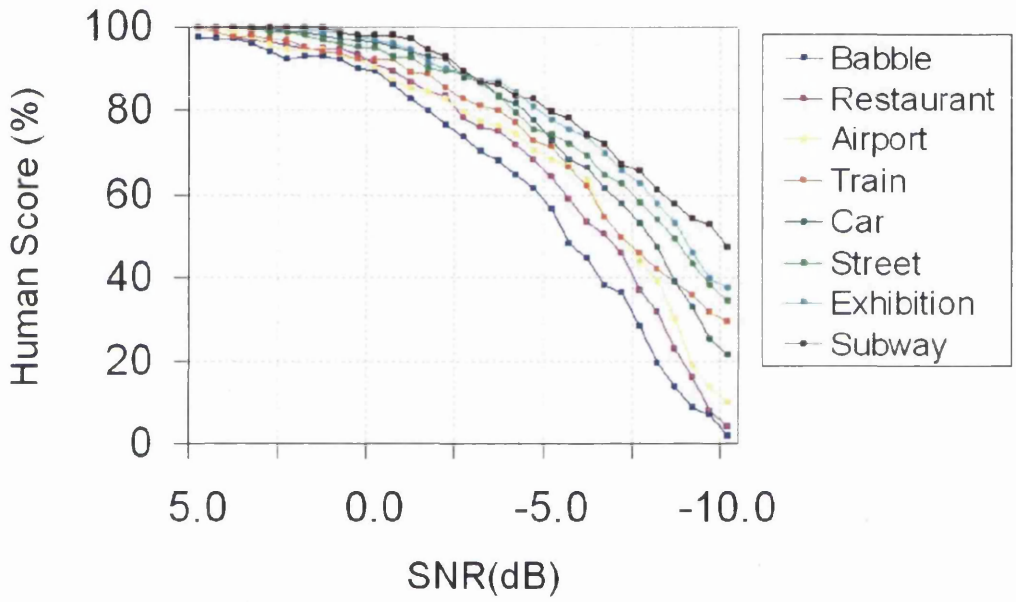


Figure B.1: Human scores for $DS1_{add}$.

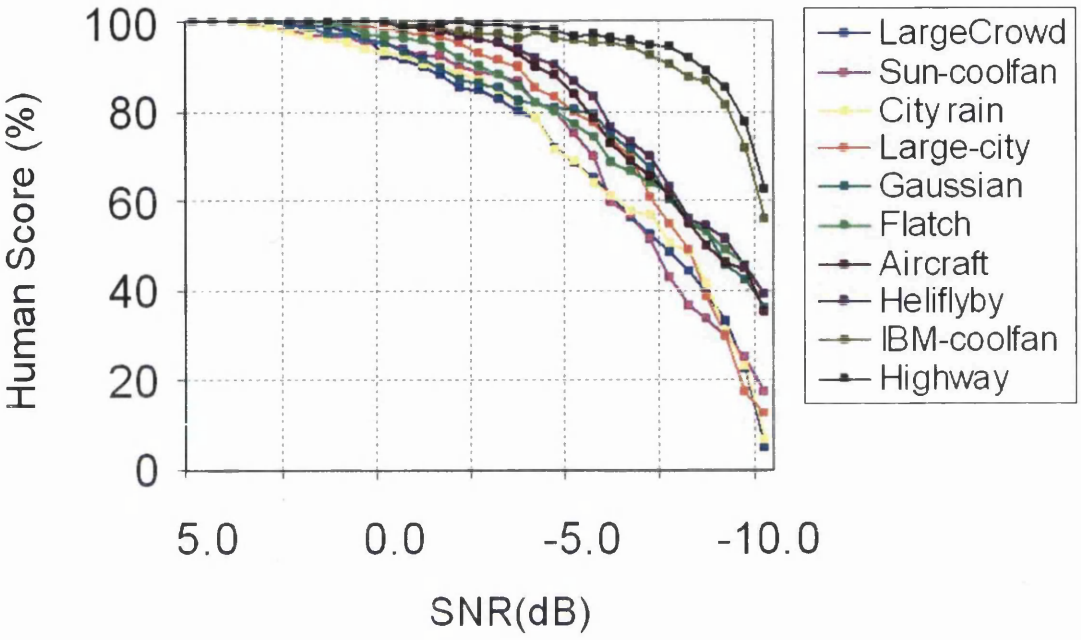


Figure B.2: Human scores for $DS2_{add}$.

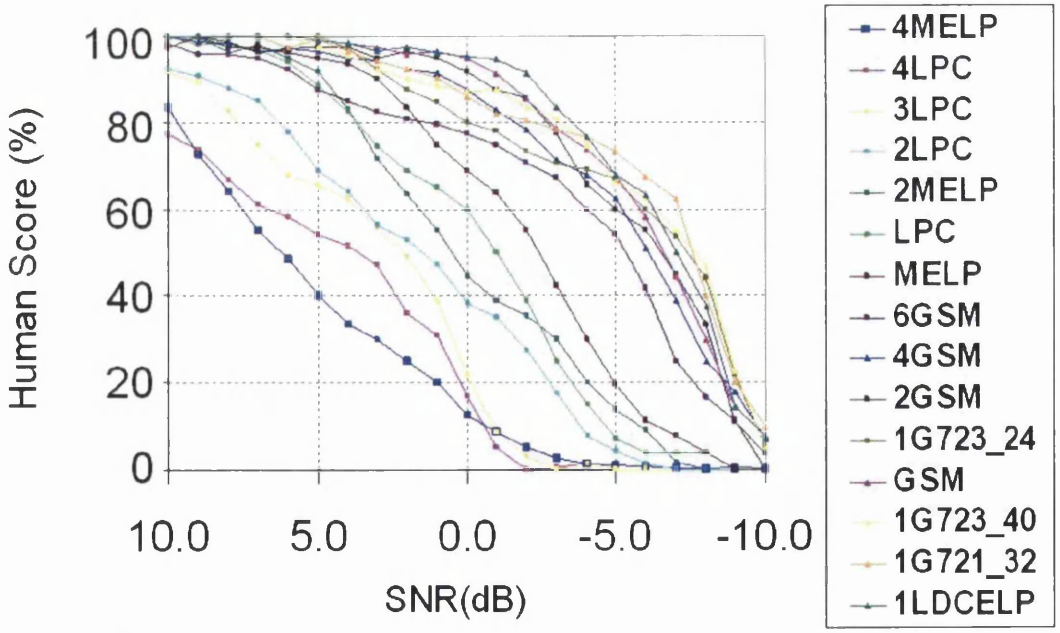


Figure B.3: Human scores for $DS3_{cod}$.

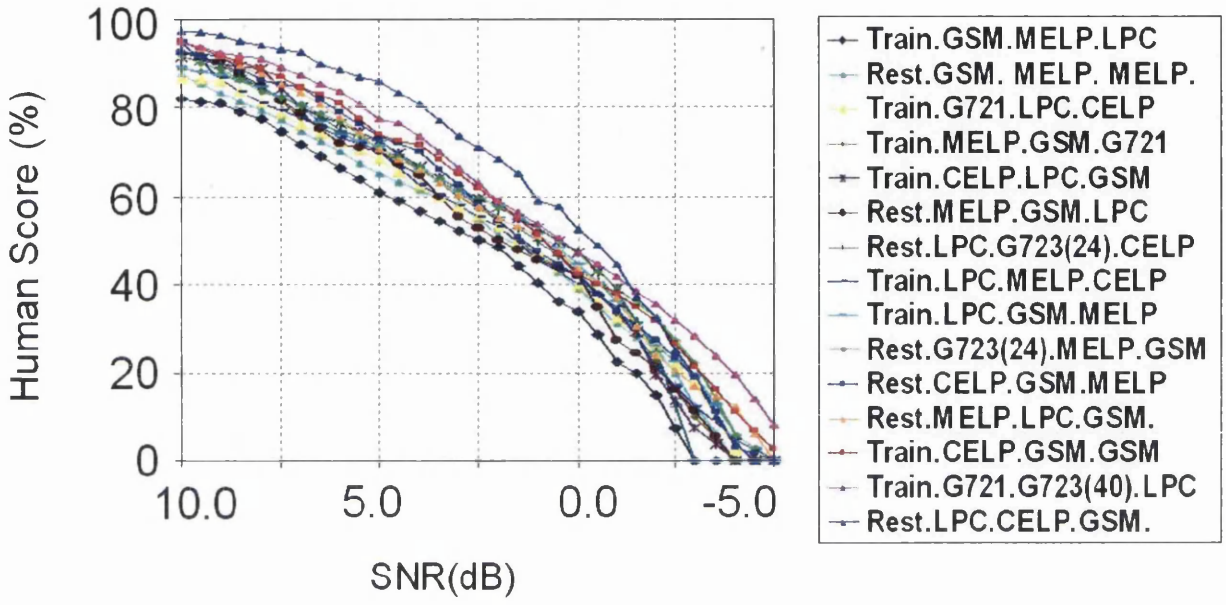


Figure B.4: Human scores for $DS4_{cod}$.

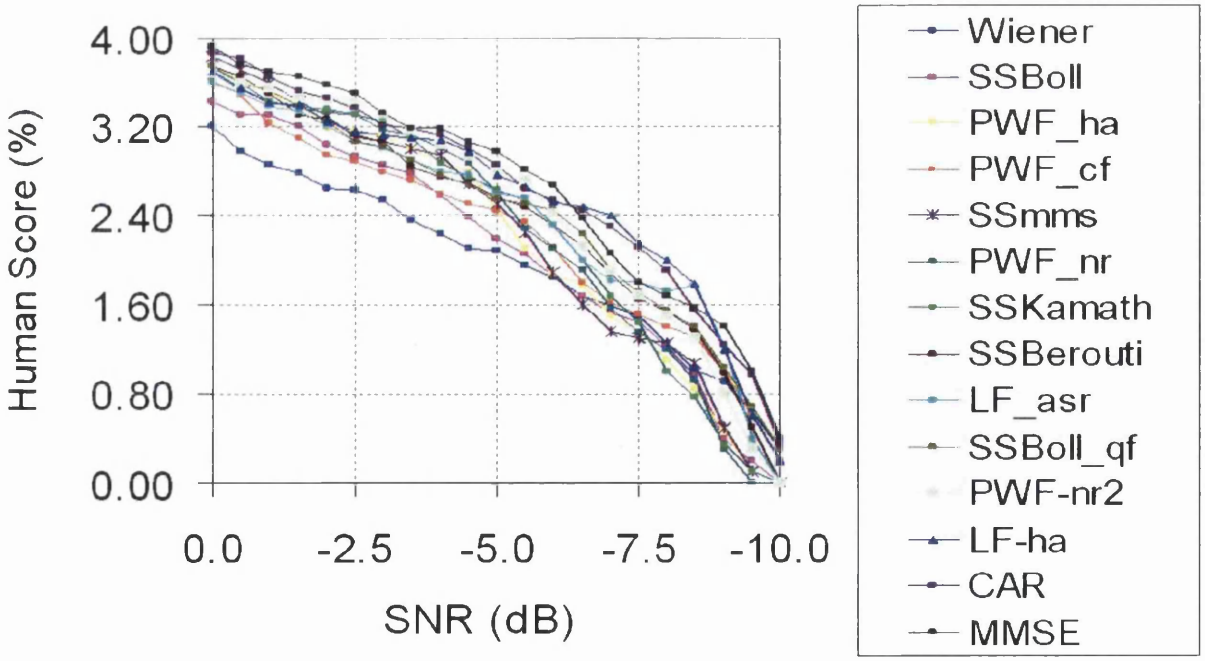


Figure B.5: Human scores for $DS5_{enh}$.

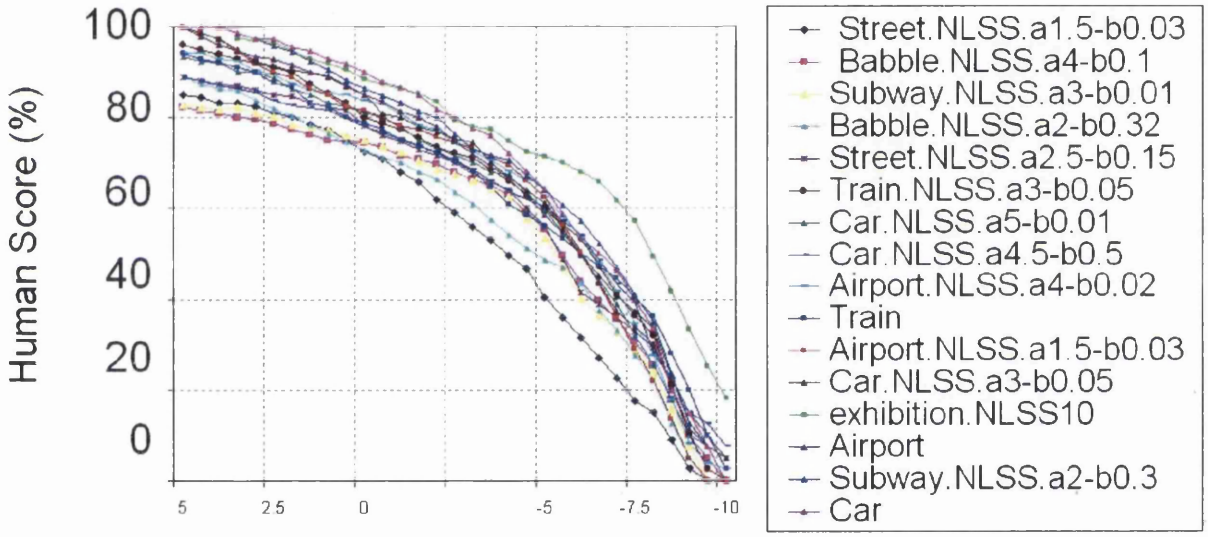


Figure B.6: Human scores for $DS6_{enh}$.

Appendix C

Part I Correlations

Pearson and Kendall$_2$ correlations obtained with the quality measures in Chapter 4 and ASR in Chapter 5 are shown here:

C.1 Correlations by Quality Measures

	CSNR	SEGSNR	IS	LAR	LLR	WSS	MNB	MBSD	PESQ
$DS1_{add}$	0.62	0.66	0.27	0.22	0.22	0.76	0.39	0.26	0.16
$DS2_{add}$	0.34	0.50	0.76	0.32	0.66	0.56	0.69	0.78	0.91
$DS3_{cod}$	0.82	0.81	0.68	0.79	0.85	0.88	0.9	0.83	0.83
$DS4_{cod}$	0.56	0.51	0.54	0.54	0.57	0.55	0.67	0.59	0.54
$DS5_{enh}$	0.46	0.42	0.45	0.33	0.34	0.74	0.54	0.49	0.54
$DS6_{enh}$	0.37	0.31	0.42	0.58	0.58	0.64	0.46	0.38	0.44
Average	0.53	0.54	0.52	0.46	0.54	0.69	0.61	0.56	0.57

Table C.1: Kendall$_2$ correlations obtained for the six test sets using quality measures.

	CSNR	SEGSNR	IS	LAR	LLR	WSS	MNB	MBSD	PESQ
$DS1_{add}$	0.33	0.33	-0.69	-0.82	-0.76	0.65	-0.32	-0.7	-0.74
$DS2_{add}$	-0.46	-0.03	0.71	-0.04	0.5	0.09	0.57	0.76	0.79
$DS3_{cod}$	0.81	0.79	0.52	0.69	0.74	0.82	0.88	0.77	0.79
$DS4_{cod}$	0.33	0.29	-0.21	0.18	0.2	0.04	0.18	-0.38	0.3
$DS5_{enh}$	0.16	0.27	0.54	0.28	0.24	0.64	0.4	-0.08	0.4
$DS6_{enh}$	-0.31	-0.39	-0.13	01.34	0.11	0.38	-0.11	-0.38	-0.12
Average	0.14	0.21	0.12	0.21	0.17	0.44	0.27	-0.01	0.27

Table C.2: Pearson Correlations obtained for test sets $DS1_{add}$ to $DS6_{enh}$ using the 9 quality measures.

C.2 Correlations by ASRs

Test Category	Test Set	WordAcc_clean		Corr_clean		Del_clean		Subst_clean		Ins_clean	
Environmental	$DS1_{add}$	0.85	0.76	0.38	0.57	0.23	0.48	0.78	0.58	0.72	0.48
	$DS2_{add}$	0.67		0.75		0.73		0.38		0.23	
Coding	$DS3_{cod}$	0.87	0.74	0.82	0.68	0.52	0.50	0.70	0.65	0.35	0.44
	$DS4_{cod}$	0.62		0.53		0.48		0.60		0.51	
Enhancement	$DS5_{enh}$	0.58	0.52	0.63	0.50	0.61	0.49	0.42	0.50	0.38	0.49
	$DS6_{enh}$	0.46		0.37		0.37		0.57		0.60	
Average		0.68		0.57		0.49		0.58		0.46	

Table C.3: $Kendall_2$ correlations obtained for test set $DS1_{add}$ to $DS6_{enh}$ using standard scores from a clean-trained ASR system.

Test Category	Test Set	WordAcc_clean		Corr_clean		Del_clean		Subst_clean		Ins_clean	
Environmental	$DS1_{add}$	0.80	0.62	-0.41	0.17	-0.74	0.02	0.79	0.14	0.74	-0.03
	$DS2_{add}$	0.44		0.73		0.77		-0.52		-0.80	
Coding	$DS3_{cod}$	0.87	0.70	0.80	0.53	0.08	-0.01	0.67	0.51	-0.34	-0.16
	$DS4_{cod}$	0.52		0.25		-0.11		0.34		0.02	
Enhancement	$DS5_{enh}$	0.31	0.12	0.42	0.03	0.34	0.00	-0.25	-0.04	-0.37	-0.06
	$DS6_{enh}$	-0.07		-0.37		-0.33		0.17		0.25	
Average		0.48		0.24		0.00		0.20		-0.08	

Table C.4: Pearson correlations obtained for test sets $DS1_{add}$ to $DS6_{enh}$ using standard ASR scores from a clean-trained ASR.

Test Category	Test Set	WordAcc_multi		Corr_multi		Del_multi		Subst_multi		Ins_multi	
Environmental	$DS1_{add}$	0.58	0.45	0.47	0.55	0.43	0.64	0.65	0.40	0.51	0.33
	$DS2_{add}$	0.32		0.62		0.85		0.15		0.16	
Coding	$DS3_{cod}$	0.88	0.80	0.86	0.74	0.15	0.27	0.89	0.77	0.93	0.77
	$DS4_{cod}$	0.72		0.62		0.39		0.64		0.61	
Enhancement	$DS5_{enh}$	0.25	0.42	0.24	0.41	0.20	0.27	0.62	0.68	0.68	0.72
	$DS6_{enh}$	0.57		0.57		0.39		0.74		0.75	
Average		0.56		0.57		0.39		0.62		0.61	

Table C.5: $Kendall_2$ correlations obtained for test sets $DS1_{add}$ to $DS6_{enh}$ using standard ASR scores from a multi-trained ASR.