Objective Speech Quality Measurement for Chinese Speech

A thesis
submitted in partial fulfilment
of the requirements for the Degree
of
Master of Science
in the
University of Canterbury
by
Fong Loong Chong

Examining Committee

Professor Dr Krzysztof Pawlikowski	University of Canterbury, New Zealand
Dr Ian McLoughlin	Tait Electronics Research, New Zealand
Associate Professor Dr Benjamin Premkumar	Nanyang Technological University, Singapore

University of Canterbury
2005

To Pei-Jung, Mum, and Dad

Abstract

Objective Speech Quality Measurement systems (OSQMs) have been found to provide high accuracy in measuring the speech quality of sound processing systems like codecs and telecom systems for English and some other European languages. However, the quality of sound systems used to process Chinese speech has not been adequately investigated to date. In order to accurately measure speech quality, speech intelligibility must first be optimised so that this attribute will not influence the measurement. While intelligibility can be high for sound processing systems in English or some European languages, this may not be true for Chinese speech due to two of its unique phonetic characteristics: the consonant-vowel-consonant (CVC) structure and use of tones. Each of these two characteristics can affect intelligibility of Chinese speech. The intelligibility issue that is related to the CVC structure is called *consonantal intelligibility* while that from the use of tones is known as *tonal intelligibility* in this research. The degradation in these two intelligibility types may not be taken into account by the OSQMs and therefore result in an inaccurate quality rating. The first purpose of this thesis was to evaluate OSQMs to investigate whether they regarded the degradation in Chinese speech intelligibility in their computation of an objective quality score. After evaluating the OSQMs, it was found that correlation between both consonant and tonal intelligibility, and quality is low. To resolve the problem of a low correlation between consonant intelligibility and quality, the second purpose of this thesis was to expose or magnify the discrepancies that cause intelligibility degradations so as to improve the OSQMs' sensitivity toward consonantal intelligibility. Two methods namely *high pass filtering* and *consonant amplification* were proposed for improvement. While both methods yielded improvements, it was concluded that the consonant amplification method is more effective than high pass filtering such that it yielded a better correlation.

Table of Contents

List of Tables

List of Figures

Acknowledgments

This project was supported financially by a Technology for Industry fellowship (TIF) from the Foundation for Research, Science and Technology through Tait Electronics Ltd. The support of both is deeply appreciated. My greatest gratitude goes to my supervisor Professor Krzysztof Pawlikowski and associate supervisor Dr Ian McLoughlin for their patient guidance. Without their help, this thesis would not have become a reality. The kind participants of my experiments will also be remembered for their time and efforts. Last but not least, I am deeply grateful to my wife and my parents, who were constantly behind me throughout this project, and my Father in heaven who led me all the way. May all the praises and glory be upon Him.

0.1 Abbreviations

ACR	. . . .	Absolute Category Rating
AD	. . . .	Auditory Distance (in MNB)
BBC	. . . .	British Broadcasting Corporation
BM	. . . .	Basilar Membrane
C1	. . . .	Initial Consonant of Chinese CVC Syllable
CB	. . . .	Critical Bandwidth
CDRT	. . . .	Chinese Diagnostic Rhyme Test
CDRT-Tone	. . . .	Chinese Diagnostic Rhyme Test - Tone
CF	. . . .	Characteristic Frequency (corresponding to lowest hearing threshold of an auditory nerve fibre)
CPA	. . . .	Chinese Phonetic Alphabet
CVC	. . . .	Consonant-Vowel-Consonant
CVR	. . . .	Consonant-Vowel Ratio
DCR	. . . .	Degradation Category Rating
DMOS	. . . .	Degradation Mean Opinion Score
DRT	. . . .	Diagnostic Rhyme Test
f0	. . . .	Fundamental Frequency
F1, F2, F3	. . . .	First, Second, and Third Formant
f_c	. . . .	Centre Frequency (of critical bandwidth)
FMNB	. . . .	Frequency MNB
HPF	. . . .	High Pass Filter
IPA	. . . .	International Phonetic Alphabet
ITU	. . . .	International Telecommunications Union
ITU-T	. . . .	Telecommunication Standardisation Sector of ITU
LD-CELP	. . . .	Low Delay Code Excited Linear Prediction
MNB	. . . .	Measuring Normalizing Block
MTF	. . . .	Modulation Transfer Function (in STI)
MOS	. . . .	Mean Opinion Score

OSQM	. . . .	Objective Speech Quality Measurement system
PESQ	. . . .	Perceptual Evaluation of Speech Quality
SNR	. . . .	Signal-to-Noise Ratio
SPL	. . . .	Sound Pressure Level
STI	. . . .	Speech Transmission Index
TMNB	. . . .	Time MNB

Chapter I

Introduction

In the search for the optimisation of transmission speed and storage, speech information is often coded, or transmitted with a reduced bandwidth. As a result, quality and/or intelligibility are sometimes degraded. Speech quality is normally defined as the degree of goodness in the perception of speech while speech intelligibility is how well or clearly one can understand what is being said. In order to assess the level of acceptability of degraded speeches, various subjective methods have been developed to test codecs or sound processing systems. Although good results have been demonstrated with these, they are time consuming and expensive due to the necessary involvement of teams of professional or naive subjects[1][56]. To reduce cost, computerised objective systems were created with the hope of replacing human subjects [90][43]. While reasonable standards have been reported by several of these systems, they have not reached the accuracy of well constructed subjective tests yet [92][84]. Therefore, their evaluations and improvements are constantly been researched for further breakthroughs. To date, *objective speech quality measurement systems* (OSQMs) have been developed mostly in Europe or the United States, and effectiveness is only tested for English, several European and Asian languages but not Chinese (Mandarin) [38][70][32].

The motivation for this research arises from the fact that Chinese (note, in this thesis "Chinese" refers to Mandarin, the official dialect of People's Republic of China also spoken widely in Malaysia, Singapore, Taiwan, and in other communities worldwide) is spoken by over a billion population throughout the world, and therefore an OSQM suited for Mandarin would benefit this enormous population. Besides this, Chinese speech has its own unique characteristics that are not found

[1] Subjects will mean human participants that participated in the subjective tests. Professional subjects will mean trained subjects while naive will mean untrained.

in most other languages. These characteristics may aggravate the degradation in speech intelligibility after processing which might not be evident to existing OSQMs in their computation of Chinese speech quality.

One might question, *"Should speech intelligibility be considered in the measurement of speech quality?"* and *"What is the relationship between these two speech attributes?"* The answer to the first question is *yes*. This answer and the answer to the second question will be discussed later. Steeneken and Houtgast stated in [77] that speech quality assessment is normally used for communications with high intelligibility. When the OSQMs regard the intelligibility of processed speech in English or some European and Asian languages to be high, they would also consider the same for Chinese speech not knowing that intelligibility could be affected by speech processing. The accuracy of quality measurements, therefore, lie in doubt. If there is indeed a relationship between speech quality and intelligibility, an effective OSQM should detect the acoustic discrepancies arising from the speech processing process that degrades intelligibility. An appropriate quality score should be computed according to the level of intelligibility. The objective of this research was firstly to evaluate OSQMs to investigate whether they regarded the degradation in Chinese speech intelligibility in their computation of an objective quality score. If indeed they did not take intelligibility into account appropriately, our second objective was to expose or magnify these discrepancies of the speech signals for the OSQMs.

The structure of this thesis is thus: Chapters 2, 3, and 4 will provide background information and context for this research. Chapter 2 will extend the context by presenting an overview of the human auditory process: information beneficial in the understanding of the perceptual model incorporated in the latest OSQMs. Proceeding this, several key aspects regarding speech will be mentioned in chapter 3. In it, the speech production process and characteristics of speech shall be discussed. Since we deal with Chinese speech, the last section in this chapter will introduce the unique characteristics of Chinese speech. After this, chapter 4 will discuss and introduce various subjective and objective speech quality and intelligibility measurement tests or systems. The respective tests or systems to be involved in our research will be discussed in more detail to conclude the introduction of the background for this research.

Chapters 5 and 6 constitute the main findings of our research. The answer to the two questions posed earlier shall be answered in Chapter 5. It also records the evaluation of two common OSQMs, namely Perceptual Evaluation of Speech Quality (PESQ) and Measuring Normalizing Blocks (MNB), with regards to the unique characteristics of Chinese speech. Two suggestions to expose or magnify the acoustics discrepancies of the processed speech for the OSQMs shall be mentioned in Chapter 6. Evaluations done for these methods will also be discussed.

Finally, this thesis concludes with a summary of the research, and suggestions for future work.

Chapter II

The Human Auditory Process

2.1 Introduction

The aim of our research is to evaluate OSQMs with a view to improving them such that they can more effectively be used to measure the objective quality of Chinese speech, and in particular to provide information on speech intelligibility of Chinese speech. Since recently developed OSQMs incorporate a perceptual model that mimics the human perception of speech (please refer to section 4.3.2), knowledge of the human auditory system aids in understanding the perception model. This chapter thus begins with an introduction to the physiology of the human auditory system followed by a discussion of the psychological aspects of human hearing otherwise known as psychoacoustics.

2.2 The human auditory system

The human ear can be considered as a complex signal processing system as it has the ability to capture sounds of complex frequencies, process them and send the processed signals to the human brain. With this ability, it allows humans to judge the differences in sound intensities, pitch frequencies, even estimate distances from which sound originate. We shall now discuss how our ears receive and process sound into signals to be interpreted by our brain.

The general human auditory system consists of two fundamental regions where auditory processing takes place (chapter 3 of [104]). The first region is the peripheral region where acoustical signals are converted into potential differences that initiate neural activity in the second region.

The second region involves neural processing that contributes to the auditory sensation where there are approximately 30,000 auditory nerve fibres in each ear

(chapter 1 of [61], and [36]) transmitting auditory information from the innermost part of the peripheral region to the human brain.

2.2.1 Peripheral region of the human auditory system

The peripheral region is made up of three parts: the outer ear, middle ear, and the inner ear. The outer ear includes the *pinna*, which is the part protruding out of the head, and the *meatus* or *auditory canal*. The pinna receives sounds to modify or filter them to be channelled to the middle ear via the auditory canal. The middle ear consists of the ear drum or tympanic membrane, and the ossicles which includes the malleus, incus, and stapes. The ossicles are known to be the three smallest bones in the human body. When sound has been channelled through the auditory canal, the ear drum vibrates and the ossicles transmits these vibrations to the inner ear. They work like a hammer (malleus) hammering the anvil (incus) that in turn causes the stirrup (stapes) to vibrate on the oval window, which is a membrane covered opening to the cochlea (inner ear). Within the inner ear is a spiral shaped cochlea (that resembles a snail) that has tough and hard walls and contains two types of fluids. The length of the cochlea is about 32mm long (chapter 3 of [104]) when it is unwound and there are two membranes that run along its length, the *Reissner's membrane* and the *basilar membrane* (BM). The BM is the membrane that relates to the frequencies of sound and the Reissner's membrane merely provides a separation between two channels in the cochlea. One end of the cochlea is known as the *base* and the other is called the *apex*. The base is where the oval window lies and the apex is the inner end of the cochlea.

The relationship between the middle and inner ear is thus: since the cochlea (inner ear) is filled with fluids (which is denser than air), when sound waves reached the oval window, most of them will be reflected instead of directly causing a vibration movement on it. In this case, no acoustic information will be passed to the inner ear. Therefore, the middle ear plays an important part in translating the vibrations caused by sound waves in the air to the vibrations in the fluids in the cochlea. Due to this difference in *acoustic impedance*, the middle ear performs an impedance matching between two different mediums. When the stapes causes a movement on the oval window, this in turn stirs up a vibration in the BM. Within the cochlea, the peak of the vibrations that arise from different frequencies, how-

ever, do not occur at the same position along the BM. Lower frequencies cause the peak to occur at the apex as it is wider and less stiff compared to the base. Therefore, higher frequencies will not cause much movement towards the apex. For this property, the cochlea is regarded as a Fourier analyser as different points on the BM counteract with different frequencies. The frequency of vibrations along the points of the BM that arise from a particular sound wave has the same frequency to that wave. However, the phase along different points where vibration occurs is different. Lying on the BM is the organ of *Corti* which contains one row of inner hair cells on one side and up to five rows of outer hair cells on the other side. On each hair cell are "hairs" know as the *stereocilia*. There are about 140 stereocilia on each outer hair cell and 40 on each inner ones. There is another membrane called the *tectorial* membrane on the other side of the hair cells. When sound waves are present which causes the BM to vibrate, this vibration causes the stereocilia on the inner hair cells to be displaced between the BM and the tectorial membrane. Consequently, potassium ions flow into the hair cell and this results in a potential difference between the inside and outside of the cells. This sparks the neural response and send signals to the second region of the auditory system (chapter 1 of [61]).

2.2.2 Neural processing in the human auditory system

The second region in the human auditory system is where neural processing occurs. Movements along the BM that were caused by the stimulation of sound were transmitted to the brain through approximately 30,000 auditory nerve fibres (chapter 1 of [61], chapter 3 of [104], and [36]) transmitting auditory information from the innermost part of the peripheral region to the human brain.

When sound is present, neural impulses (spikes) are transmitted through these nerve fibres. The impulse rate or number of spikes depends on the loudness level. There are, in fact, several properties of the auditory nerve fibres in relation to the neural impulse rates and they will be mentioned as followed.

Tuning curves and tonotopic organisation

Each nerve fibre corresponds to a certain position on the BM. This means to say that each fibre carries a range of frequencies (neural tuning curves (chapter 1 of

[61], and [47])) and different fibres corresponds to different frequency ranges. A particular frequency known as the *characteristic frequency* (CF) corresponds to the lowest hearing threshold of an auditory nerve fibre. This CF is also the frequency which causes the greatest vibration on a point along the BM. Not only does each fibre relate to a particular part along the BM, even the orientation of the fibres is related. The organisation of the nerve fibres, known as a *tonotopic organisation* [36], is such that the fibres along the outer edge of the fibre bunch associate with higher CFs and those at the centre of the bunch with lower CFs.

Spontaneous firing rate

It was realised that even without any sounds, neural impulses existed at a significant but slower rate which is called the *spontaneous firing rate*. The spontaneous firing rate also varies between different nerve fibres which ranges from approximately 0 to 150 per second (chapter 1 of [61]). Usually, nerve fibres with a lower neural threshold have high spontaneous rates and vice versa, where the threshold is the lowest sound level which causes a stimulation on each nerve fibre.

Phase locking

Phase locking occurs for frequencies below 4-5 kHz at each nerve fibre. When a pure tone was heard, the impulse responses in the fibres seem to be synchronous with the frequency of the tone. For example when a 1 kHz tone (period of 1 ms) was heard, the peaks of the impulses also occurred at intervals of approximately 1 ms. This phenomena of phase locking, however, disappears at about 4 kHz or slightly higher. This is due to decreasing intervals between impulse peaks with increasing frequency (decreasing period) until a point where no distinct peak occurs (chapter 1 of [61]).

Two tone suppression

At the presence of a tone whose frequency is approximately equivalent to the CF of a nerve fibre mentioned in section 2.2.2, a burst of impulses will occur followed by a period of steady impulses that is lower than the initial burst. When another tone is introduced, a change to the rate of impulses occurs according to

the frequency of that tone. If the frequency lies within the tuning curve of that fibre, this will lead to an increase in the impulse rates. However, if the frequency of that tone lies marginally outside of the tuning curve, the impulse rates that arise from the first tone will be reduced or suppressed until the second tone is removed [36].

2.3 Psychoacoustics

In order to understand the human auditory process, knowing merely the anatomy of our auditory system and how it works is insufficient. The relationship between the physical properties of sound and its human perception is also vital. This relationship enables researchers in the audio processing field to develop models or systems that could simulate our complex auditory system. The science that studies this relationship is known as *Psychoacoustics*.

The term "sound" relates to three basic properties: intensity, frequency, and timbre. There are various issues relating these three properties which should be noted in the design of auditory models. The study of Psychoacoustics, therefore, provides a deeper insight regarding humans' perception of sound that will aid the designing process. The following issues regarding humans' sound perception will be briefly discussed in the proceeding subsections:

- Loudness perception
- Concept of critical band
- Masking
- Pitch perception

2.3.1 Loudness perception

It is difficult for one to describe the loudness of sound in terms of a certain scale as it is a subjective sensation almost differing among human beings. More reliable is for humans to give a rating of it on a numerical scale to match loudness against a given reference tone (for example a 1 kHz sinusoidal) to that of the tone being

tested. The latter, although requires some effort, has been put to good use (see *Equal-loudness contours* in next paragraph). This measure, known as the *loudness level* was introduced by Barkhausen in the twenties (chapter 8 of [104]). The unit of this scale is "phon" which is equivalent to the sound pressure level (SPL) of a 1 kHz tone in dB SPL. To determine the loudness (phon) of another tone, the loudness of the 1 kHz[1] reference tone is adjusted to match that of the tone being tested. Hence this loudness level is not exactly how loud the tone being tested, rather, how loud a 1 kHz tone would sound to match the loudness of this tone.

A set of *Equal-loudness contours* (figure 2.1) can be derived from the loudness level. The 1 kHz tone is set to a certain value, sound pressure levels for a range of frequencies that matches the loudness of this 1 kHz were determined. Hence any frequency along this contour will sound equally loud and they shared the same phon value. The SPL of the 1 kHz tone was then increased to another fixed level and SPLs for the range of frequencies were again recorded. This procedure is done for a range of SPLs for the 1 kHz tone. The lowest curve among the equal-loudness contours represents the *absolute hearing threshold* which is the lowest loudness level of a tone our ears can detect. The opposite of the absolute hearing threshold is the *threshold of pain* where it is the loudest sound level a human being could bear. This upper hearing threshold lies approximately at 140 dB SPL regardless of frequency (chapter 3 of [59]). We also realised from the contours that the curves seem to be have higher variations (of loudness level) at lower levels and are flatter at high levels. This explains why we could hardly hear the bass of an audio signal when the volume is relatively low but it could sound as loud as the higher frequencies when volume is high. The equal-loudness contours have been used in areas like the designing of amplifiers, objective speech quality measurement systems [43][93], etc.

When measuring the loudness level of complex sounds, it would be inaccurate to calculate the average from the sum of loudness levels across a frequency range. As it is previously mentioned, the variations on an equal-loudness curve is greater at lower sound levels and hence lower, and perhaps higher (approximately > 16 kHz) frequency sounds seem neglected by the human ears. To compensate for

[1] The 1 kHz tone is often used as a reference or common standard tone in electro- and psycho-acoustics (chapter 8 of [104]).

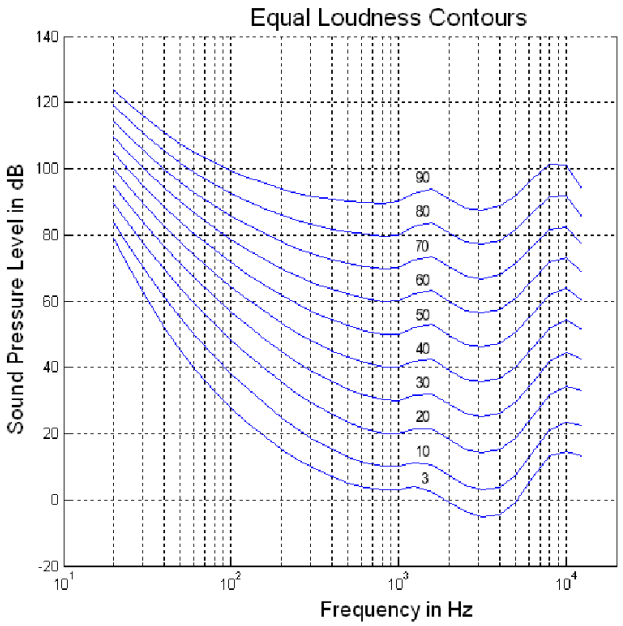

Figure 2.1: Equal loudness Contours. *Based on ISO 226.*

the lower and higher frequencies in the measurements, an A-weighted decibel (dBA) scale is adopted which takes into account the insensitivity for lower and higher frequencies at lower sound levels. The A-weighted decibel (dBA) scale is based approximately on the 30 phon contour and below (chapter 4 of [61]). For high sound levels, where the equal-loudness contour is flatter, a C-weighting which treated low and high frequencies fairly equal in loudness is used. The dBC weighting is generally used for loudness level above 85 phons. The median B-weighting is used for loudness level of around 70 phons.

In order to scale loudness so that linearly increasing the loudness scale would lead to a linearly proportional increase in subjective loudness (for example, doubling the unit on the scale will cause the subjective loudness to be doubled), the "sone" loudness scale was introduced. The sone scale starts at 40 dB SPL of a 1 kHz tone (i.e. 1 sone = 1 kHz @ 40 dB SPL). It was experimentally determined that a 10 dB increase in sound level equals the effect of doubling the subjective loudness (chapter 8 of [104]). Therefore a 1 kHz at 50 dB SPL would be 2 sones, 60 dB SPL 4 sones, and so on. The relationship between sones and phons can be approximated by the equation (chapter 6 of [94]):

$$phon = 40 + 10\log_2(sone) \qquad \text{or} \qquad sone = 2^{\frac{(phon-40)}{10}}$$

2.3.2 Concept of critical band

It was mentioned in section 2.2.1 that the cochlea acts like a Fourier analyser as different positions along the basilar membrane (BM) counteract with different frequencies. The BM can thus be viewed as having a series of bandpass filters with different centre frequencies along its length. The passbands of the filters also overlap one another. When a signal masked with background noise is presented to a human subject, it is assumed that a particular filter along the BM with centre frequency (f_c) nearest to the frequency of the masked signal receives this signal. When the bandwidth of the background noise centred at the signal is broadened, the threshold of this signal increases. This increase will happen until a point where the threshold will remain almost constant even when bandwidth still increases. The bandwidth of noise at which no further increase in signal threshold occurs is called the *critical bandwidth* (CB) (chapter 3 of [61]). This CB is also the bandwidth of the filter at which the signal was captured. Therefore when a complex sound is heard, the respective filters along the BM would receive the particular signal whose frequency is nearest to their f_cs.

The CBs, however, are not constant along the length of the BM, i.e. not constant as frequency increases. Figure 2.2 shows that CBs from 0 Hz to approximately 500 Hz are constant with a bandwidth of 100 Hz. From 500 Hz to about 3 kHz, the increase in CB is lower than that of frequency, and after 3 kHz, CB increases faster. It is sometimes assumed that there is no overlapping of the bandwidths of filters where the upper cutoff frequency f_u of a filter is exactly the lower cutoff frequency f_1 of the next filter. Table 2.1 shows the experimentally determined values of lower and upper cutoff frequencies corresponding to the respective filters with given centre frequency. A value is given for each frequency where the f_u of one filter is the f_1 of the next ranging from 0 to 15500 Hz where there are 24 critical bands along this range. These range of values are known as the *critical-band rates* scale having its unit as *Bark*[2]. These critical-band rates also

[2] Named after Barkhausen, a scientist who studied the auditory perception of loudness (chapter

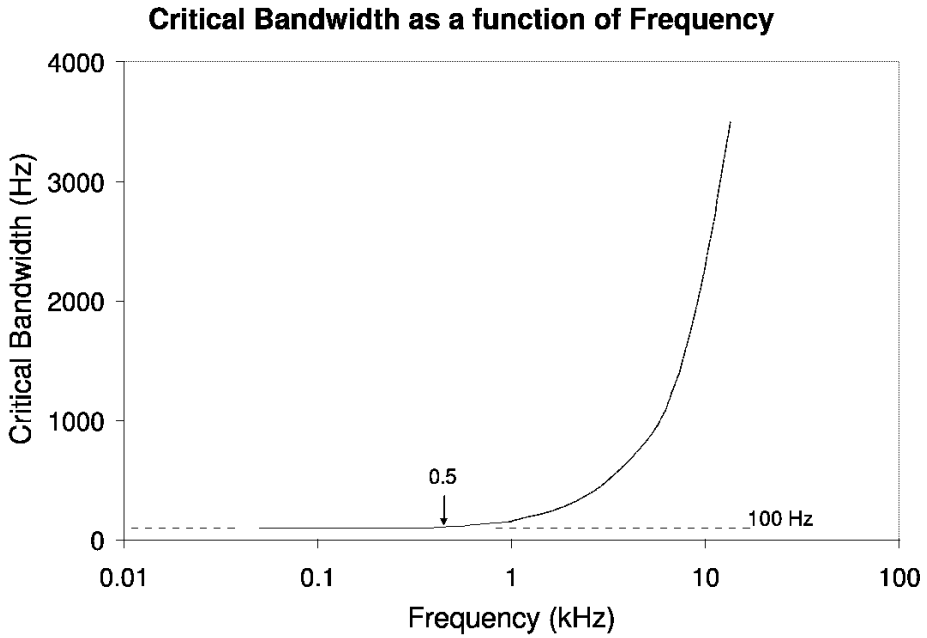

Figure 2.2: Critical Bandwidth as a function of frequency. *Redrawn from figure 6.8 in chapter 6 of [104].*

allow us to understand the length along the BM that corresponds to different frequencies. When the unwound BM is compared against the critical-band rates and frequencies ranging from 0 to 16 kHz, matching the 32 mm length to 24 Barks and 16 kHz ranges (figure 2.3[3]), it was realised that the frequency scale is not proportional to the length along the BM but rather adopts the relationship between critical-band rates and frequency. 1 Bark corresponds to about 1.3 mm along the BM. Near the apex end of the cochlea, the BM corresponds to lower frequencies and the frequency scale was linear up to about 500 Hz. After that, the frequency scale is approximately logarithmic up until reaching the base end (oval window). This relationship between the length of BM, and the frequencies associated with it along its length, and the critical-band rates is important to studies in the electro- and psycho-acoustical fields (chapter 6 of [104]).

6 of [104])

[3] Also shown in this figure is the length along the BM that corresponds to ratio pitch ranging from 0 to 2400 mel, to be discussed later in section 2.3.3.

Table 2.1: Critical-band rate z, lower (f_1) and upper (f_u) cutoff frequencies of critical bandwidths Δf_G, with centre frequency at f_c.

z Bark	f_1, f_u Hz	f_c Hz	z Bark	Δf_G Hz	z Bark	f_1, f_u Hz	f_c Hz	z Bark	Δf_G Hz
0	0				12	1720			
		50	0.5	100			1850	12.5	280
1	100				13	2000			
		150	1.5	100			2150	13.5	320
2	200				14	2320			
		250	2.5	100			2500	14.5	380
3	300				15	2700			
		350	3.5	100			2900	15.5	450
4	400				16	3150			
		450	4.5	110			3400	16.5	550
5	510				17	3700			
		570	5.5	120			4000	17.5	700
6	630				18	4400			
		700	6.5	140			4800	18.5	900
7	770				19	5300			
		840	7.5	150			5800	19.5	1100
8	920				20	6400			
		1000	8.5	160			7000	20.5	1300
9	1080				21	7700			
		1170	9.5	190			8500	21.5	1800
10	1270				22	9500			
		1370	10.5	210			10500	22.5	2500
11	1480				23	12000			
		1600	11.5	240			13500	23.5	3500
12	1720				24	15500			
		1850	12.5	280					

Data taken from chapter 6 of [104].

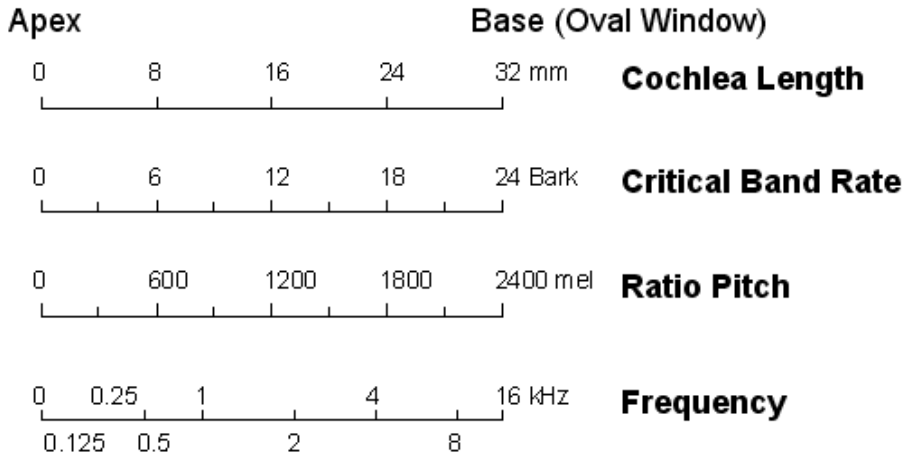

Figure 2.3: Scales of Critical-Band Rate, Ratio Pitch, and Frequency compared against the length of the unwound Cochlea. *Redrawn from figure 6.11 in chapter 6 of [104].*

2.3.3 Masking

Masking is the phenomenon whereby an audible sound is suppressed by another sound causing the original to appear weaker or inaudible. This phenomenon reflects the frequency selective ability within our ears. If our ears cannot effectively select the wanted tone (frequency) among a complex sound or noise, the wanted tone is said to be masked. In order for that tone to be heard, its loudness level must exceed a threshold value called the *masked threshold* (chapter 4 of [104]). The formal definition of masking by the American Standards Association is [14]:

1. The process by which the threshold of audibility for one sound is raised by the presence of another (masking) sound.

2. The amount by which the threshold of audibility of a sound is raised by the presence of another (masking) sound.

A given test tone(s) can be masked by noise, another pure tone, or complex tones all of which are known as maskers. The masker can be present either simultaneously (simultaneous masking) with, before (pre- or backward-masking), or after (post- or forward masking) the test tone(s). When the test tone(s) is totally

inaudible, total masking occurs while partial masking occurs when the loudness of the test tone(s) is reduced but still audible (chapter 4 of [104]).

Simultaneous masking

Simultaneous masking occurs when the whole duration of the test tone is being masked. There are two factors present in simultaneous masking (chapter 3 of [61]):

1. Swamping

2. Suppression

Swamping refers to the overwhelming of auditory information within a critical band (or an auditory bandpass filter) by the masker resulting in the test tone being left out or undetected. Hence the effect of swamping occurs when both the masker and the tone lies in the same CB. Suppression, however, occurs when the frequency of the test tone is above or below the maskers', lying in different CBs. The effect is similar to that of two tone suppression (section 2.2.2) where the tone in an auditory nerve fibre is being suppressed by a masker which will not cause auditory impulses to occur in the same fibre (critical band). When the masker itself covers a wide frequency range, for example wide band noise, both swamping and suppression occur in simultaneous masking.

Non-simultaneous masking

Non-simultaneous masking refers to masking where the masker is presented before or after the test tone. When the test tone is presented before the masker, it is called backward masking or premasking. Forward masking or postmasking refers to the case where the masker is presented before the test tone.

Backward masking is usually less obvious and it only occurs for a time 20 ms or less after the commencement of test tone. It exists when the build up time of the test tone (lower loudness level or faint ones) is slow and that of the loud masker is fast. In this case, the loud masker would be heard earlier than the test tone and if the masked threshold is not exceeded by the test tone, it is masked

(chapter 4 of [104]). In our context, backward masking may occur where the softer phonemes of a Chinese syllable (the initial consonant) is masked by the proceeding one (vowel) that is relatively louder.

Forward masking occurs when the test tone exists within 200 ms after the masker is switched off. It may be due to the time after the cessation of the masker where masking still exists within this period. After the masker is switched off, there is a residual "ringing" effect which lasted for about 150 to 200 ms. When this "ringing" is sufficiently loud, masking occurs. Another reason for forward masking may be due to fatigue of the auditory system after the presentation of the loud masker. Hence the test tone is neglected when the human subject has not recovered from this fatigue.

Pitch perception

Similar to loudness, pitch is also a subjective sensation which cannot be measured directly. It is defined by the American Standards Association as *"that attribute of auditory sensation in terms of which sounds may be ordered on a musical scale"* [14]. It is related to the frequency or fundamental frequency of a pure or complex tone. It is also related to the sound pressure level. Pitch increases with increasing frequency but for increasing loudness, pitch decreases for lower tone frequencies (approximately < 2 kHz) and increases for higher ones (approximately > 4 kHz). A generally accepted view of how our auditory system perceives pitch is the *place theory of hearing* (chapter 6 of [61]) where different places along the BM vibrates according to its associated frequency. It is assumed that the pitch corresponds to the place on the BM where this vibration is maximum (which relates to the CF). This in turn causes information to be transmitted to the brain through specific auditory nerve fibres that carry those frequencies.

A ratio pitch in *mels* (one mel is defined as an equal distance from one pitch to another along the scale using a subjective judgement [78]) is often used to measure pitch of pure tones. It begins with a subjective perception of what it sounds to be half the pitch of a test tone. A tone with a known frequency was presented to a human subject. The subject was then required to adjust the frequency of the tone until the new pitch sounded half of that of the original test tone. This half pitch frequency was collected for a range of frequencies. A relationship curve

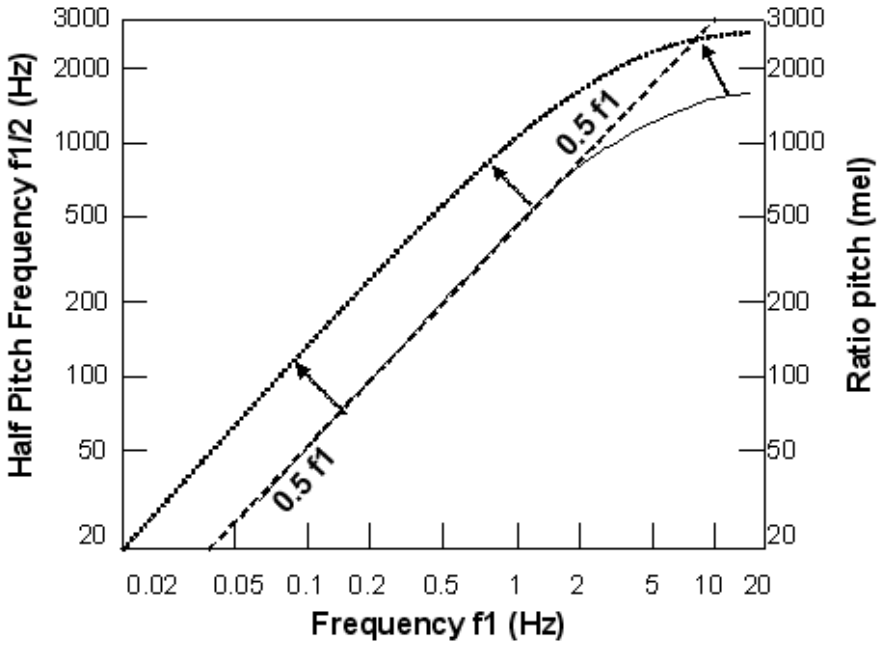

Figure 2.4: Relationship between Half Pitch frequency, Ratio Pitch and Frequency. *Redrawn from figure 5.1 in chapter 5 of [104].*

between the original frequency and the frequency that produces the half pitch was determined. It was realised that the frequency of the half pitch is almost half of that of the original tone's for frequencies below 500 Hz. Above that, the frequency of the original tone increases more than the half pitch's to get the same half pitch sensation. This relationship is similar to that between the critical band rate and frequency (please refer to figure 2.3 in section 2.3.2). This curve was then shifted by a factor of 2 to match the scale of the half pitch's and this half pitch scale became the mel scale (chapter 5 of [104]). Figure 2.4 depicts this relationship.

For a complex tone where the higher frequencies are harmonics of the lower one, for example, a complex tone containing frequencies 200, 400, 600, 800 Hz, ..., etc. The pitch of this complex tone is close to that of the fundamental frequency, in this case a low 200 Hz pitch. One might assume that removing the 200 Hz will yield a pitch of another frequency. However, the part that changes is the timbre of the tone instead of the pitch. The pitch sounds rather similar to that of the 200 Hz. Using the same 200 Hz harmonics tone where higher frequencies exists, removing all harmonics except those of the mid frequency ones like 1800,

2000, and 2200 Hz still give us the same pitch. The timbre, however, changes drastically. This similar pitch is known as the *residue pitch* and is different from that of the fundamental frequency though it sounded close. The positions on the BM that vibrate are also different from that which is caused by a pure tone. This means to say that the positions on the BM that responds to the middle or higher frequencies also allow a listener to hear a low pitch (chapter 6 of [61]).

2.4 Conclusions for the human auditory process

A brief introduction of the human auditory process was presented in this chapter. It included the physiological and psychological aspects of hearing. Physiologically speaking, the human auditory system consists of two fundamental regions: the peripheral and neural processing regions. Sound signals from the peripheral region are transmitted to the neural processing region through auditory nerve fibres. Regarding the psychological aspect of human hearing, four issues were mentioned. They were *loudness perception*, the *critical band* concept, *masking*, and *pitch perception*. The appreciation of these concepts gave us foresight into the perceptual models used by the objective speech measurement systems in chapter 4.

In the next chapter, we shall discuss issues regarding the general aspects of the speech. These include speech production, general characteristics of speech, and specifically the characteristics of Chinese speech.

Chapter III

Speech

3.1 Introduction

Speech is one of the elementary methods of communication. Besides speaking face-to-face, speech can also be propagated by other means. In today's world, speech communications can be through telephony, recording systems (cassettes, CDs, DVDs, and their players), the Internet, and so on. The design and operation of such systems requires knowledge of the characteristics of human speech in order to effectively and efficiently convey vocal content. This chapter will provide an introduction concerning general aspects of speech (predominantly based on English). We will first introduce the production of speech and then briefly discuss its characteristics in general. Since we are dealing with Chinese speech in particular, the unique characteristics of Chinese will be discussed at the end of the chapter.

3.2 Speech production

The structures in our body that together enable the production of speech sounds is known as the *vocal organs*. These consist of the lungs, trachea (windpipe), larynx (where the vocal cord or glottis is located), pharynx (throat), mouth, and the nasal cavities (chapter 9 of [61] and chapter 1 of [52]). Using the vocal organs, speech sounds are produced by two essential and one optional functional processes namely, *initiation*, *articulation*, and/or *phonation* (chapter 1 of [19]).

3.2.1 Initiation

For speech sounds to occur, air has to be present and it is usually provided by the lungs in our body. There are three types of initiation to the provision of air among

all languages of the world (chapter 2 of [19]):

1. Pulmonic, which involve the lungs,

2. Glottalic, which involves the vocal cord or glottis, and

3. Velaric, which involve the tongue, and the velum or soft palate (located at the top inner part of the mouth).

In both English and Chinese speech, only pulmonic initiation is adopted.

3.2.2 Articulation

After initiation, articulation takes place to transform the airflow into acoustic elements, that form different types of sound. Articulation can be performed by the glottis, upper surface of the vocal tract, teeth, tongue, and/or lips. Different methods of articulation exist for consonants and vowels. For consonants, air that flows from the initiation process is obstructed whereas for vowels, it remained relatively unobstructed. The places of articulation for producing basic English consonants are given in table 3.1. At most of these places of articulation, there are also various ways to articulate (chapter 1 of [52]):

- Stop - where airflow is stopped by the articulators to prevent air from escaping the mouth.

    1. Nasal Stop (Nasal) - Air is allowed to flow out of the nose by releasing the soft palate even though it is stopped in the oral cavity. Examples of nasal stops are the beginning of English words '**m**e' (bilabial closure), '**n**ight' (alveolar closure), and the end of word 'ha**ng**' (velar closure). Another term used by phoneticians for nasal stops is *"nasal"*.
    2. Oral Stop (Stop) - Air is completely stopped in this case where no air flows out of the mouth or nose. Air pressure is build up in the oral cavity and subsequently released in bursts. Examples of oral stops that occur at the beginning of English words are **p**it and **b**oy (bilabial closure), **t**ee and **d**ye (alveolar closure), **k**ite and **g**ird (velar closure).

Table 3.1: Places of articulations for producing English consonants.

Name	**Parts used for articulation**	**Consonants produced**[a]	**Examples**
Bilabial	Upper and lower lips	/p,b,m/	**p**it, **b**oy, **m**eat
Labiodental	Lower lip and upper front teeth	/f,v/	**f**ive, **v**owel
Dental	Tongue tip or blade and upper front teeth	/th/	**t**hese, **t**he
Alveolar	Tongue tip or blade and alveolar ridge	/t,d,n,s,z,l/	**t**ee, **d**ye, **n**ight, **s**ign, **z**eal, **l**oud
Retroflex	Tip of tongue and back of alveolar ridge	/r/	**r**ight, **r**ead
Palato-Alveolar	Tongue blade and back of alveolar ridge	/sh/	**s**heep
Palatal	Front of tongue and hard palate	/y/	**y**ellow
Velar	Back of tongue and soft palate	/h,k,g/	**h**ack, **k**ite, **g**ird

[a] Alphabets shown between /-/ represent English alphabets producing the sounds shown in the corresponding examples.

[b] The articulatory places are listed in ascending rows where the parts used to articulate is nearest to the outside of the mouth.

Words with oral stops at the end are ‘te**d**’ (alveolar stop) and ‘da**n**’ (alveolar nasal). The term *“stop”* commonly refers to oral stops.

- Fricative - where turbulent airflow is created due to partial obstruction that arose from close proximity of two articulators.

    1. Sibilants - Sibilants are louder in intensity and have higher pitches. Examples of sibilants are /s/ in ‘**s**ign’ and /z/ in ’**z**oo’.
    2. Non-sibilants - They are softer and have lower pitches than sibilants. Some examples are /th/ in ‘**th**ese’ and /f/ in ‘**f**it’.

- Approximant - similar to fricatives except that articulators are not so close as to producing a turbulent airflow. Examples of approximants are the beginning of ‘**y**ellow’ and ’**w**illow’.

- Lateral (Approximant) - An approximant produced by partial obstruction between one or both sides of the tongue and the roof of the mouth. Examples of laterals are the beginning of ‘**l**ie’ and end of ‘pa**le**’.

- Affricate - A stop followed by a fricative. An example is the beginning and end of ‘**ch**ur**ch**’ (palato-alveolar affricate).

- Flap (Tap) - A single tap by the tongue on the alveolar ridge. An example will be the middle of the word ‘be**tt**er’ when it is pronounced quickly (more common in American English).

- Trill (Roll) - A repeating or trilling action of the ‘r’ sound. Not so common in English.

For vowels, the airflow is smoother than consonants in that obstruction is not as great. Articulation involves the tongue and the lips. There are three classes in which a vowel can be classified (chapter 1 of [52]):

1. Position of tongue - The position of the tongue’s highest point within the mouth (e.g. f**ee**t (front), th**e** (centre), and g**oo**d (back)).

2. Height of tongue - The height of the body of tongue or the proximity between the tongue and the roof of the mouth (e.g. b**ee**t (high or close[1]), b**i**t (mid-high or close-mid), b**e**d (mid-low or open-mid), and b**a**d (low or open)).

3. Shape of lips - How "rounded" are the lips (e.g. f**ee**t (unrounded), h**oo**d (rounded)).

3.2.3 Phonation

Phonation refers to the voicing of a sound which relates to the vibration of the vocal cord or glottis. Although phonation is optional in speech production, it occurs in a non-negligible fraction of speech sounds. Excluding whispers, all English and Chinese vowels are voiced. Out of 24 English consonants from table 2.1 in chapter 2 of [52], 15 (62.5%) are voiced. For Chinese, however, only 4 (19%) out of 21 consonants (table 3.8 in section 3.4.1) are voiced. Some of the Chinese consonants that arise from the same articulation as its counterpart in English are unvoiced (for example consonants /b/, /d/, and /g/).

3.3 Characteristics of speech

Speech can be considered as a translation from one form (written or psychological) into comprehensible sounds of a particular language. Of the fact that it is based on sound introduces the various aspects of loudness, pitch, and so on, which were mentioned in the previous chapter. In this section, emphasis will be given on categorising speech sounds, and to discuss the characteristics of each category mentioned.

Each English word is made up of one or more syllables where a syllable is defined as *a minimal pulse of initiatory activity bounded by a momentary retardation of the initiator* by Catford in chapter 9 of [19]. This retardation is usually caused by an articulation of a consonant. However, a syllable itself seldom consists of purely one basic sound but can generally be broken up into yet smaller

[1] The first description in this bracket refers to the height of the tongue and the second relates to the proximity of the tongue to the roof of the mouth (this second description is used in the International Phonetic Alphabet).

units of elementary sounds. These elementary sounds are known as phonemes or basic speech sounds (chapter 9 of [61]). There are two categories of Phonemes: vowels (including diphthongs[2]) and consonants (including semi-vowels).

3.3.1 Phonetic transcription

In the English language, the Latin alphabet is used to denote phonemes, such that a word or syllable can be pronounceable by concatenating a few alphabetical characters. However, the same word or syllable in other languages, for example Chinese[3], does not necessarily use similar alphabetical means to represent sounds. In order for phoneticians to understand and pronounce speech sounds for different languages, a set of special alphabets developed by the *International Phonetic Association* [2] called the *International Phonetic Alphabets* (IPA) is used to represent most, if not, all speech sounds. Figure 3.1 reproduces the full IPA chart. Table 3.2 and 3.3 shows the phonetic transcription (IPA) for English (British[4]) consonants and vowels.

3.3.2 Consonants

Consonants are produced by the articulation of the upper surface of the vocal tract, teeth, tongue, and/or lips to obstruct the air that flows from the initiation process. Due to this obstruction and minimal vocal resonance (which is shorter in duration for voiced consonants), the relative intensity or power of consonants is generally lower than that for vowels. The duration of some consonants like stops is also very short compared to the vowel in a monosyllabic consonant-vowel-consonant (CVC) word (all single Chinese characters are monosyllabic (CVC) in pronunciation). These two relatively minute acoustic features make consonants more susceptible to masking and intelligibility loss. Despite the lower intensity and shorter duration that makes them easier to be confused, the consonants are more important for intelligibility [60][74]. Table 3.4 shows the power of some vowels and consonants. It was shown in the table that the average power of selected vowels is over 20 times more than that of consonants. In both English and Chinese, there

[2] Diphthongs and semi-vowels will be discussed in the respective vowels and consonants section.
[3] Before the romanisation process (please refer to section 3.4.1), and in its original literature.
[4] There are some differences in the pronunciation of English dialects.

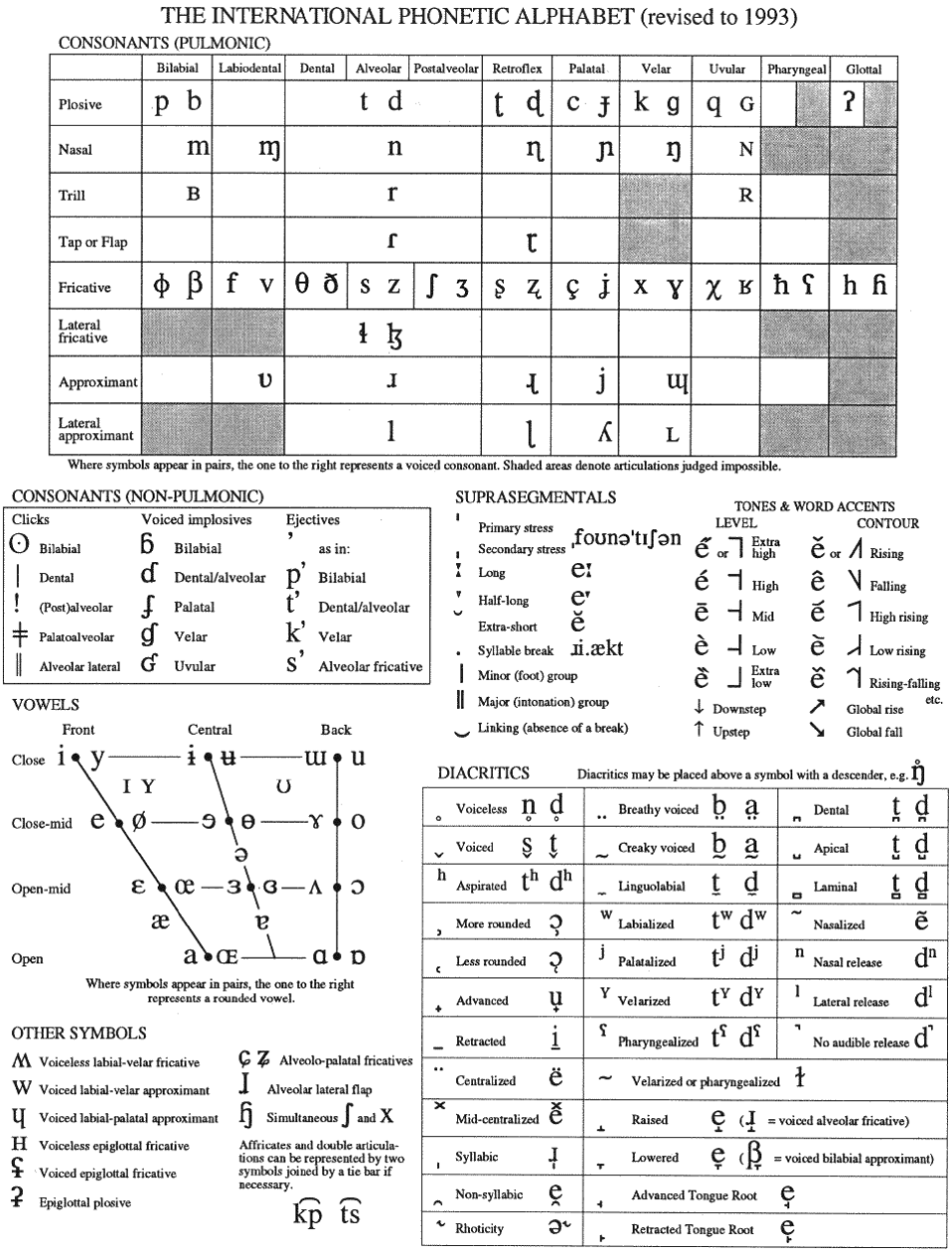

Figure 3.1: Full chart of the International Phonetic Alphabet (Revised to 1993, Updated 1996). *Image from International Phonetic Association (Department of Theoretical and Applied Linguistics, School of English, Aristotle University of Thessaloniki, Thessaloniki 54124, GREECE)[1]*

Table 3.2: IPA transcription of English consonants before vowels *e* and *ai*, or as an end consonant, with their articulation type.

IPA Symbol	Vowel **e**	Vowel **ai**	End	Articulation
b	bet	buy		stop
d	debt	die		stop
g	get	guy		stop
p	pet	pie		stop
t	ten	tie		stop
k	ken	kite		stop
w	wet	why		approximant
j(y)	yet			approximant
l	let	lie		approximant
r(ɪ)	retch	rye		approximant
m	met	my	ra**m**	nasal
n	net	nigh	ra**n**	nasal
ŋ			ra**ng**	nasal
f	fed	fie		fricative
θ			thigh	fricative
s	set	sigh		fricative
ʃ	shed	shy	mi**ss**ion	fricative
h	hen	high		fricative
v	vet	vie		fricative
ð	then	thy		fricative
z	Zen	Zion	mi**zz**en	fricative
ʒ			vi**s**ion	fricative
tʃ	Chet	chime		affricate
dʒ	jet	jive		affricate

Table taken from table 6.1 in chapter 6 of [53].

Table 3.3: IPA transcription of BBC English vowels and their corresponding examples between a pair of consonants.

IPA Symbol	**Examples between pair of consonants**			
i (iː)	b**ea**d	b**ea**t	h**ee**d	
ɪ	b**i**d	b**i**t	h**i**d	k**i**t
eɪ	b**aye**d	b**ai**t	h**aye**d	K**a**te
e	b**e**d	b**e**t	h**ea**d	
æ	b**a**d	b**a**t	h**a**d	c**a**t
ɑ (ɑː)	b**ar**d	B**ar**t	h**ar**d	c**ar**t
ɒ	b**o**dy	b**o**ttom	h**o**d	c**o**t
ɔ (ɔː)	b**aw**d	b**ough**t	h**awe**d	c**augh**t
ʊ	b**u**ddhist		h**oo**d	
əʊ	b**o**de	b**oa**t	h**oe**d	c**oa**t
u (uː)	b**oo**ed	b**oo**t	wh**o**'d	c**oo**t
ʌ	b**u**d	b**u**t	H**u**dd	c**u**t
ə (ɜː)	b**ir**d	B**er**t	h**ear**d	c**ur**t
aɪ	b**i**de	b**i**te	h**i**de	k**i**te
aʊ	b**owe**d	b**ou**t	h**ow**dy	
ɔʊ	B**oy**d		ah**oy**	q**uoi**t
ɪə	b**ee**r	p**ee**r	h**ere**	
eə	b**a**re	p**ea**r	h**ai**r	c**a**re
aə	b**y**re	p**y**re	h**i**re	
ʊɪ	b**oo**r	p**oo**r		

Table taken from table 3.3 in chapter 3 of [53].

Table 3.4: Average conversational power of speech sounds in microwatts.

Vowels		**Diphthongs**		**Semi-Vowels**		**Consonants**	
ɔ	47	əʊ	22	n	2.11	ʃ	1.83
ɑ	34	aɪ	20	m	1.85	tʃ	1.44
ɛ	17			ŋ	0.35	s	0.94
ʌ	15			l	0.33	z	0.72
u	13					dʒ	0.47
i	12					k	0.34
ə	10					t	0.14
æ	9					d	0.08
ɪ	9					f	0.08
						v	0.03
Average		18.9[a]		**Average**		0.8	

[a] This average value includes both vowels and diphthongs. Semi-vowels are included in the calculation of the average consonant power.

[b] *Values taken from table 3 in chapter 2 of* [59].

are some phonemes, which sound like an incomplete (non-syllabic) vowel, called *semi-vowels* (chapter 9 of [52]). They are produced by a rapid glide to its proceeding vowel. Since they appear in the same position as a consonant in a syllable (best seen in the CVC structure of a Chinese syllable), we shall consider them as consonants in our discussion and subsequent calculation of consonant power and duration. Some examples of semi-vowels are the /w/ and /y/ in the Chinese Hanyu Pinyin system, and the nasals.

3.3.3 Vowels

Generally, vowels are produced by the vibration of vocal cords from a pulmonic initiation with a relatively less obstructed articulation. Since this involves the vibration of vocal cords, vowels are voiced (excluding the whispering of vowels). A vowel sound is in fact a combination of resonating frequencies called formants. The first two formants (F1 and F2) are important in the determination of vowel intelligibility while the third (F3) contributes to its quality to some extent [39]. Formant frequencies of similar vowels produced by different speakers are quite

similar regardless of female or male voice (chapter 2 of [59]) although the pitch for a woman is generally about an octave higher than that for a man [81].

Another attribute related to vowels is pitch whose height is determined by the fundamental frequency, f0 (chapter 8 of [19]). This is the so-called base frequency we hear in the event of a complex sound where higher frequencies are harmonics of this f0. In tonal languages such as Chinese (including various dialects), pitch is the component that give tones to the syllables. Therefore, a distortion in pitch during a speech coding or transmission process will result in a possible change of tones (loss in tonal intelligibility).

A *diphthong* is a consecutive sequence or combination of vowels within one syllable (chapter 6 of [19]). Although a few vowels are concatenated, a diphthong sounds as a single vowel where the sound of the one vowel glides to the next. Some examples of diphthongs are the [aɪ][5] in bide and [aʊ] in bowed. Diphthongs will be considered as vowels in our research.

3.3.4 Frequency range of intelligible speech

During a telephone conversation, there are times where words are wrongly heard. For example, the sentence *"My name is Fong"* can sometimes be heard as *"My name is Thong"* or *" ... Hong"*. This is partly because the telephone bandwidth is band limited to a range from about 300 Hz to 3400 Hz [44][74] while the range of frequencies found in speech is from about 50Hz to over 10,000Hz [65]. Speech frequencies out of this telephone band are therefore removed or attenuated and hence either inaudible or distorted. To prevent this loss of intelligibility, frequency ranges of speech, in particular vowels and consonants, should be known (of course, for practical reasons like saving bandwidth, sometimes intelligibility have to be compromised).

It was earlier mentioned that the first two formants of vowels are important in the determination of vowel intelligibility and the third its quality. The frequency range of F1 for English vowels of a male speaker ranges from about 270 to 730 Hz, F2 from 840 to 2290 Hz, and F3 1690 to 3010 Hz (table 3.2 in [68]). Therefore, vowel intelligibility would be preserved as long as the frequencies from 270 Hz

[5] Alphabets shown between [-] represent the International Phonetic Alphabets (IPA) while those shown in /-/ represents English or the Chinese Hanyu Pinyin (to be discussed later).

to 2290 Hz are present.

For consonants, stops like /b,d,g,k/ have their greatest intensities within the telephone bandwidth range. /t/ has a slightly higher frequency for its peak intensity at about 4000 Hz[6]. Approximants /w, y, r/ have their frequency ranges corresponding to their F1s and F2s and all are within the telephone band. /l/, however, has got some formant energies below 500 Hz and at about 1500 Hz. The higher energies occurs at frequencies higher than 4000 Hz. For nasals, since they are voiced, their intelligible frequency range falls within that as vowel intelligibility is preserved. Unvoiced fricatives are the ones where the intelligible frequencies are higher than the vowels. This is especially so for /th [θ]/,/s/, and /f/ where their most intense energies lie above 4000 Hz outside the telephone band [74]. Hence those consonants that usually cause errors in telephone conversations are the fricatives and perhaps some of the stops as their duration is rather short.

Figure 3.2 shows the relationship between the percentage of correct syllables in an intelligibility test and cutoff frequencies of low- and high-pass filters. At a cutoff frequency of 2 kHz, we realised that 75% of the syllables were correct for both filters. At the highpass cutoff frequency of 6 kHz, no correct syllables were heard. Similarly, any frequency below 200 Hz is unintelligible when the low-pass filter is applied at that cutoff frequency. Therefore, in order to get a high intelligibility (say 95%), it is safe to retain frequencies above 700 Hz and below 4000 Hz. The other 5% that is unintelligible would very likely be the higher frequency consonants.

3.3.5 Loudness of intelligible speech

In a totally quiet environment, a soft whisper at a distance of say 1 m can be heard. However, in environments with substantial amount of background noise, no longer can the whisper be heard. Rather, volume has to be increased for clear communication. Usually, audiologists relate this clear or intelligible communication with a factor known as the *signal-to-noise* ratio (SNR), which is the ratio between sound pressure level of speech signals to ambient noise. Generally for an effective (intelligible) communication, an average SNR of at least +6 dB (the

[6] Frequency information for this paragraph are based on notes and interpreting spectrograms from chapter 6 of [53]

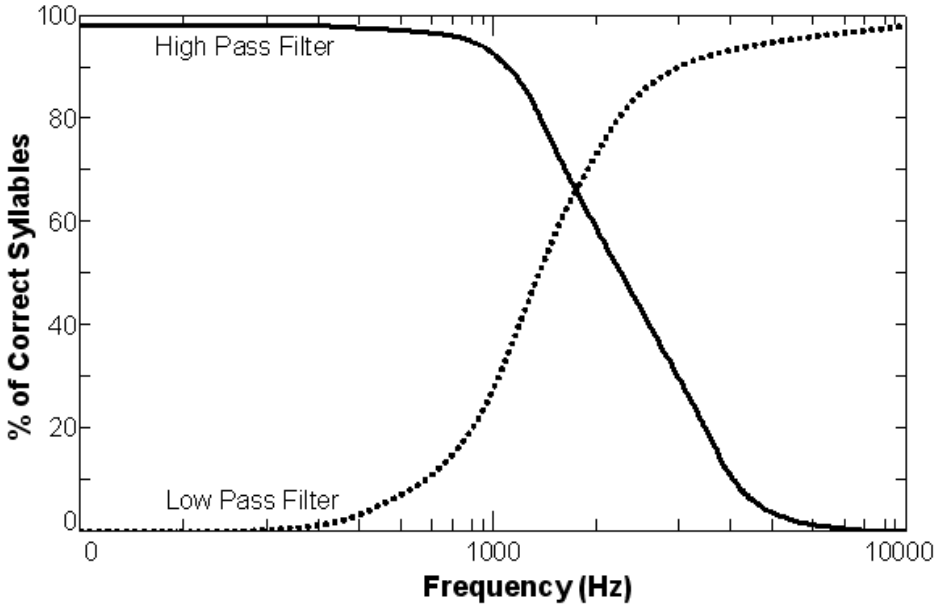

Figure 3.2: Relationship between Cutoff frequencies of Low- and High-Pass filters and percentage of correct syllables. *Redrawn from figure 23 in chapter 3 of [59].*

average speech level is 6 dB louder than noise) must be achieved (chapter 9 of [61]). However, this only applies to environments with noise level ranging from 30 to 110 dB. At high noise levels that exceed 110 dB, intelligibility will be affected having the same SNR (chapter 14 of [94]). A list of sound pressure levels for common indoor and outdoor noises is given in table 3.5

3.3.6 Speech contexts

Speech is not made up of merely one syllable or word. Usually, a speaker has to speak in a length of a phrase or sentence to properly convey a message. When phrases or sentences are spoken, words within it usually contribute to a common message or context. Because of this, someone could actually guess or anticipate a missing word in a sentence. For example, the sentence *"I _____ in the Computer Science and Software Engineering faculty at the University of Canterbury"* with a missing word, one would have contemplated the missing word to be *"study"* or *"lecture"*. This is because these words fit into the context of the sentence. It will not sound logical to guess the missing word as "jump" or "hitch-hike". We would

Table 3.5: Common Indoor and Outdoor Noises.

Indoor	**Sound Pressure Level**	**Outdoor**
Rockband at 5 m	110 dB	
	105 dB	Jet Flyover at 300 m
Inside Subway Train (New York)	99 dB	
	95 dB	Gas Lawn Mower at 1 m
Food Blender at 1 m	89 dB	
	84 dB	Diesel Truck at 15 m
Garbage Disposal at 1 m	81 dB	
	79 db	Noisy Urban Daytime
Shouting at 1 m	76 dB	
Vacuum Cleaner at 3 m	70 dB	Gas Lawn Mower at 30 m
Normal Speech at 1 m	66 dB	
	64 dB	Commercial Area
Large Business Office	56 dB	
Dishwasher Next Room	51 dB	
	50 dB	Quiet Urban Daytime
Small Theatre, Large Conference Room (Background)	41 dB	
	40 dB	Quiet Urban Nighttime
	34 dB	Quiet Suburban Nighttime
Library	33 dB	
Bedroom at Night	26 dB	
	24 dB	Quiet Rural Nighttime
Concert Hall (Background)	22 dB	
Broadcast and Recording Studio	14 dB	
Threshold of Hearing	3 dB	

Values estimated from a chart in [23].

also reckon that the missing word should be a verb rather than a noun or an adjective. The issue of an improved speech intelligibility in contextual speech has been mentioned in chapter 2 of [59], chapter 9 of [61], and many other sources. Indeed, when a word is unintelligible when presented by itself, it might sound intelligible when it is presented in a sentence. This so-called increase in intelligibility does not only apply to English alone, but other languages as well. Considering the fact that much of our communications are contextual, one may doubt the importance of this research since we deal with the intelligibility of single Chinese syllables. However, we must remember that ambiguity also exists in contextual speeches. For example, in the telephone conversation quoted not long before, if I were to say, *"My name is Fong"* on the telephone, the other party might have heard it as *" ... Thong" or " ... Hong"*. Or if I emphasise, *" 'F''o''n''g' Fong"*, the other party might record it as *" 'S''o''n''g' Song"*. In both cases, we know the context surrounds a name, however, the ambiguity is great. An example in Mandarin would be the easily confused numbers 1 /yi1[7]/ and 7 /qi1/[tɕi]. A considerable amount of ambiguity will arise if someone's telephone number is 371-7174. Therefore, speech intelligibility at an individual word or syllable level is also crucial for effective communication and testing it in this level is also worthwhile. The key is, if speech intelligibility is high in the word or syllable level, similarly it should be high, if not, higher in the contextual level.

3.4 The Chinese language

The Chinese languages are the languages of the Han people residing mainly in China, Taiwan, and South East Asia. It belongs to the family of *Sino-Tibetan* languages [30], and are spoken by more than a billion people in the world. There are seven major Chinese language groups or dialects which include *Mandarin*, *Wu dialect*, *Xiang dialect*, *Gan dialect*, *Hakka*, *Yue dialect* or *Cantonese*, and *Min dialect* (chapter 8 of [64]). The Mandarin Chinese ("/Pu3/ /Tong1/ /Hua4[8]/" in China and "/Guo2/ /Yu3/" in Taiwan) are spoken by most of the Chinese popu-

[7] Notation to be discussed later.

[8] These three syllables are an alphabetic representation of Chinese syllables called the Chinese Phonetic Alphabets or Hanyu PinYin. The number behind each syllable denotes the tone associated to that syllable. The mentioned alphabetic representation and tones will be discuss in subsection 3.4.1

lation and it is their common language or official dialect. Our research is based on Mandarin Chinese and for simplification purposes, when the term “Chinese” is used in any subsequent part of this thesis, it will refer to Mandarin Chinese.

3.4.1 Characteristics of the Chinese language

The Chinese language has got its own set of characteristics that differ from English and most European languages. The written form is made up of distinct characters instead of alphabets. All Chinese words are formed by one or more characters (morphemes) and all these characters are monosyllabic. In fact, these monosyllabic characters form a major proportion of all its morphemes. Some examples of the mono-character word are the commonly used /ni3[9]/ (you), /wo3/ (I, me), /shi4/ (yes, is), and /ren2/ (man). Examples of multi-character words are /fei1/-/ji1/ (aeroplane), /dian4/-/shi4/-/ji1/ (television), /zheng4/-/fu3/-/ji1/-/gou4/ (government organisation), and /dian4/-/shi4/-/lian2/-/xu4/-/ju4/ (TV serial). From the fact that a majority of the monosyllabic characters are morphemes, many of the multi-character words are formed by concatenating a series of morphemes (chapter 1 of [64]). Take the example of /fei1/-/ji1/ (aeroplane), /fei1/ in Chinese means fly and /ji1/ means machine. Concatenating them will produce a flying machine that is an aeroplane. Another example is /dian4/-/shi4/-/ji1/ (television), /dian4/ means electricity, /shi4/ means vision or looking at, and /ji1/ means machine. Therefore piecing them together makes an electric visual machine that is a television.

Usually, Chinese characters have only one pronunciation, but there are several cases where one character has more than one pronunciation. Which pronunciation to use depends on the context. However, there are almost always many characters sharing the same pronunciation. Since the written form is not alphabetic, one has to memorise the pronunciation and tone for every Chinese character. Although there are some rules for pronunciation for characters having the same basic strokes, these rules often only lead to either the correct consonant or

[9] Please refer to tables 3.6 and 3.7 for the transcription of Chinese Phonetic Alphabet with International Phonetic Alphabet for the pronunciation of these few Chinese words. Generally, it sounds close (but sometimes not similar) to the English pronunciation of these alphabets with a lexical tone which in this case, the pronunciation of “ni” with tone 3 (tones will be discussed later in the Tones subsection).

Table 3.6: Transcription of Chinese Phonetic Alphabet (CPA) for Chinese consonants with International Phonetic Alphabet (IPA).

CPA	**IPA**	**CPA**	**IPA**
b	[p]	z	[ts]
p	[p‘]	c	[ts‘]
m	[m]	s	[s]
f	[f]	j	[tɕ]
d	[t]	q	[tɕ‘]
t	[t‘]	x	[ɕ]
n	[n]	zh	[tʃ]
l	[l]	ch	[tʃ‘]
g	[k]	sh	[ʃ]
k	[k‘]	r	[ʒ]
h	[x]	(ng)	[ŋ]

Transcription taken from figure 2 in [56].

vowel. Hence, memory and practise are the only reliable methods for recognising Chinese characters. In order to ease pronunciation of Chinese characters, romanisation of the Chinese language was performed as early as the mid 19th century. Some examples of romanisation systems for Chinese include the *Wade-Gile*, *Yale Romanization*, *Gwoyeu Romatzyh*, *TongYong PinYin*, and *Hanyu PinYin* system. The *HanYu Pinyin* or the *Chinese Phonetic Alphabet* (CPA) system was approved by the government of the People’s Republic of China in 1958 and was officially adopted in 1979 [80]. This system, however, is not used in Taiwan. Instead, the Taiwanese used the locally created *TongYong PinYin* system [8][3]. In this thesis, the HanYu Pinyin system is used to represent Chinese words. A transcription of the Chinese Phonetic Alphabet (CPA) with the IPA is given in table 3.6 for Chinese consonants and table 3.7 for vowels. Speech-wise, Chinese is a tonal language and all Chinese syllables have a similar phonetic structure. We shall discuss the phonetic structure and tones of Chinese speech with more detail in the following subsections.

Table 3.7: Transcription of Chinese Phonetic Alphabet (CPA) for Chinese vowels with International Phonetic Alphabet (IPA).

CPA	**IPA**	**CPA**	**IPA**	**CPA**	**IPA**	**CPA**	**IPA**
a	[A]	ai	[ai]	iao	[iau]	uan	[uan]
o	[o]	ei	[ei]	iou	[iou]	uen	[uən]
e	[ɣ]	ao	[ɑu]	ian	[iɛn]	uang	[uaŋ]
ê	[ɛ]	ou	[ou]	in	[in]	ueng	[uəŋ]
i	[i]	an	[an]	iang	[iaŋ]	ong	[uŋ]
-i (front)	[Ì]	en	[ən]	ing	[iŋ]	üe	[yɛ]
-i (back)	[ļ]	ang	[aŋ]	ua	[uʌ]	üan	[yɛn]
-u	[u]	eng	[əŋ]	uo	[uo]	ün	[yn]
ü	[y]	ia	[iʌ]	uai	[uai]	iong	[yŋ]
er	[ər]	ie	[iɛ]	uei	[uei]		

Transcription taken from figure 2 in [56].

Phonetic structure

In Mandarin Chinese, each syllable has a Consonant-Vowel-(Consonant) (CV(C)) structure which consists of an initial consonant (we shall name it C1 for the rest of this thesis), a vowel, and a probable final consonant. The initial consonant (known as *"initial"* both in [54] and [102]) of a Chinese syllable is either one of 21[10] consonants or a null (this is a special case with a vowel as an initial). Unlike English, most (81%) of these consonants in Mandarin Chinese are unvoiced [58]. Plosives like /b/, /d/, and /g/ and some other consonants that are voiced in English are unvoiced when pronounced in Chinese. Please refer to table 3.8 for the 21 consonants and their phonetic classifications.

According to Zhang [102], the later part (V(C)) of a Chinese syllable, which was named a *"final"*, consists of a medial, a kernel vowel, and a coda. There are a total of 10 kernel vowels of which either one must be present in any syllable while the medial and the coda can be optional. The only consonant sounds that will appear in a final of a Chinese syllable are the nasals /n/ and /ng/ and these only happen in the coda. The final will consist of no more than three phonemes and 39 finals can arise from the combination of the three (or less) components in

[10] There are in fact 23 consonants in the written Chinese phonetic alphabets (CPA) two of which are semi-vowels (/w,y/), and they were excluded from the list by both [54] and [102].

Table 3.8: Chinese Consonants and their phonetic classifications.

	Unvoiced			**Voiced**	
	Unaspirated	**Aspirated Fricatives**	**Voiceless Fricatives**	**Voiced**	**Nasals**
Labial	b	p	f		m
Alveolar	d	t		l	n
Velar	g	k	h		
Palatal	j	q	x		
Sibilant	z	c	s		
Retroflex	zh	ch	sh	r	

the finals. Including the initials and finals, there are generally a maximum of four phonemes in every Chinese syllable [103]. All the combinations that can arise from the CV(C) structure are: V, CV, VC, and CVC.

Excluding tones, there are about 408 basic syllables constructed from the CV(C) combination that cover all the phonemic pronunciation of the Chinese language. Although there are only approximately 408 basic syllables, due to the monosyllabic CV(C) structure, the recognition of syllables in hearing might not be a trivial task. The reason being that a consonant is usually the part that bears crucial information that is important to the intelligibility of speech [60][74]. However, since the average power of consonants is less than 20 times below that of the vowels (section 3.3.2), they are highly susceptible to noise and masking effects. Furthermore, it is sometimes not easy to even distinguish consonants under normal hearing conditions. Hence consonants can easily be confused. In Chinese speech, 39 confusing sets of vocabulary arise from the 39 finals which might impair the intelligibility of the syllables. An example of a confusing set is the /a/ set which includes: [a], [ba], [pa], [ma], [fa], [da], [ta], [na], [la], [ga], [ka], [ha], [zha], [cha], [sha], [za], [ca], [sa]. Within this confusing set, there are pairs of rhyming syllables that are more prone to confusion within the pair. Li *et al.* listed six phonetically rhyming pairs for Chinese speech in [56]. The pairs are Airflow-No Airflow, Nasal-Oral, Sustained-Interrupted, Sibilated-Unsibilated, Grave-Acute, and Compact-Diffuse. Syllables within these pairs sounded phonetically very close. In this case, the intelligibility of the consonants (we shall name it consonantal

intelligibility) can be easily confused.

Tones

One can still not master correct Chinese pronunciation by simply learning the phonics, because each Chinese syllable carries a tone. Every Chinese syllable is thus defined by both its constituent phonemes and a single tone. Two syllables sharing identical phonemes will have different meanings if the tones associated to the phonemes are different. There are a total of four lexical tones and one neutral tone [100]. The primary difference of the five mentioned tones is in their pitch contours that alter the fundamental frequency, f0, against time. While the four lexical tones have specific patterns in their contours, the neutral tone has not. Tone 1 is a high-level tone, tone 2 is mid-rising, tone 3 is mid-falling-rising, tone 4 is high-falling (please refer to figure 3.3), and the neutral tone depends on the tone of its previous syllable. An example of a basic syllable with tones is /ma1/ (mother), /ma2/ (numbness), /ma3/ (horse), /ma4/ (scold), and /ma/[11] (The second syllable for mother[12]). Since most of the Chinese consonants are unvoiced, the tonal elements are carried in the vowels. Therefore the tone of a Chinese syllable can be determined by extracting pitch information from its vowel [101][24].

From the 408 basic syllables, about 1345 syllables can be constructed when the five tones are included into the basic syllables and since each pronunciation may be associated with many characters, the corpus of Chinese characters is enormously large. However, only about 3,000 to 4,000 characters are commonly used by an ordinary Chinese literate (chapter 3 of [64]).

3.5 Conclusions for the chapter regarding speech

The production of speech was briefly covered in this chapter. Speech sounds are produced by two essential and one optional functional processes namely, *initiation*, *articulation*, and/or *phonation*. We also discussed the various types of initiation and articulation and explained what phonation means.

[11] There is no tone number associated with the neutral tone.

[12] In Chinese, people usually use the combination of /ma1/ /ma/ to address mother. There are also occasions where only the single syllable /ma1/ is used

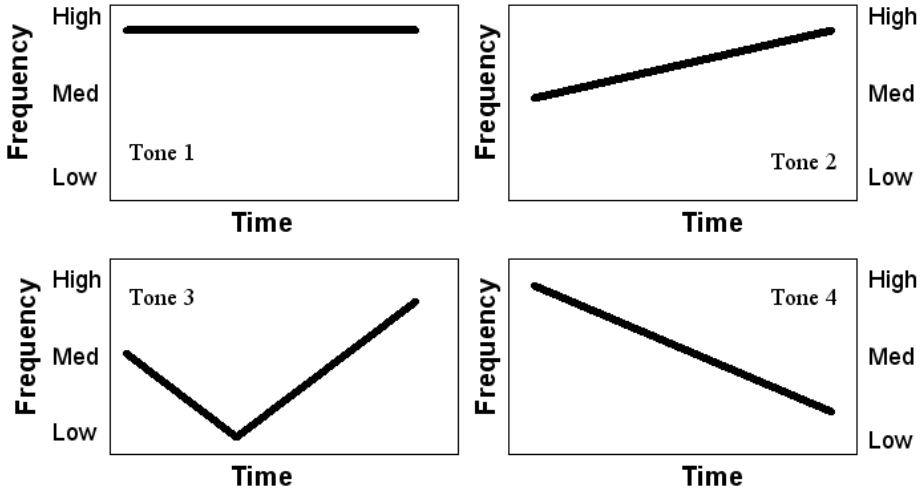

Figure 3.3: Pitch contours of the four Chinese lexical tones.

The characteristics of speech were also introduced. As not all languages use a Latin alphabet to represent its phonemes, the International Phonetic Alphabets is defined to denote most, if not, all speech sounds a human could possibly produce. The production and characteristics of consonants and vowels were dealt with followed by the frequency and loudness of intelligible speech. The influence of speech context to intelligibility was also briefly mentioned.

Lastly, an overview of the Chinese language plus its unique characteristics were discussed as an important context for our research.

Chapter IV

Speech Quality and Intelligibility Measurements

4.1 Introduction

Due to the probability of information loss in speech transmission networks or speech processing systems through transmission error, speech coding loss, bandwidth limitations, and so on, the quality and intelligibility of a piece of processed speech may well be degraded through the process. This degradation may be undesirable at times when specific properties of that piece of speech are required, for example the loss of intelligibility in a telecommunication system where speech intelligibility is essential to the users. Therefore, it is often desirable to test the quality and/or intelligibility of such systems in such a way as to provide a benchmark for their performance. Of course, before one can determine a measure of quality and intelligibility degradation, it is appropriate to first define the meaning of *Speech Quality*, and *Speech Intelligibility*. The definitions of the root words *Speech*, *Quality*, and *Intelligibility* are as followed:

- **Speech [noun]:** The ability to talk, the activity of talking, or a piece of spoken language [7]

- **Quality [noun]:** The degree of goodness or worth [4]

- **Intelligibility [noun]:** From intelligible [adjective] (of speech and writing), clear enough to be understood [7]

From the above definitions, the complex word or phrase can be determined:

- **Speech Quality:** The degree of goodness in the perception of speech

- **Speech Intelligibility:** How well or clearly one can understand what is being said

Steeneken in [76] defines them more technically as:

- **Speech Quality:** Quality of a reproduced speech signal with respect to the amount of audible distortions

- **Speech Intelligibility:** The amount of speech items that are recognised correctly

Here we realise that speech quality and speech intelligibility are different attributes in relation to the perception of speech. Though they are differing attributes, they are not totally exclusive of one another as there exists some form of relationship between them (please refer to section 5.2). However, in the measurement of these two attributes, it is generally recognised that different measurement approaches must be used to test them individually. These approaches can be divided into two categories for both attributes: *subjective* tests and the *objective* tests. Subjective tests involve a group of human listeners to rate either of the two attributes while objective tests involve some computerised mathematical calculations of the physical parameters of speech signals to determine them. In the next few sections, both subjective and objective speech quality and speech intelligibility measurement approaches will be briefly considered.

4.2 Subjective tests

Subjective tests or listening tests involve a pool of human subjects to rate or provide opinions on either attribute. Depending on the objective or the test, the minimum number of human subjects used differs. Generally, the higher the number of human subjects used, the higher the confidence in the test outcome. Since human subjects are used, tests are performed in real-time (no simulation or time warping is done computationally). Subjects have to listen to all test speech in order to provide opinions or rate the system being tested. This category of tests can be considered more accurate than machine-judged tests since humans can easily and repeatedly perceive quality or intelligibility, using complex auditory processes

which are still not fully understood, or able to be replicated by machine. Well known subjective intelligibility tests include the *Diagnostic Rhyme Test* (DRT) [87], *Modified Rhyme Test* (MRT) [12], and *Phonetically Balanced Word Lists* (PB) [11] [27], and subjective quality tests include the *Diagnostic Acceptability Measure* (DAM) [86], *Mean Opinion Score* (MOS) [42], and *Degradation Mean Opinion Score* (DMOS) [42][28].

4.2.1 Subjective intelligibility tests

We have defined the term "Speech Intelligibility" as 'how well or clearly one can understand what is being said'. In other words, it is the degree of recognition of a piece of spoken speech. It was previously mentioned that a basic monosyllabic piece of speech is made up of phonemes and a complex one consists of a string of spoken words. Hence in subjective intelligibility tests, the materials used in the test may be at the level of basic phonemic units, words (meaningful or nonsense), or even sentences. When nonsense words are used, they usually consist of a combination of consonant-vowel-consonant (CVC) (similar to the structure of a Chinese syllable). They may also exist in the form of VCV, VC, CV, CVCC, or CCVC. The phonemes are selected so that a specific range of vocal attributes can be tested [77]. The test material can be presented to the subjects in different forms, for example, an individual word or a sentence may be played, or the word to be tested might be embedded in a carrier phrase [10]. There also exist various methods in which the subjects respond to the tests in an interactive fashion. This can be an open or closed response. In the open response situation, the subjects are required to give responses as to what messages or phonemes they actually perceived in the listening test, while in the closed response situation, subjects are only required to make a selection of what they have heard, usually from a list of candidate sounds.

Examples of subjective intelligibility tests at phoneme or word level are the rhyme tests like the *Diagnostic Rhyme Test* (DRT) and the *Modified Rhyme Test* (MRT). These tests require a closed response from the subjects where they would have to choose the word that is played from a list of two (DRT), or six (MRT) rhyming words presented to them on the display. The initial consonants are being tested in DRT while both consonants and vowels are used in MRT. The advantage

of such tests is that the procedure is simple and subjects used can be untrained or "naive". This type of testing is usually used for systems where the basic level of intelligibility is not very high.

In the case of the open response test, subjects have to state what they hear and nonsense words are usually used in such tests [29]. Different combinations of consonants and vowels are used depending on the language or the particular diagnostic information required for the system under test. Sometimes, words used are embedded in a carrier phrase. This is to take the effects of echoes and reverberation into consideration as such effects will occur in the carrier phrases. Subjects participating in these tests must be thoroughly trained. This type of test is advantageous in testing high-end systems where the basic level of intelligibility is high. An example of an open response monosyllabic word test is recorded in [29].

At sentence level, subjects are required to give a rating to the overall intelligibility of the entire sentence as in the *Mean Opinion Score Test* (MOS) where subjects are asked to rate the intelligibility of the sentence according to a five-point scale (bad, poor, fair, good, and excellent), or to give an estimation in percentage (0% to 100%) of the number of intelligible words in the sentences. One example of a sentence level test is the *Speech Reception Threshold* (SRT) [66]. In SRT, the subjects will listen to sentences masked by noise. When a subject recognises a sentence, the noise level of the next sentence will be increased by 2 dB. This increase in noise level proceeds until the subject cannot recognise the sentence where the noise level will decrease by 2 dB at this point. This procedure will continue until 50% of the sentences are correctly recognised. The advantages of this test are that untrained subjects can be used, and results can be easily reproduced while the disadvange is that accuracy of this test may be affected by training effects and fatigue due to the significant length of test.

Since our research is primarily concerned with Chinese speech, we shall use the subjective tests specially designed for the testing of systems processing Chinese speech. The proposed *Chinese Diagnostic Rhyme Test* (CDRT) [56] and its extension *CDRT-Tone Test* [26] designed by Li *et al.* and Ding *et al.* to evaluate the intelligibility of Chinese speech processed through sound processing systems are used in this research. As the CDRT was developed based on the principles of DRT, we shall briefly introduce the DRT, CDRT, and CDRT-Tone tests in the

following subsections.

Diagnostic rhyme test

The *Diagnostic Rhyme Test (DRT)* [87][89] developed by William D. Voiers uses a corpus of 192 words in 96 rhyming pairs. Each rhyming pair defers from its counterpart in only one aspect, the initial consonant. The DRT test only tests consonants because the consonants are more important in the intelligibility of words and are more easily confused than vowels [60][74]. They are also more susceptible to masking effects. In the DRT, six elementary phonetic attributes of the English consonants are tested. The attributes are: voicing, nasality, sustention, sibilation, graveness, and compactness.

- **Voicing:** To test whether the consonant in a pair with the same oral articulation is voiced or not. Examples of consonants in this pair are /v/-/f/, /z/-/s/, and /g/-/c/.

- **Nasality:** To test whether the consonant contains a nasal component or is purely oral. Half of the pairs in this category involve a grave phoneme pair, e.g. /m/-/b/, and half an acute pair, e.g. /n/-/d/.

- **Sustention:** To test whether the consonant is sustained or interrupted. Half of the pairs in this category are a voiced phoneme pair and half unvoiced.

- **Sibilation:** To test whether the consonant is sibilated or not. A sibilated consonant contains high frequency components with significant energy level e.g. /s/, /z/. Half of the pairs in this category are voiced phoneme pairs and half unvoiced.

- **Graveness:** To test whether the consonant is grave or acute. A grave consonant contains a high proportion of low frequency components. Part of the pairs in this category are voiced, unvoiced, sustained, and interrupted.

- **Compactness:** To test whether the consonant is compact or diffused. A compact consonant is articulated behind the alveolar region of the mouth. Some examples are /j/, /k/, /g/, /h/.

During the DRT test, one word of a pair is audibly reproduced, and the pair of words displayed on a computer monitor. Subjects will choose one of the two displayed words according to their perception of which one was heard. All the original and processed[1] 192 words will be played to the subjects at least once. The diagnostic information of the system to be tested can be determined by the DRT test when the results of the test are categorised. It can be realised which specific acoustic attribute was improperly processed or mis-transmitted in that system. An overall score can also be computed if the overall performance of the tested system is required. This score is useful in comparing the overall performance of different systems. The equation for this overall score is:

$$S = \frac{100(R-W)}{T}$$

where S is the "true" percent-correct responses, R is the observed number of correct responses, W is the observed number of incorrect responses, and T is the total number of items involved.

The DRT test is internationally recognised and is very widely used around the world especially in the evaluation of speech coders. It is also useful in comparing different systems in terms of overall performance, or specific phonetic attributes. The test is simple to administer and is easily reproducible.

Chinese diagnostic rhyme test

Adopting the philosophy and methodology of the DRT which has various advantages and is very popular, the *Chinese Diagnostic Rhyme Test* (CDRT) was proposed to evaluate the intelligibility of Chinese speech transmitted through communication systems. It is effectively the DRT applied to Chinese. It uses a corpus of 192 words in 96 rhyming pairs. From this 96 rhyming pairs, six elementary phonemic attributes are tested. They are *airflow*, *nasality*, *sustention*, *sibilation*, *graveness*, and *compactness*. The elementary phonemic attributes are identical to that of the DRT's except for the attribute of *voicing*. Since most Chinese consonants are unvoiced, this attribute are not directly applicable in this case. Therefore

[1] Speech files that have been processed by the sound system or coded and decoded by the speech coder being tested.

the attribute of *Airflow* is tested instead. Chinese consonants of the airflow-no airflow pair include /p/-/b/, /t/-/d/, /q/-/j/, /c/-/z/, and /ch/-/zh/. For nasality, the pairs /m/-/b/ and /n/-/l/ are used because a considerable fraction of Chinese speakers tend to confuse the pronunciation of /n/ and /l/. The CDRT test procedure is similar to that of DRT in which a Chinese syllable is played to each subject while the CDRT pair in which the played syllable exists is displayed on the monitor. The subject is required to make a closed response decision by selecting whichever of the two Chinese syllables displayed matches what he/she heard. The corpus of Chinese characters in the CDRT is given in [56].

By obtaining results on which attribute fails, a system's flaws can be easily identified. Since this process is very similar to DRT, the CDRT inherits many of the advantages of the DRT. However, although the DRT is rather extensive in testing important attributes of English speech, the CDRT does not test all the characteristics of Chinese speech because Chinese, differing from English, is a tonal language. Since CDRT only discriminates consonants, vowels and tones are not tested. Hence one cannot form a concrete conclusion concerning the intelligibility of Chinese speech in a particular system solely based on CDRT results.

CDRT-tone

From section 3.4.1, in the Chinese language, most syllables can be pronounced with one of five different tones such that pronouncing a syllable with a different tone imparts different, and usually totally unrelated, meanings. By testing a system using phonemic measures alone (CDRT) cannot conclusively determine whether that system reliably processes Chinese speech. Therefore, as an extension to CDRT, the CDRT-Tone test was proposed by Ding *et al.* [26] to test the tonal intelligibility of Chinese syllables. It consists of 40 pairs of Chinese syllables divided into four categories according to the similarity of pitch height and contour of the four lexical tones[2]. The categories are: *(tone 1-tone 2), (tone 1-tone 3), (tone 2-tone 3),* and *(tone 3-tone 4).* Categories like *(tone 1-tone 4)* and *(tone 2-tone 4)* are omitted because their pitch heights and contours are significantly different. With the addition of the CDRT-Tone test, the intelligibility of Chinese

[2] Since the fifth tone a light tone and not a lexicon by itself, it is not included in the CDRT-Tone test (please refer to section 3.4.1).

speech transmitted through a particular system can be more confidently concluded compared to use of CDRT alone. The 40 pairs of Chinese characters are given in [26]. Testing procedures in CDRT-Tone follow identical methodology to that of DRT and CDRT.

4.2.2 Subjective quality tests

Differing from speech intelligibility, speech quality is the overall impression of a piece of speech which has been considered by some researchers as a unidimensional phenomenon [67], while others as multidimensional to include dimensions of clarity (intelligibility), fullness, brightness, softness, spaciousness, nearness, extraneous sounds, and loudness[3] [35]. Hence a subjective speech quality test is usually a preference measurement that evaluates the output of the sound system being tested as a unidimensional entity, or in terms of the listed dimensions. It could also be an evaluation of the amount of degradation in speech quality caused by the system. The most commonly used types of subjective quality measurement are the *Absolute Category Rating (ACR)* and the *Degradation Category Rating (DCR)* methods. In the ACR tests, subjects are required to give an immediate rating for the quality of a piece of speech from the system being tested. No reference speeches are given to which the subjects can take reference from for the quality of speech they rate. Ratings are given on a five point scale called the *Mean Opinion Score* (MOS) previously mentioned in section 4.2.1. A variation on the ACR test is the DCR test. In the DCR test, a segment of reference speech is presented to the subjects to anchor the result of the judgement given to the piece of speech from the system being tested. By doing so, the amount of degradation with respect to the original speech signal can be determined. Therefore, the five point scale from the MOS is used here as a *Degradation Mean Opinion Score* which rates the amount of degradation of the processed speech. The DMOS scores range from 5 to 1 (5 - Degradation is inaudible, 4 - Degradation is slightly audible but not annoying, 3 - Degradation is slightly annoying, 2 - Degradation is annoying, and 1 - Degradation is very annoying). The advantage of the DCR emerges over the ACR when the system being tested is of a high quality where quality differences between output speech is small. The DCR is also useful in the case where the quality of the

[3] Different researchers might give varying terms and definitions to these dimensions.

original speech is not very good. Both the ACR and DCR methods do not require trained speakers. A larger pool of subjects is therefore desirable for a stronger confidence in the experiments. The test material used in ACR and DCR consists of simple, meaningful, short sentences which can be easily understood. No obvious connections in the meaning of consecutive sentences is allowed. Details for conducting the ACR and DCR are given in [42].

Another well known subjective quality test is the *Diagnostic Acceptability Measure* (DAM) [86]. It allows subjects to rate overall quality as well as to identify separate quality dimensions within the test speech material. It requires the subjects to give separate ratings for the speech signal itself, the background effect, and total effect. A total of 21 ratings are given for a speech file where ten ratings concern the quality of the speech signal itself, eight concerning the quality of the background, and three concerning intelligibility, pleasantness, and overall acceptability. These 21 ratings provide an extensive evaluation of the system being tested. Trained subjects are required due to the complexity of this test.

4.3 Objective tests

It is a well-published fact that the most accurate types of measurement for speech quality and intelligibility are subjective tests [92][84]. However, it requires considerable effort and time in order to form a conclusive opinion using subjective tests. A significant number of human subjects is necessary with each spending upwards of twenty minutes performing listening tests for each system they test. A substantial cost is therefore also required to compensate the subjects time, and this sum becomes greater when trained subjects are required. On top of that, the same significant number of native speakers for each language under test is also required. Therefore, automated objective measurements have been developed to ease this situation allowing a better time and cost effective evaluation solution for sound systems. However, an objective speech quality or intelligibility measurement system cannot replace the human brain in the perception of speech and it does not directly measure the speech attributes. These objective systems merely predict the speech attributes based upon physical parameters of the speech signal itself, and perform mathematical calculations to obtain an estimated quantity. Hence these systems can only provide an estimate of the perception of humans

and it is not easy (if not, impossible) to achieve a 100% accurate estimation. In the following subsections, various objective measurement concepts will be briefly introduced for measuring both speech quality and intelligibility.

4.3.1 Objective intelligibility tests

French and Steinberg suggested in 1947 [34] that the intelligibility of sounds can be computed from the intensities of the speech signal plus other components such as distortion and noise as a function of frequency. They mentioned that the various sounds in a speech signal differ from each other in their build-up, decay, length, total intensity, and distribution of intensity with frequency, and that in speech, it is a continuous variation of intensities across the whole frequency range. Thus, they proposed that speech signals be divided into different distinct frequency bands and that the intensity of each band carries a contribution to a whole quantity known as the *Articulation Index* (AI). The level of intelligibility can therefore be estimated from the value of the articulation index. Their work was later validated and adopted by various researchers in their search for measurements of speech intelligibility [51][46]. Their research hence opened the way for objective intelligibility measurement systems that evolved throughout history. Examples of well known objective speech intelligibility measurement systems are the *Speech Transmission Index* (STI) [77] and the *Speech Intelligibility Index* (SII) [13].

The Speech Transmission Index (STI) works on the basis of generating an artificial test signal that replaces the original speech signal. The artificial signal consists of seven separate octave band signals where six of the seven octave band signals made up an artificial speech signal while the seventh is a test signal. The signal-to-noise ratio of these seven octave bands contributes to the value of the STI.

In a piece of speech, a sequence of phonemes is present and each phoneme can be represented by a specific frequency spectrum. For a piece of speech to be intelligible, the differences in consecutive spectrals must be preserved. These differences in the spectrum are related to fluctuations in the envelope functions that are determined from the sequence of phonemes. When the speech signal is affected by noise or other distortions, the differences in the spectrals are reduced which also leads to a reduction of the fluctuations in the envelope functions. The

fluctuations in the envelope function are depicted in an envelope spectrum. The reduction of the fluctuations is reflected by a reduction factor known as the *Modulation Transfer Function* (MTF) which is calculated from the difference between the envelope spectrum of an octave band in the original signal and that of a modulated signal. The MTF shows the reduction factor of the modulation index as a function of modulation frequency. A total of 14 modulated frequency signals that varies within the range of the fluctuations in speech (0.63 to 12.5 Hz) are used to determine the MTFs of the seven octave bands. This means to say that a 14 X 7 matrix consisting of MTFs of each octave band and modulated signal is calculated. A weighted summation of the MTFs from the seven octave bands is then calculated to obtained the STI value which ranges from 0 to 1 [76].

The *Speech Intelligibility Index (SII)*, which is a later version of AI, works in a manner rather similar to the STI. However, SII also takes into account the physical properties of the system being tested when it estimates speech intelligibility.

4.3.2 Objective quality tests

A number of objective speech quality measurement systems have been developed to ease the human effort required to assess the output speech quality of sound systems or speech coders. Although they might not perform as accurately as the subjective methods, they do provide a good indication of speech quality of the systems, and this is especially helpful in the development process and in mutual comparison between systems. The aim of a good objective speech quality measurement system (OSQM) is to achieve a high correlation with subjective quality ratings. The objective measurement systems can be classified into three categories namely the *Time Domain Measures*, *Frequency Domain Measures*, and the *Perceptual Domain Measures* [93].

Time domain measures such as Signal-to-Noise Ratio (SNR) and segmental SNR (SNRSEG) [48] are widely used when the coding of speech signals is based upon waveform shape matching. These methods deem distortion to be a physical entity and hence provide an estimation on the speech quality in terms of the quantity of this entity (noise) which is present. It is achieved by calculating the mean squared error between the original and coded speech waveforms. However, when speech coding technology advanced to the point where waveform coding

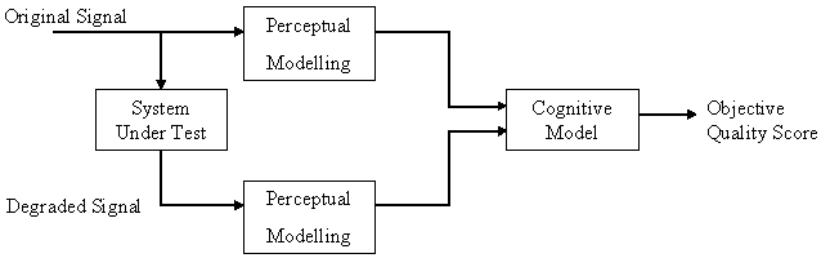

Figure 4.1: General structure of a perceptual domain based OSQM.

was no longer an optimum means to code speech, and far more complex speech handling methods arose, these simplistic quality methods were no longer capable of providing a reasonable estimate of quality [48]. This is especially true when bit-rates of speech coders can reduce to 4 kbps or less (compare with waveform coding which can only work effectively above about 16 kbps) [95].

The second category of OSQMs are the frequency domain measures. These tend to perform better than time domain methods in cases where time alignment errors or phase shifts occur. However, this class of measures calculates speech quality by predicting the failure in the speech production models used in the codecs, and hence their performance is limited by the constrains of the models [84]. Linear Predictive Coding based measures are examples of this type, as is the *Cepstral Distance Measure* (CD) [49] which is regarded as one of the best performers.

The third category is the perceptual domain measures where speech quality is predicted using human auditory perceptual models. The general structure of a perceptual domain measure is given in fig 4.1. Since the auditory pattern of the human ear is very similar across a range of people, it may be reliable to adopt such a model for quality measurement. The aim of such models is to mimic several key features of the human auditory system. The key features are:

- **Frequency Based Perception:** The human ear perceives sound as a collection of frequencies similar to a Fourier analyser

- **Frequency Warping:** Frequency in Hertz is warped into the critical band scale (Bark) due to the non-linearity of the distance in the basilar Membrane to the frequency ranges

- **Loudness Perception:** Different loudness perception for sounds with same intensity but different frequency

This method is by far the most effective and reliable to objectively estimate speech quality, and much research was done to investigate and develop different auditory models. Examples of this type of speech quality measurement systems include the *Bark Spectral Distortion* (BSD) [93], *Perceptual Speech Quality Measure* (PSQM) [43], *Measuring Normalizing Block* (MNB) [90][91], *Deustche Telekom Speech Quality Estimation* (DT-SQE) [32], *PSQM+* [15], *Modified Bark Spectral Distortion* (MBSD) [98][97][99], *Perceptual Analysis Measurement System* (PAMS) [73][69], *Enhanced Modified Bark Spectral Distortion* (EMBSD) [96], *Perceptual Evaluation of Speech Quality* (PESQ) [45][71][16], etc. Each of these systems has its own advantages and disadvantages which are worthy of consideration when used to assess speech quality objectively. Currently, PESQ is considered one of the most advanced system being used around the world [79][22]. Issues or problems that arise in modern systems like packet loss, variable delay, and speech codecs, are being considered in PESQ. Our research will be based on the application of PESQ and MNB so that a contrast can be obtained between the more advanced system and the an older one. A brief description of both systems is given in the following subsections.

Perceptual evaluation of speech quality (PESQ)

As technology advances, new speech processing systems and speech codecs are continually being developed, and these give rise to newer issues that will affect the measurement of speech quality objectively. Older systems like the BSD, PSQM, and MNB will no longer meet up with the requirements of the present day OSQMs as they do not account for the conditions of current speech processing systems such as lower bit rate speech codecs, packetised audio, variable delays, etc. Adopting and integrating concept from PSQM+ and PAMS, PESQ was developed to appease such issues. This new system became the new ITU-T (International Telecommunications Union) recommendation as a method to objectively assess or measure speech quality. This new ITU-T P.862 recommendation replaces its former P.861 which defines the PSQM. Stated in the ITU-T P.862 recommendation, the PESQ achieves a correlation of 0.935 with subjective scores.

The goal of PESQ is to mimic the perception of speech in real life using a psychoacoustic model. PESQ is an intrusive objective speech quality measurement method that compares the coded (degraded) signal from the codec or network with the original (reference) signal using a computer. The physical signals that are input to the computer are transformed into internal representations to be mapped onto psychophysical representations. These psychophysical representations closely resemble the auditory perception of a human in terms of perceptual frequency measured in Barks and loudness measured in sones (these units were explained in sections 2.3.2 and 2.3.1). The structure of PESQ model is given in figure 4.2. The steps taken to achieve this are [45][72]:

- Level alignment,
- time alignment,
- time-frequency mapping,
- frequency warping, and
- loudness mapping.

Level Alignment: The gain of systems being tested differs between systems and this information is not input into PESQ for calculation. Therefore the signals are scaled so that the effects caused by the system gain can be compensated and both the original signal and the coded signal can be normalised to a similar level for processing. Scaling is done by calculating the different gains and applying them to both signals.

Time Alignment: When the original signal is passed through a sound processing system, there is a time lag between the original signal and the coded signal. If the original signal is compared directly to the coded signal simultaneously, the objective measurement system may not generate an accurate result since different parts of the messages are being compared. Therefore, time alignment is required between the original and coded signals to ensure that the corresponding parts of both can be compared.

Time-Frequency Mapping: Since the human ear acts like a Fourier analyser in a sense that it perceives sound as a collection of frequencies, the equivalent psychoacoustic model used in objective measurement systems also works with frequencies. Hence a time-frequency transformation is performed. This is achieved by performing a short-term Fourier Transform (STFT) with a Hann Window with a size of 32 ms (a frame length of 256 samples for 8kHz sampling or 512 samples for 16kHz sampling) [45]. A 50% overlap between successive windows (frames) are used in this STFT.

Frequency Warping: Different frequencies produce their maximum effects at different locations along the basilar membrane in the human ear, so that each location responds only to a limited range of frequencies. The effective frequency range to which a given location responds is its critical band (chapter 3 of [61] and chapter 6 of [104]). Hence in the objective measurement system, the frequency scale in Hertz (Hz) is warped and mapped onto the critical band rate scale. This produces a pitch power density representation within each STFT frame. These power representations are then summed up and normalised.

Loudness Mapping: Due to the phenomena that the human ear perceive sounds of different frequencies having similar intensities as different level of loudness (section 2.3.1), the intensity axis should be warped to the loudness scale based on the absolute hearing threshold. In order to obtain an accurate measure, the psychoacoustic loudness scale must be calibrated according to the loudness level in *phons* which gives equal loudness throughout the range of audible frequencies instead of sound pressure levels (chapter 8 of [104], and [93]). The calibrations are performed with a reference of a 1000Hz pure sine wave at a level of 40 dB SPL[4] to give a loudness value of 1 sone. An increase of 1 sone represents the doubling of the loudness sensation and is equivalent to a increase of 10 phons.

After mapping the original and coded signal onto the psychophysical domain, the audible error between the two signals is calculated and aggregated into disturbance values over time and frequency. A quality score is then calculated by

[4] Since dB represents an intensity or power ratio, it is not an absolute intensity. To specify the absolute intensity of a sound, we need to specify it N dB above or below a certain reference level. A sound level specified using this reference level is referred to as a Sound Pressure Level (SPL). For example, a sound at 30 dB SPL is 30 dB higher in level than the reference level of 0 dB (chapter 1 of [61]).

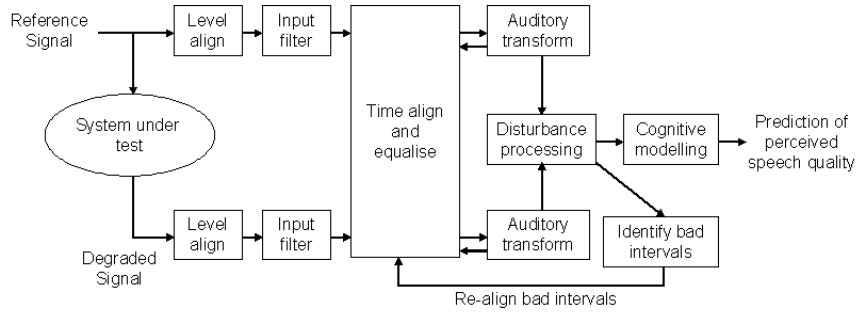

Figure 4.2: General structure of PESQ. Redrawn from [72].

substracting the disturbance values from the total score of 4.5. Thus, the quality scores range from -0.5 to 4.5, where -0.5 indicates very poor quality and 4.5 indicates perfect quality. If the output of a transmission system or codec is exactly similar to the input, i.e. no degradation is detected, this will yield a result of 4.5. Normally, the quality of speech files will range from 1 to 4.5. Values seldom fall below 1, except in cases where quality degradation is extreme.

Measuring normalizing block (MNB)

The MNB algorithm was developed by Stephen Voran in 1997 and included into the appendix of the ITU-T P.861 recommendation. This is another intrusive method that uses a hierarchy of measuring normalizing[5] blocks that model a perceptual transformation and a distance measure to determine speech quality. The MNB first transforms the original and degraded signal into the perceptual domain. The transformation into the perceptual domain is quite similar to the concept in PESQ which does a time-frequency mapping, frequency warping, and loudness mapping. After this process, a distance measure that measures the auditory distance between the two perceptually transformed signals is done by using a hierarchy of MNBs. There are two methods of MNB: the *Time MNB* (TMNB) and the *Frequency MNB* (FMNB) where spectral deviations are measured at multiple time and frequency scales. The TMNB integrates over some frequency scale, and at multiple times

[5] The MNB was developed in the United States of America and hence the spelling used to name this method, particular the middle word “Normalizing”, is in American English. Henceforth when this name or term is used in any part of this thesis, the American spelling will be used to preserve the originality of its name

measures differences and normalises them. Then, the positive and negative measurements are integrated over time. The FMNB works in a reciprocal approach: It integrates over some time scale, and then at multiple frequencies, measures differences and normalises them. Similarly to TMNB, the measurements are integrated over frequency here. In this way of working from larger time or frequency scales down to smaller one, a human's auditory patterns of adaptation and reaction to spectral differences can be emulated. Hence after a series of MNBs (two structures were proposed: structure one consists of 12 MNBs and structure two of 11 MNBs), a full set of linearly independent measurements can be formed and linear combinations of these measurements are used to determine the auditory distance (AD) which is an estimate of the perceptual distance between the original and degraded signal (in other words, the perceived speech quality):

$$AD = \sum_{i=1}^{12(or11)} w_i.m(i)$$

where w_i is the weight and m is the measurement of each block. The value of AD begins at 0 when the original and degraded signal are identical, and increases when the perceptual distance of the signals is greater (lower quality). To map the AD values into a finite range to obtain a better correlation with the MOS or DMOS scores, the logistic function which ranges from 0 to 1 is used:

$$L(AD) = \frac{1}{1+e^{a.AD+b}}$$

When $a > 0$, $L(AD)$ is a decreasing function of AD.

4.4 Conclusions for the chapter regarding speech quality and intelligibility measurements

In the beginning of this chapter, we defined the terms speech quality and speech intelligibility. To measure these two speech attributes, there are two main categories of testing methods: subjective and objective measurements. Various subjective and objective speech quality and intelligibility measuring tests or systems were also discussed in this chapter. A more detailed introduction was given for the tests

or systems involved in our research. They were the subjective *Chinese Diagnostic Rhyme Test* (CDRT) and *CDRT-Tone* intelligibility tests, and the objective *Perceptual Evaluation of Speech Quality* (PESQ) and *Measuring Normalizing Blocks* (MNB) quality measurement systems. The end of this chapter thereby concludes our background or contextual information portion of this thesis.

The next two chapters report our main findings. They begin with the evaluation of the OSQMs for processed Chinese speech in the proceeding chapter.

Chapter V

Evaluation of Existing Objective Speech Quality Measurement Systems

5.1 Introduction

Chapter 3 presented two characteristics of the Chinese language that differ from English and most European languages. They are the CV(C) phonetic structure and the use of tone, both of which are closely related to the intelligibility of Chinese words or syllables. Arising from the CV(C) phonetic structure are 39 confusing sets of Chinese sounds (section 3.4.1), which impair the recognition or intelligibility of Chinese syllables. We shall name this *Consonantal Intelligibility*. There are also up to four lexical tones that are associated with any Chinese syllable to give it a distinct meaning. We shall call the recognition or intelligibility of Chinese words through tones the *Tonal Intelligibility*. Our approach in this research is first to investigate the relationship between speech quality and speech intelligibility. If there is indeed a relationship between these two speech attributes, we shall define and establish this relationship. This relationship shall be used as a basis to evaluate an objective speech quality measurement system (OSQM) to determine whether the two mentioned characteristics of Chinese speech are taken into consideration in the measurement of Chinese speech quality.

In this research, two OSQMs will be tested: The *Perceptual Evaluation of Speech Quality* (PESQ)[1] described in ITU-T recommendation P.862 [45] which is a more recent version based on a *cognitive perceptual* model, and the *Measuring Normalizing Block* (MNB)[2] method [90], included into the appendix (II) of the

[1] The ANSI-C reference implementation of PESQ used for evaluation purposes in our research was obtained from ANNEX A of ITU-T recommendation P.862.

[2] Software implementation of MNB is downloaded from http://www.icir.org/hodson/mnb/. Usage of this algorithm for this research is with permission from Stephen Voran, the author of the

ITU-T recommendation P.861 [43][3].

5.2 Relationship between speech quality and intelligibility

Since we are dealing with an issue regarding the correspondence between Chinese speech intelligibility and quality, one important question to ask is, *"Should speech intelligibility be considered when OSQMs assess the quality of a piece of speech?"* One may claim that a piece of totally unintelligible speech, for example speech spoken in a foreign language, can be of high quality if the fidelity of it is excellent, hence speech intelligibility should not affect quality. However, it must first be realised that both speech quality and intelligibility mentioned in our scope of study are the outputs from a speech processing system and are assessed with respect to the original speech at the input of the system. Hence what we are interested in is the quality or intelligibility of the processed speech affected by the changes (degradations) made to the original speech. Looking back into our definitions in section 4.1 (where speech intelligibility is defined as how well or clearly one can understand what is being said, or the amount of speech items that are recognised correctly, and speech quality is the degree of goodness in the perception of speech, or quality of a reproduced speech signal with respect to the amount of audible distortions), we assume that speech intelligibility has a narrower scope of just the recognition or understanding of speech while speech quality encompasses a broader scope which we assume includes intelligibility. If this is so, the degradation or distortion that causes the loss of intelligibility would normally also cause a decline in quality but a loss in quality does not necessarily result in an intelligibility loss.

There are several items of evidence to prove this point:

1. When searching for a satisfactory method to evaluate the quality of processed speech, Voiers in [86] stated that, *"It is a matter of common observation that user acceptance of voice communications equipment depends on*

algorithms.

[3] Although ITU-T P.861 has been made obsolete and replaced by P.862, MNB was added for informative reason in appendix II of [43]. Hence it was not totally binding to the standards described in P.861. Furthermore, we could also use it in our research to give a more detailed block-by-block analysis to determine a contrast between the updated and outdated objective speech assessment systems

factors other than speech intelligibility, intelligibility being a necessary but not sufficient condition of acceptability.". He realised that assessing speech quality based on ratings or scores for speech intelligibility is insufficient to determine quality. Therefore, he proposed the *Diagnostic Acceptability Measure* (DAM) that combines an isometric (direct) and a parametric (indirect) approach to determine speech quality. Speech quality is regarded as an overall acceptability entity in the isometric approach whereas in the parametric approach, speech quality is viewed as a multidimensional entity that includes the quality perception of the speech signal itself, background effects, and total effects. At that time, a total of 20 individual ratings were given to these three attributes[4] and intelligibility was included as one of the three items listed in the total effect attribute. When proving the validity of DAM, a high (but curvilinear) correlation was obtained between the isometric acceptability (quality) rating and the intelligibility rating in which when the level of intelligibility increases, quality also increases (with the points at the centre slightly biased toward intelligibility). Here it appears that Voiers considered intelligibility as an aspect in the determination of processed speech quality, and that there is a correlation between them.

2. In another study made by Voiers in 1980 [88], he investigated the relationship between intelligibility ratings from the DRT and quality scores from the DAM. In his findings, he suggested that speech quality can be predicted by other factors besides intelligibility and that in the DAM test, the attributes or items that associate most with quality are:

   (a) Perceived Distortion,

   (b) signal and background flutter, and

   (c) signal high-pass and signal nasality.

   He continued by mentioning that this finding is consistent with intuition, and results of other research of which is not stated. Finally, he concluded

[4] It was in 1977 when the DAM was first developed; one more item was added and hence in section 4.2.2, a total of 21 items instead of 20 was mentioned

that, *"overall acceptability" or "quality" is heavily but not totally dependent on measured intelligibility"*. Again, we can see the strong link between speech quality and intelligibility, and we can infer from Voiers' findings that speech quality covers a broader scope which includes speech intelligibility.

3. Preminger and Van Tasell mentioned in [67] that there are two approaches to investigate speech quality - A multidimensional approach and a unidimensional approach. In the multidimensional approach, speech quality is viewed in a multidimensional perspective and the dimensions are listed as clarity (intelligibility), fullness, brightness, softness (the antonym of sharpness), spaciousness, nearness, extraneous sounds, and loudness [35]. These dimensions allow one to realise the specific aspect in which speech quality is being affected in a piece of perceived speech and an alteration to one or more of these dimensions will actually affect the quality of speech. In the unidimensional approach, speech quality measurement became merely a form of preferential measurement. The listening subjects' preference, however, may be influenced by one or several individual quality dimensions stated in the multidimensional approach but in this approach, the specific dimension is not recognised. This approach, however, is adopted by many researchers in their research and development in assessing speech quality. The subjective MOS test and the objective PESQ [75] are examples of this approach. One point that directly contradicts the multidimensional view is that speech quality and speech intelligibility are sometimes considered as separate or sometimes even conflicting entities in the unidimensional view. The relationship between speech quality and intelligibility, and the importance of speech intelligibility to speech quality is therefore a question in this approach. To answer this question, Preminger and Van Tasell performed two experiments (reported in the same paper) specifically to quantify the relationship between speech quality and intelligibility. In both experiments, subjects were required to rate five speech quality dimensions as a function of changes to the frequency response of a listening system. These dimensions are:

    - **Intelligibility:** Percentage of spoken words a subject can understand

- **Pleasantness of Tone:** How pleasing the tonal quality of the speech sounds to the subject
- **Loudness:** How loud the speech seems to the subject
- **Listening Effort:** The amount of effort the subject needs to give to the listening task in order to understand as much of the speech as he/she can
- **Total Impression:** The overall quality or fidelity of the speech

In the first experiment, intelligibility was allowed to vary over a wide range from 25% to 100%, and it was realised that correlation between intelligibility and other dimensions was high and subjects' ratings for all dimensions except for pleasantness were remarkably similar. This high correlation suggests that in this experiment, the quality of a piece of speech could be predicted with confidence on the basis of its perceived intelligibility when intelligibility is allowed to vary widely. This again supports the claim that intelligibility is an important consideration in the measurement of speech quality. In the second experiment, intelligibility was held constant at 100% and this time, inter-subject and inter-dimensional similarities and correlation were reduced. This is due to the fact that intelligibility is the key factor in producing the high correlation in the first experiment and therefore when the influence of this factor is removed, the relationship between dimensions were greatly affected. The above further emphasises the importance of speech intelligibility in terms of quality and therefore intelligibility issues should not be taken trivially in speech quality measurement systems.

4. The experiment conducted by Licklider in [57] also illustrated this point. In his experiment to understand the effect of amplitude distortion upon the intelligibility of speech, he found out that when amplitude distortion affects intelligibility, speech quality is also affected and the degree is more severe than it affects intelligibility. The explanation to his finding is that other dimensions of speech quality are affected more than intelligibility. His result was also quoted by Voiers in [88], which we have cited not long ago, as an example to suggest that speech quality covers a broader scope than intelligibility.

5. In the subjective *speech reception threshold* (SRT) test mentioned in section 4.2.1 where the intensity of masking noise added to a word or sentence was increased until the subject cannot recognise that word or sentence. When noise was added to the speech, its quality clearly is affected but intelligibility remains high until a threshold point. This substantiates the evidence that speech quality encompasses a broader scope.

6. Steeneken stated in [77] that *"Speech quality assessment is normally used for communications with a high intelligibility, for which most tests based on intelligibility scores cannot be applied because of ceiling effects"*. This statement corresponds to the second experiment of Preminger and Van Tasell whereby when speech intelligibility was held constantly at a very high level, inter-subject and inter-dimensional similarity was reduced hence giving a fair opinion of quality in the absence of the intelligibility factor.

From the above evidence, it can be comprehended that speech quality encompasses a broader scope that includes intelligibility. Furthermore, there exists a strong correlation between speech quality and intelligibility in that the level of intelligibility relates to determination of quality. Although speech intelligibility possesses a narrower scope and can even be considered a dimension of speech quality, it is by no means inferior since it is the intelligibility of the information content that is often of primary importance in speech. Thus, the answer to our question in the beginning of this section is 'yes' and it should be as we have mentioned that *"intelligibility of the information content is of primary interest in speech."*

In our research, we shall build upon the above points to define a few relationships relating speech intelligibility to quality:

1. When intelligibility is held constantly at a high level, speech quality cannot be predicted with confidence from a measure of intelligibility, i.e. speech quality can be high or low.

2. When intelligibility varies, speech quality tends to correlate with speech intelligibility in that:

(a) high intelligibility generally yields a higher quality score, and

(b) low intelligibility generally yields a lower quality score.

In our case, we are more concerned with the second relationship since intelligibility of Chinese speech should be taken into account in the measurement of speech quality. We shall use the defined relationships (especially the second one) in our research to adjudicate the OSQMs (PESQ and MNB) and also to evaluate any improvements made to these systems.

5.2.1 Pearson's product moment correlation coefficient

The above mentioned relationship between speech intelligibility and quality can be considered as an association between these two quantities. Assuming a linear relationship between them, this relationship can be assessed by a statistical measure known as the *Pearson's product moment correlation coefficient, r* or *Pearson's correlation* in short (we shall simply call it correlation) where r is defined by the formula [50]:

$$r = \frac{n(\sum_{i=1}^{n} X_i Y_i) - (\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y_i)}{\sqrt{[n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2][n\sum_{i=1}^{n} Y_i^2 - (\sum_{i=1}^{n} Y_i)^2]}}$$

where n is the number of Chinese speech files in a dataset, X_i are the subjective intelligibility ratings (the total number of intelligibility errors recorded for individual Chinese syllables or the amount of intelligibility degradation for each CDRT or CDRT-Tone category), and Y_i are the objective quality scores in the cases of experiments 1, 2, 4, and 5 in this chapter, and the correlations from the improvements in the next chapter. r ranges from -1 to +1. A positive r means that there is a positive association between both quantities that both increase together along their axis, while a negative means that while one quantity increases, the other decreases. A zero correlation means that there is little association between the two (please refer to figure 5.1). In our case, a strong negative correlation is desired in that when the number of intelligibility error increases (a decrease in intelligibility), quality decreases.

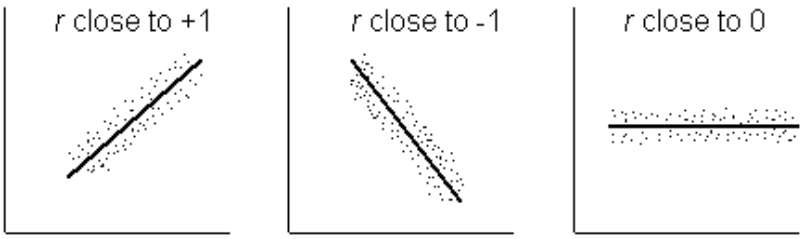

Figure 5.1: Examples of a positive, negative, and zero correlation or association.

Significance of the difference between two correlation coefficients from independent samples

When we have two correlations from independent samples (for example, the difference between PESQ's correlation with subjective intelligibility ratings in the noisy condition and PESQ's correlation with subjective intelligibility ratings in the noiseless condition), the steps to determine whether or not the difference in correlation is significant are as follows (p. 405-406 of [17]):

1. Convert both correlation coefficients (r) to their respective Fisher's z values using the r to z table (table K, pp 573-574 of [17]) or by the Fisher's transformation formula [9]:

$$z = \frac{1}{2} ln \frac{1+r}{1-r} = 1.1513 log_{10} \frac{1+r}{1-r}$$

2. Calculate the standard error of difference between the two z's by the formula:

$$\sigma_{z_1 - z_2} = \sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}}$$

   where N_1 and N_2 are the sample sizes of the two independent samples.

3. Calculate the Z value by dividing the difference between the z_1 and z_2 by the standard error:

$$Z = \frac{z_1 - z_2}{\sigma_{z_1 - z_2}}$$

4. Find out the value for the limits of the resulting confidence interval using a

table of the standardised normal probability distribution (for example, use table C[5], p. 558 of [17]). At 95%, that limit value equals 1.96 for two-tailed and 1.6449 for one-tailed significance test. If the calculated Z value is higher than the limit value from the table, the difference between the correlations is significant.

Significance of the difference between two correlation coefficients from the same sample

When we have two dependent correlations arising from the same sample (for example, the difference between PESQ's correlation with subjective intelligibility ratings and MNB's correlation with the same intelligibility ratings in our case), a *t*-test (p. 407 of [17]), where

$$t = (r_{xy} - r_{zy}) \sqrt{\frac{(N-3)(1+r_{xz})}{2(1-{r_{xy}}^2-{r_{xz}}^2-{r_{zy}}^2+2r_{xy}r_{xz}r_{zy})}} \quad , \tag{5.1}$$

with $N-3$ degrees of freedom, is used to determine the significance of the differences. Using the same example, r_{xy} would be the correlation coefficient between PESQ scores and subjective intelligibility ratings, r_{zy} the correlation coefficient between MNB scores and intelligibility ratings, and r_{xz} the correlation coefficient between PESQ and MNB scores. A one-tailed *t*-test is used in this chapter and the next to determine the significance of differences of correlations from the same intelligibility ratings.

5.3 Experiments 1 and 2: Determination of correlation between consonantal intelligibility and objective speech quality of Chinese speech

In our research, we conducted two larger scale experiments to investigate whether consonantal intelligibility is measured by PESQ and MNB in their determination of speech quality for Chinese speech. A number of smaller experiments were performed in addition to these to calibrate the results and experimental procedures. Both main experiments shared the same procedure with differences in the exper-

[5] In this table, total area under the normal curve is 10,000 instead of 1. Therefore, values for the respective confidence intervals have to be divided by 10,000.

imental parameters to reflect different conditions. Experiment 1 was performed initially to investigate the correlation between the quality of a set of processed Chinese speech recordings and intelligibility, where intelligibility is low due to the masking effect of additive noise. Experiment 2 was performed later to determine the same relationship but using a set of processed files with higher intelligibility by not including noise. More subjects were required in experiment 2 since there were fewer errors and thus more results needed overall to maintain statistical confidence (so as to yield a substantial number of intelligibility errors assuming fewer errors under noiseless conditions).

5.3.1 Parameters and procedures of experiments 1 and 2

Experiments 1 and 2 each consisted of two parts with the objective of each part being:

1. To evaluate the intelligibility of Chinese speech processed with a speech codec (with or without noise) using the CDRT test.

2. To determine correlation between subjective consonantal intelligibility ratings from part 1 and quality scores from OSQMs.

All subjects used were native Chinese speakers with no hearing impairments. In experiment 1, five subjects were used since sufficient intelligibility errors can be collected from 5 (subjects) $\times$ 192 (speech files) data points. All source files (original datasets) were recorded in an anechoic chamber with sampling rate of 16 kHz (downsampled to 8 kHz) and 16-bit resolution. The original datasets were coded and then decoded using the *GSM* [33] speech coder after which noise (simulated to a relative vehicle engine power level of 4%) was added to obtain a set of processed speech files (processed datasets). In experiment 2, 40 subjects were used. A sampling rate of 16kHz was used to record the source files and these files were also recorded in an anechoic chamber. An ITU-T *G.728* [41] LD-CELP speech coder was used this time to obtain the processed datasets. No noise was added in this experiment. A summary of the experimental parameters are shown in table 5.1

Table 5.1: Experimental parameters for Experiments 1 and 2.

Experiment	**1**	**2**
No. of Subjects	5	40
Sampling Rate	8kHz	16kHz
Speech Coder	GSM	ITU-T G.728
Noise added to processed files	Yes	No
Amount of Noise	4% relative power	—

In the first part of these two experiments, the consonantal intelligibility of Chinese Speech processed with the *GSM* (plus noise) and *ITU-T G.728* speech coders were evaluated using the CDRT intelligibility test (section 4.2.1). This evaluation was performed using a laptop computer and a high quality Philips HS900 headphone (Frequency range: 10-28,000 Hz, Sensitivity: 102 dB, Impedance: 32 Ohm). The source files (original datasets) were the 96 pairs of phonetically rhyming Chinese syllables from CDRT. These 192 (96 pairs) original plus 192 processed (coded-decoded) speech files were played to the subjects in random sequence. Each subject was required to make a closed response by selecting from two words displayed on a monitor the one which he/she perceived that was played through the headphones. To reduce errors caused by the effect of fatigue, they were given a two minute break after every 32 words played. Results of the evaluation part for experiment 2 using 30 subjects (10 more subjects were later added to increase the confidence in this evaluation so that sufficient data points can be collected. This was not needed in experiment 1 as confidence level is sufficient from the 5×192 data points.) were published in [21].

In the second part of each experiment, quality scores for the all the CDRT syllables were computed using PESQ and MNB on a computer. This process was performed by inputting the original speech files and the processed (coded) speech files into PESQ and MNB after which quality scores of the processed files were computed and output by both OSQMs with respect to the original files. The Pearson's correlation, r, was calculated between the intelligibility scores in part 1 and the computed quality scores in this part. For evaluation on each phonemic category, the quality and intelligibility scores were averaged according to the six phonemic categories listed in CDRT.

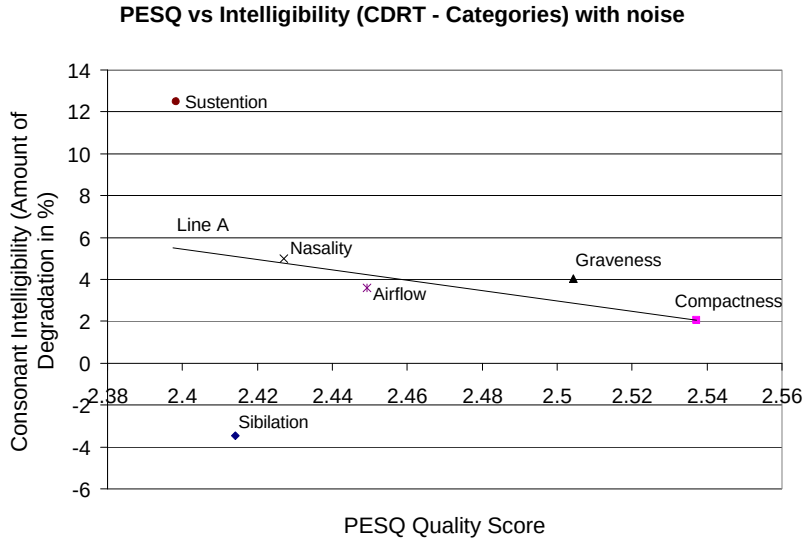

Figure 5.2: PESQ vs Consonant Intelligibility (with noise) for the six phonemic categories (Experiment 1).

5.3.2 Results

Figures 5.2 and 5.3 showed the scattered plots relating subjective intelligibility and objective quality for each phonemic category for experiment 1 and figures 5.4 and 5.5 for experiment 2. This is to illustrate the relationship between intelligibility and quality for the six phonemic categories from both experiments. The subjective intelligibility ratings were calculated as the percentage difference between intelligibility of original speech files and processed ones among each category. The amount of degradation in intelligibility, averaged quality scores, and the Pearson's correlation coefficients for each category are listed in tables 5.2 and 5.3 for experiments 1 and 2 respectively.

According to the established relationships, we would expect the categories with higher percentage of degradation to obtain a lower quality score. From experiment 1, this relationship can be vaguely seen in the PESQ vs intelligibility plot (figure 5.2) but is not conclusive from the MNB vs intelligibility plot (figure 5.3) at first sight (without considering the calculated trend line - line A). All points in figure 5.2 showed a marginal trend of a negative correlation except one

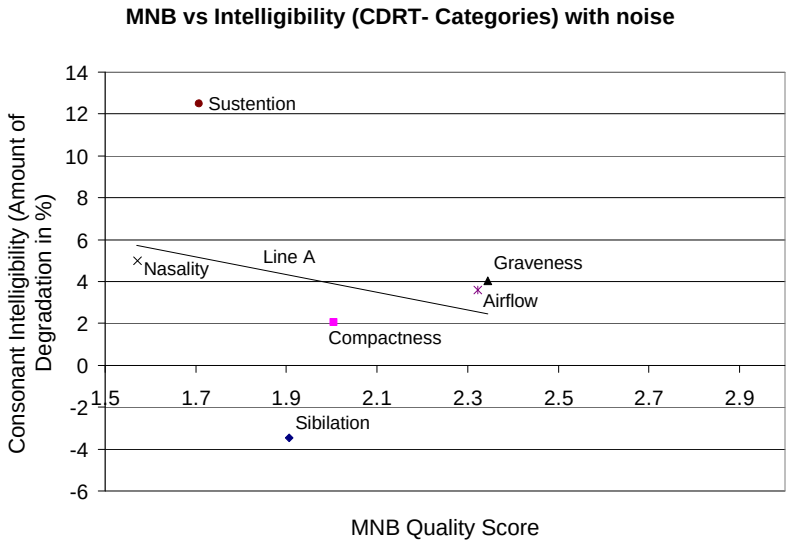

Figure 5.3: MNB vs Consonant Intelligibility (with noise) for the six phonemic categories (Experiment 1).

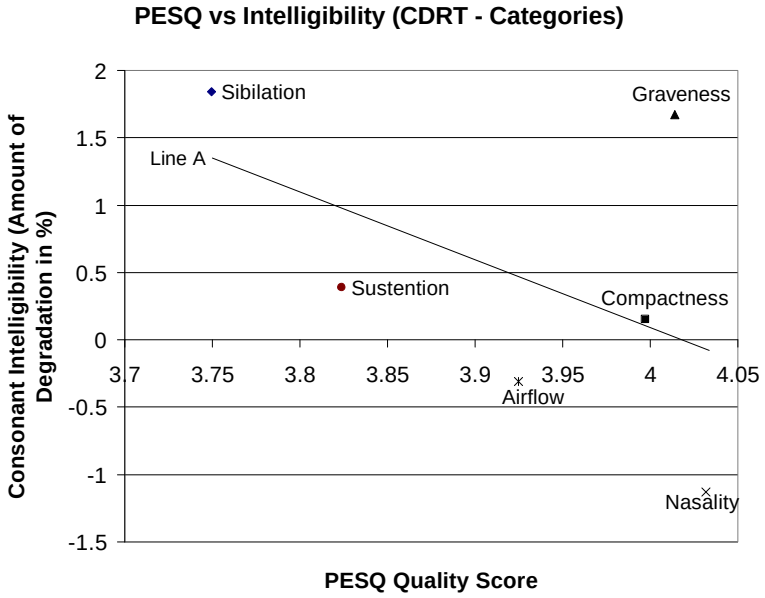

Figure 5.4: PESQ vs Consonant Intelligibility (without noise) for the six phonemic categories (Experiment 2).

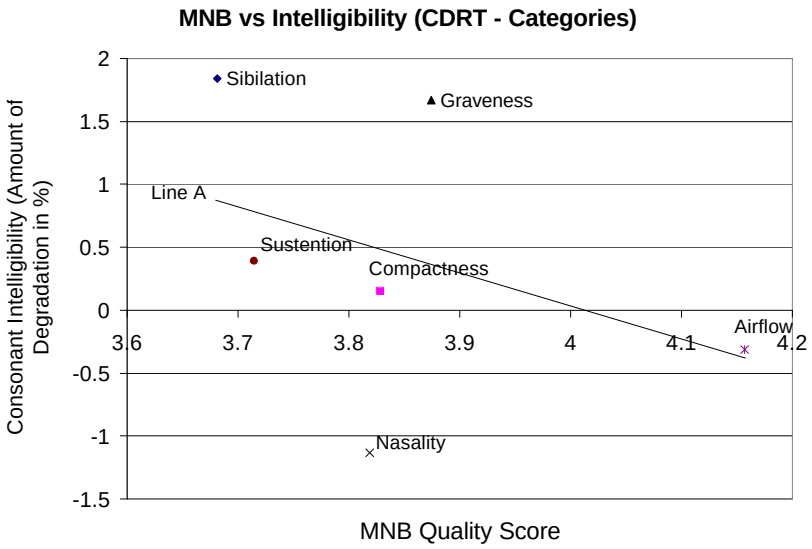

Figure 5.5: MNB vs Consonant Intelligibility (without noise) for the six phonemic categories (Experiment 2).

Table 5.2: Amount of degradation in intelligibility of phonemic categories, their averaged objective quality scores, and the correlation between amount of intelligibility degradation and quality scores for Chinese syllables with noise (Experiment 1).

Phonemic Categories	**Amount of Degradation in Intelligibility (%)**	**PESQ Quality Score**	**MNB Quality Score**	**Correlation (CDRT & PESQ)**	**Correlation (CDRT & MNB)**
Sibilation	-3.45[a]	2.41	1.91	-0.072	0.021
Compactness	2.05	2.54	2.00	-0.111	-0.090
Graveness	4.05	2.50	2.35	-0.289	-0.142
Nasality	5.01	2.43	1.57	0.024	-0.268
Airflow	3.59	2.45	2.32	-0.197	-0.097
Sustention	12.49	2.40	1.71	0.088	0.086

[a] The negative result here means that the number of intelligibility errors in the original dataset is higher than that of the processed dataset in CDRT.

[b] All objective quality scores were rounded to two decimal places and correlation coefficients listed in this table were rounded to three decimal places due to small figures.

Table 5.3: Amount of degradation in intelligibility of phonemic categories, their averaged objective quality scores, and the correlation between amount of intelligibility degradation and quality scores for Chinese syllables without effect of noise (Experiment 2).

Phonemic Categories	**Amount of Degradation in Intelligibility (%)**	**PESQ Quality Score**	**MNB Quality Score**	**Correlation (CDRT & PESQ)**	**Correlation (CDRT & MNB)**
Sibilation	1.84	3.75	3.68	-0.133	0.196
Compactness	0.15	4.00	3.83	0.035	-0.034
Graveness	1.67	4.01	3.87	-0.016	-0.101
Nasality	-1.13[a]	4.03	3.82	-0.199	-0.115
Airflow	-0.31	3.92	4.16	-0.100	0.148
Sustention	0.39	3.82	3.71	0.264	-0.234

[a] The negative result here means that the number of intelligibility errors in the original dataset is higher than that of the processed dataset in CDRT.

[b] All objective quality scores were rounded to two decimal places and correlation coefficients listed in this table were rounded to three decimal places due to small figures.

point (Sibilation). This case was not so for the MNB vs Intelligibility plot (figure 5.3). The Pearson's correlation coefficient, *r*, computed to reflect this relationship shows that for the PESQ case, a correlation of -0.257[6] was calculated and -0.288 for MNB.

The established relationship was also not clearly depicted from the plots (figures 5.4 and 5.5) for experiment 2. It only displayed a slight trend in this relationship. Worst, the category "graveness" has the second highest percentage of degradation in intelligibility (which is a considerable difference compared to the other four categories) but ranks second in quality from both OSQMs. This means to say that a category with poor intelligibility is high in quality and this clearly defies the second relationship stated in section 5.2. A correlation of -0.486 was calculated between PESQ and intelligibility, and -0.387 for MNB.

Let us narrow the examination down into the level of individual syllables. The

[6] A negative correlation value was obtained because we are calculating the correlation between the amount of degradation in intelligibility (in %) and the quality scores in which when amount of degradation increases, quality should decrease. This also applies to correlation coefficients listed in tables 5.2 and 5.3

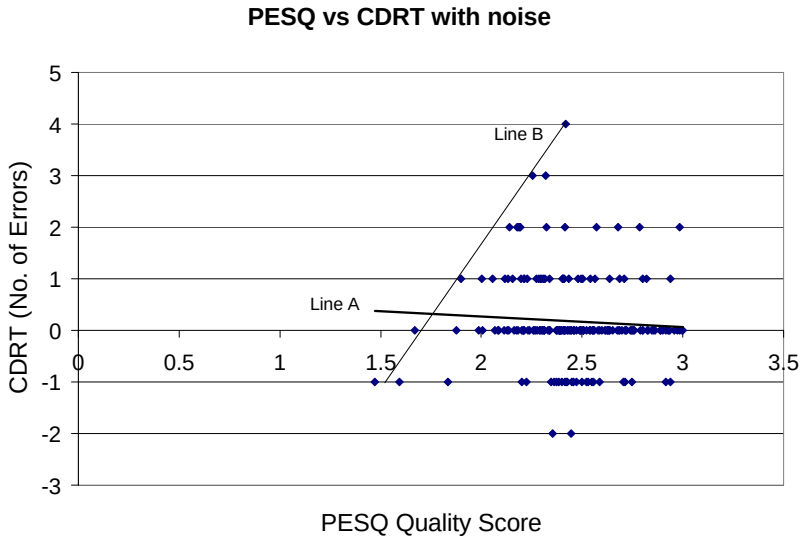

Figure 5.6: PESQ vs Consonant Intelligibility (with noise) for each individual syllable (Experiment 1).

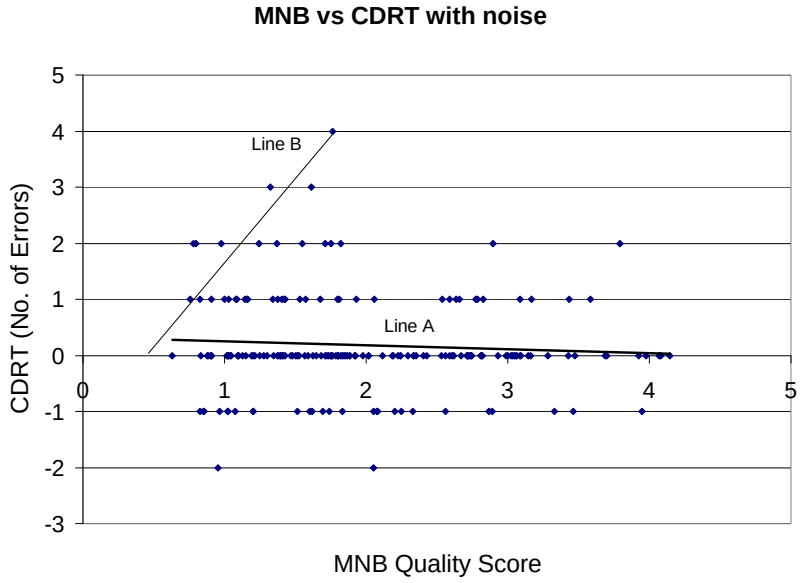

Figure 5.7: MNB vs Consonant Intelligibility (with noise) for each individual syllable (Experiment 1).

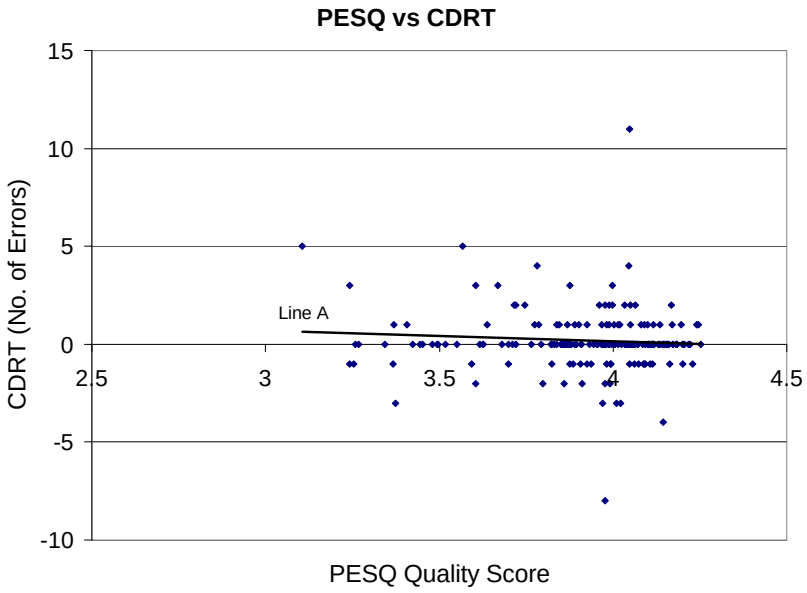

Figure 5.8: Experiment 2: PESQ vs Consonant Intelligibility (without noise) for each individual syllable (Experiment 2).

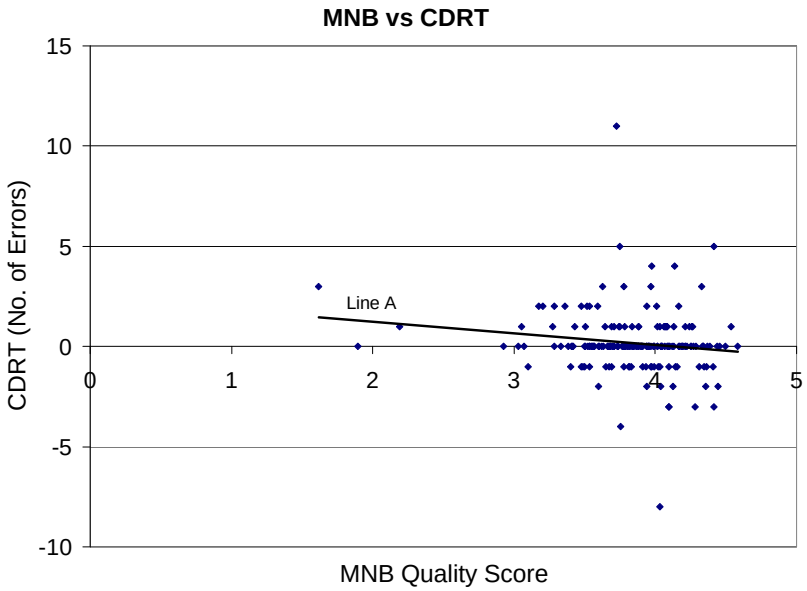

Figure 5.9: MNB vs Consonant Intelligibility (without noise) for each individual syllable (Experiment 2).

level of intelligibility (number of CDRT errors) of each syllable from the set of 192 is plotted against its own objective quality score (please refer to figures 5.6 and 5.7 for experiment 1, and figures 5.8 and 5.9 for experiment 2). From the plots for the noisy condition, signs of positive correlations can be seen at the lower end (line B in figures 5.6 and 5.7) of the quality scale. However, this phenomenon occurred mainly in the positive intelligibility scale. It is interesting to note that for MNB in the noisy case, a clear division was observed in the middle of the quality scale. Although we can see a larger fraction of syllables with errors residing at the bottom half of that scale, a considerable fraction of syllables with good intelligibility (zero or negative on the intelligibility scale) also appeared there. The case with PESQ seems worse as it was shown that the distribution of syllables with 0, 1, and 2 errors was fairly equally distributed on the quality scale. There are syllables with 1 or 2 errors having quality scores as high as or even higher than those with no or negative number of errors. The overall trend from the noisy case is a marginal negative correlation (line A in both plots). The correlation between PESQ and intelligibility is -0.065 and MNB's is approximately the same (-0.068).

Quality scores from experiment 2 are relatively higher than experiment 1's. However, nowhere in these two scattered plots showed signs of good negative correlation except in the calculated trend line (line A) which showed a slight indication. We can also see syllables with higher number of errors obtaining relatively high objective quality scores. In figure 5.8, the syllable with the highest number of errors is even ranked among the higher quality ones. Furthermore, its number of errors is twice as many as the syllables with the second highest number of errors. MNB made a fairer calculation for this same case but computed high quality scores for some syllables with low intelligibility (syllables with 3, 4, and 5 errors). The correlation coefficient between the number of intelligibility errors and objective quality scores are -0.087 for PESQ and -0.150 for MNB at the individual syllable level. These experiments relating intelligibility with quality showed some signs that these two objective systems did not pay enough emphasis to the intelligibility aspect when calculating quality scores for Chinese speech. A summary of correlation coefficients at the individual syllable level is given in table 5.4

Table 5.4: Summary of Correlation coefficients at individual syllable level.

Experiment	PESQ	MNB
1	-0.065	-0.068
2	-0.087	-0.150

5.3.3 Discussions

Looking at the phonemic category level, it is interesting to see that there is a huge contrast between the amount of degradation in intelligibility for sibilation categories between both experiments. One would naturally expect that when noise is added to the processed files, it would cause a greater reduction in intelligibility than for the case where no noise is added. In fact, the inverse behaviour is seen here. Among all Chinese consonants, sibilation can be considered to be one of the most easily confused. Many native Chinese speakers cannot differentiate between a sibilated and unsibilated syllable when they are speaking or hearing. This phenomenon is shown in experiment 2 as the sibilation category yields the highest amount of degradation. The reasons why experiment 1 did not illustrate this phenomenon are discussed below. One important factor that influences recognition of sibilation is the environment. Many Chinese speakers speak more than one Chinese dialect. Mandarin Chinese is usually one of these as it is the official dialect as mentioned in section 3.4. Although Mandarin Chinese is the official dialect, many Chinese speakers use a local dialect more frequently. As some dialects, for example, Hakka and Cantonese, do not or seldom consider sibilation differences, a Mandarin Chinese speaker might be heavily influenced by their own local dialect hence not paying sufficient emphasis to the discrimination of sibilations. To explain the case for the sibilation category in experiment 1 where noise is present, let us look into the issue of whether or not noise has a detrimental effect on the discrimination of sibilants. The sibilation category contains mainly high frequency components with a significant level of energy, for example the fricatives. As the amplitude of vehicular noise is higher at lower frequencies usually below 2kHz (interpreted from figure 6.3 in [18], and figures 3 and 12 in chapter 9 of [94]), thus noise does not cause as much reduction in the intelligibility of higher frequencies. Another explanation could be that the amount of degra-

dation is the percentage difference of intelligibility errors between the processed and original speech, therefore the number of errors in the original speech itself also contributed to this result. In experiment 1, a sampling rate of 8 kHz (16 kHz downsampled to 8 kHz) was used for recording the original speech files. This means to say that the *Nyquist* frequency is 4 kHz which is insufficient to record the sounds for high frequency speech components (section 3.3.4). Nevertheless, this limitation is common to most telecommunications systems. From the massive 48 errors recorded for the original speech files (the highest number of errors among the six categories with the second highest being only a third of this with 16 errors), it was shown that intelligibility is not desirable using this sampling rate. Since the intelligibility of the original speech files is low, effects of noise will not be as detrimental as when intelligibility is high. In other words, intelligibility of the processed files will be affected more when the intelligibility of the original speech files is higher, and less affected when intelligibility is lower. This, therefore, explains the situation of better intelligibility for the sibilation category in this condition where 44 intelligibility errors were recorded for the processed files.

Regarding the measurement of quality, since both PESQ and MNB are intrusive methods that compute quality based on perceptual differences between the original and processed speech files, the lowest amount of degradation in intelligibility should have received the highest (if not, a higher) quality score. This is not true in the case of experiment 1 (especially for the sibilation category). We even see an inverse of this trend in MNB for the *sibilation*, *compactness*, *airflow*, and *graveness* categories and the *nasality* and *sustention* pair. This shows that MNB does not regard intelligibility that highly in this case.

The plots for the categories in both experiments (figures 5.2, 5.3, 5.4, and 5.5) showed a slight conformance (see line A) to the perfect correlation trend line assuming a trivially linear relationship of $y = -x + 5$. Besides the sibilation category previously mentioned, the graveness category in experiment 2 is also worthy of note as it is the one which mostly defies the quality-intelligibility relationship for both PESQ and MNB. It yields the second highest amount of degradation that is slightly less than the amount for sibilation (Note the huge gap between it and the category of sustention which has the next lower degree of degradation). However, it ranked second in terms of quality despite its position in the intelligibility scale

from both OSQMs. The case with PESQ is worse as its quality is very close to that which yields the highest quality whereas in MNB, the same situation is diffused by the larger lapse in the quality scale. This phenomenon can be explained by the attributes of the consonants. Firstly, consonants in this category contains a high proportion of low frequency energy. Furthermore, half of the consonants are plosives in which their length is shorter than other consonants. This makes the calculation of quality, where intelligibility is to be taken into account, for this category difficult as the wavelength of the consonants are long and half of the words have short durations. This means to say that even when intelligibility is low when the consonant portion is missed by the hearers, speech quality will not be affected as much as intelligibility. Such cases in an OSQM might be considered as a trivial distortion in terms of quality. The plots from experiment 1 also showed a rather similar trend for the graveness category.

Looking into the correlation between categories, we observed that all except two for both PESQ and MNB are negative in both experiments. A negative correlation is desirable in our case since an increase in the amount of intelligibility degradation will lead to a decrease in quality. This means to say that the more negative the correlation coefficient, the better the system in our case. Looking at commonalities, the correlations for the sustention categories are positive in both PESQ and MNB in experiment 1 and in PESQ for experiment 2. Despite its worst intelligibility in experiment 1 with the lowest averaged PESQ and second lowest averaged MNB score, the positive correlation may have resulted from a marginally decreasing degree of intelligibility with increasing quality scores. This could be due to the length of the consonants in this category that may affect intelligibility. Half of the syllables are longer and the other half shorter in length. The above deduction will make sense if syllables with short C1 durations yet yielding better quality scores than other syllables in the category. The correlations for this category in experiment 2 are even more interesting as two extremes occurred. Correlation for PESQ is the worst (most positive) and MNB is the best (most negative) while the average quality scores from both OSQMs are the second lowest among the categories. The differences in computing quality scores with respect to intelligibility can therefore be clearly seen here between the two OSQMs. MNB works better in this case with a lower (better) correlation than PESQ. Another

commonality can also be seen in the sibilation categories for MNB in both experiments where both correlations are positive. It was shown here that MNB could not handle consonants of such acoustic characteristics (previously briefly mentioned), which is common in Chinese speech, as well as PESQ.

At the level of individual syllables, the PESQ vs CDRT plot for experiment 1 seemed to illustrate an opposing trend to the established relationship for syllables at the lower end of the quality scale (line B) with decreasing intelligibility. Despite this feature, we cannot conclude that PESQ does not take consonantal intelligibility into account because we must consider all points in the plot besides those mentioned. For MNB, a clearer distinction can be seen by the division in the middle of the the quality scale that there are more syllables with higher intelligibility errors residing in the lower half of the quality scale. However, there are also syllables that are highly intelligible residing on the same side and this offsets the correlation. Therefore, its correlation is approximately the same as PESQs. The situation for MNB in experiment 2 resembles PESQ in experiment 1 as most points were clustered at one region with only few exceptions.

There is a particular Chinese syllable from experiment 2 that draws our attention. From the plots at the individual syllable level (figures 5.8 and 5.9), we observed that this syllable has the highest number of intelligibility errors. However in the PESQ vs CDRT plot, it lies on the upper end of the quality scale. In MNB, it lies in the middle. This is the Chinese syllable /bing3/ with a plosive consonant for the graveness category. Although it has 11 intelligibility errors (slightly more than twice the second highest number of 5 intelligibility errors), its PESQ score is 4.05 which is higher than the average scores of all categories (table 5.3). A probable reason for this is due to the the relative duration of its initial consonant C1 and its relative intensity or amplitude with respect to the whole signal. Let us examine the oscillogram of this syllable in figure 5.10. The occurrence of C1 is approximately from sample 10896 to 11008[7]. From the figure, we can see the contrast in terms of duration and amplitude between the initial consonant and the whole syllable that they are indeed negligible compared to the vowel and the end or coda consonant (sample 11109 to 18880). Time-wise, the percentage

[7] This range is obtained manually by listening precisely to the sound file segments corresponding to the enlarged part of the signal.

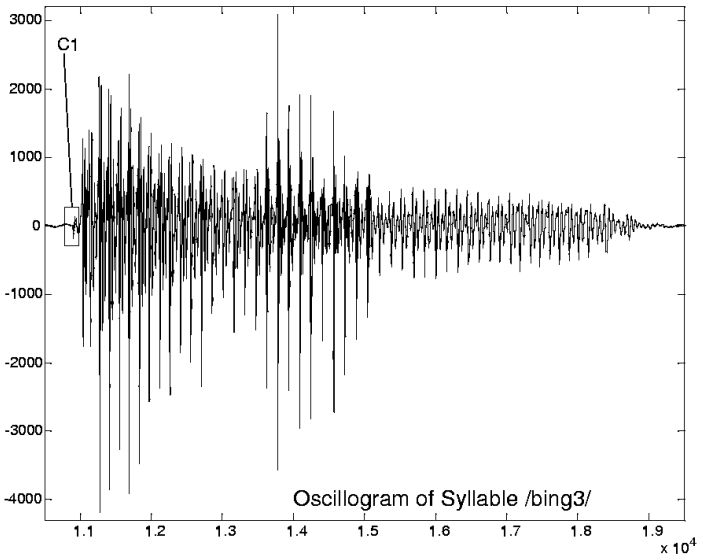

Figure 5.10: Oscillogram of Chinese syllable /bing3/.

of occurrence of the initial consonant is about 0.59% of the whole syllable and amplitude-wise, about 4% of the highest amplitude. Due to this minute fraction in duration and amplitude, it would not be surprising that less attention will be given to this part which is most important to the intelligibility of the syllable. Therefore a high quality was attributed by PESQ for this least intelligible syllable. MNB made a fairer calculation computing a score of 3.73. Figure 5.11 shows the reason why its intelligibility was low as it was shown that the amplitude of the consonant of the processed signal was attenuated significantly. Since the amplitude of the initial consonant for the original is already low, a further attenuation would cause it to be more susceptible to noise and masking effects. Therefore its intelligibility was affected. However, since the intelligibility bearing consonant constitutes only a minor portion in terms of time and amplitude of the whole syllable, quality is considered to be affected less than intelligibility.

One positive sign we can see from all the plots is that the overall quality from both PESQ and MNB decreases with the addition of noise. The average PESQ score is 2.46 and 3.92 for the cases with and without noise, and the average MNB

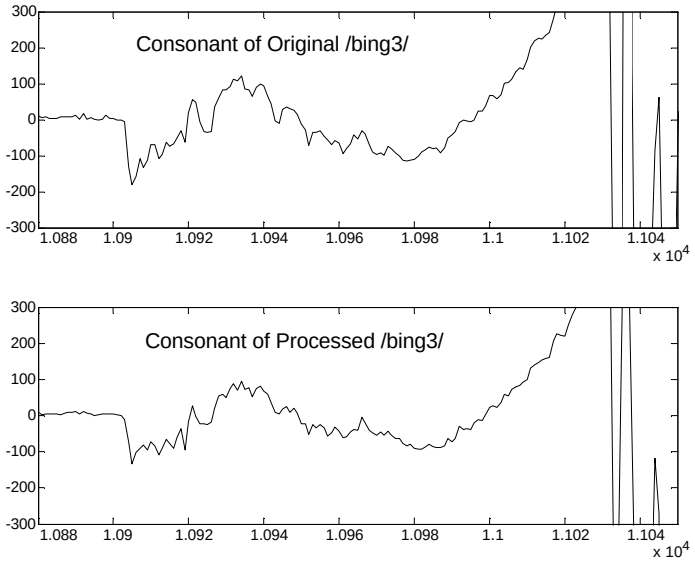

Figure 5.11: Oscillogram of the consonant for the original and processed signal of /bing3/.

scores are 1.98 and 3.85. It can be clearly seen in here that when noise is added that degrades intelligibility, quality also decreases. Of course, the decrease in quality may not necessarily be due to the reduction in intelligibility, however, there is still a possibility if we deduce it from the relationships we established in section 5.2. When this possibility stands, we can apparently see the relationship being fulfilled in the inter-condition (with and without noise) level. However, at the inter-syllable level, this is not so. From these results, we realised that neither PESQ nor MNB take consonantal intelligibility into account as much as they did for noise.

5.3.4 Conclusions

In conclusion for these two experiments, there is no visual evidence illustrating a good correlation between speech quality and intelligibility. Neither were desirable correlation coefficients obtained from the calculation. Therefore, we prove our conjecture with these two experiments that the performances of both PESQ

Table 5.5: Averaged Objective Quality scores for CDRT speech files with C1 removed and their average percentage of C1 duration.

Category	PESQ	MNB	% Duration
Sibilation	2.773	4.948	11.4 %
Compactness	2.550	4.943	7.2 %
Graveness	2.699	4.936	7.0%
Nasality	2.524	4.935	5.5
Airflow	2.781	4.947	5.8%
Sustention	2.557	4.947	11.1%
Overall	2.649	4.943	8.0%

and MNB are unsatisfactory with regards to consonantal intelligibility in their computation of speech quality.

5.4 Experiment 3 - Objective speech quality measurement on Chinese syllables with initial consonant (C1) replaced by silence

5.4.1 Introduction and results

The aim of this experiment was to investigate the measurement of speech quality by the two mentioned OSQMs for processed Chinese speech which displayed loss in consonantal intelligibility. In this experiment, the processed dataset was obtained by replacing the initial consonant (C1) from the original dataset of CDRT by silence. No other distortions were caused so that the measurement of quality is purely under the influence by the loss of intelligibility. Quality scores were computed using PESQ and MNB for these speech files. Average PESQ and MNB scores for all syllables and syllables in each of the six phonemic categories are recorded in table 5.5 together with their average percentage of C1 duration with respect to the whole syllable.

As shown on the table, MNB failed to give a reasonable score as it calculated a nearly perfect result of 4.935 and above [8] (where 4.95438085 is the perfect score

[8] MNB's output range is from 0 to 0.99087617 (rounded to 1) and the scores given for the Chinese speech files were multiplied by 5 to provide a direct comparison with the MOS scores. Although this is far from a prefect scaling, it does provide a sufficient estimate for approximate evaluative purposes.

for two exactly the same speech files, i.e. the original speech file is unaltered) for the averages of all phonemic categories and the overall average. The lowest score among all the syllables is 4.821 which is still a very high score.

Based on the high correlation between the intelligibility and quality of speech mentioned, we can conclude that this set of results from MNB is highly undesirable. This means to say that MNB would rate a totally unintelligible Chinese syllable as a high quality one. PESQ is more reasonable in its calculation compared to MNB. However, all the average PESQ scores listed were higher than fair or reasonable quality (2.5) given that the PESQ output range was from -0.5 to +4.5. Of all the CDRT speech files, 91.1% of the total number of speech files yielded a quality score higher than 2 (centre point of the PESQ output score) and nearly 60% (59.9%) yielded a quality score equal or higher than 2.5 (fair quality). At the higher end of the scale, 5.2% of all files obtained a score higher than 3.5 which is considered to be good quality. The average PESQ score for all files is 2.649. These results showed that although intelligibility was lost for these files, PESQ rates them with a fair or reasonable quality on the average, and for the 5.2% of speech files mentioned, PESQ regards them as good quality. When this is the case, an inaccurate picture will be portrayed for codecs or other telecommunication systems showing them to be good quality systems even if intelligibility is totally lost when they processed their speech. This might reduce the credibility of OSQMs since their evaluation of sound processing systems are inaccurate in terms of speech intelligibility, recalling as we have previously mentioned in section 5.2 that intelligibility is often of primary importance in speech communications.

5.4.2 Discussion

The replacement of C1 by silence from the speech files in this experiment is in fact the condition of temporal clipping of speech listed in table II.1 of the ITU-T P.861 [43] and table 3 of P.862 [45] recommendations. In the notes relative to table II.1 of the ITU-T P.861 recommendation, it was mentioned that for the case of temporal clipping of speech, *"insufficient information is available about the accuracy of the objective measures with regard to this variable"*. The high quality scores we obtained clearly showed that MNB inaccurately measures this kind of distortion. A probable explanation is that to MNB, the effect of a temporal clipping

of speech to any portion of it would be of equal importance regardless whether it is the consonants or vowels, unless some form of discrimination algorithm is incorporated in the system to discriminate them. Noting the low average percentage (8%) of the duration of C1 with respect to the whole syllable, a loss of this 8% might have been regarded as a trivial degradation and thus seem to be equally unimportant throughout the whole syllable regardless of whether it occurs on C1 or vowel. Therefore, when this kind of distortion happened to C1 in the processed speech files, MNB will provide an inaccurate computation for those unintelligible Chinese speech.

For PESQ, it is stated in the notes of table 3 in the ITU-T P.862 recommendation that *"PESQ may be less sensitive than human subjects to regular, short time clipping"*. As shown in our results, when C1 (which is the critical information bearing part in speech intelligibility) is being silenced, a reasonable quality rating was still given. Coming back to the point we mentioned in the previous paragraph, a question to ask is, *"does PESQ treat temporal clipping of the consonants as a more disastrous form of degradation than the temporal clipping of the vowels or does it regards them as of equal?"* If they are regarded as equal, the results here only show the measured speech quality under the condition of a general temporal clipping of speech regardless which portion of speeches were clipped. If it is does treat the temporal clipping of consonants as more disastrous, then the results would indicate an unsatisfactory computation made by PESQ for the loss of C1.

As a side track from our research, one point we could gather from these results is that for PESQ, the duration of the clipped part does not have a great influence on the quality score. From table 5.5, we behold that the sibilation category which has the longest C1 duration obtains the second highest PESQ score while the highest (Airflow) PESQ score is calculated for the second shortest C1 duration. A low correlation of 0.12 was computed between the averaged PESQ scores and duration from the table. From this observation, it is probable that other factors like the nature of the signal could be a greater influence to the PESQ scores for the effect of temporal clipping of speech which we will not cover here.

5.4.3 Conclusions

From this experiment, we cannot conclude whether PESQ and MNB account for the loss of intelligibility due to the condition of temporal clipping of the initial consonant in their calculation of speech quality. However, we realised that there is an effect on PESQ for this kind of distortion whereas MNB is insensitive. Nevertheless, we have found out that PESQ and MNB would inaccurately measure speech quality when temporal clipping coincidentally removed C1 from a Chinese syllable making it unintelligible. Therefore, this calls for further research into developing a system that accounts for this condition.

5.5 Experiments 4 and 5: Determination of correlation between tonal intelligibility and objective speech quality of Chinese speech

It was previously mentioned in section 3.4.1 that two Chinese syllables, which shared the same phonemic pronunciation, have different meanings if their tones are different. That is to say if the tone of a processed Chinese syllable is seriously distorted, its meaning will also be lost. When this situation occurs, we can conclude that there is a serious degradation in speech quality (since the intelligibility of it is affected) even if that processed syllable sounded perfectly clear (high fidelity) and noiseless. Since English and most other European languages are not tonal, an OSQM that was designed for these languages may not be sensitive to Chinese tones. To investigate this claim, two further experiments (experiments 4 and 5) were conducted in our research. The experiment setup and procedure of both were similar to that of experiments 1 and 2 mentioned in section 5.3. This time, instead of investigating whether consonantal intelligibility is taken into consideration by PESQ and MNB in their determination of speech quality for Chinese speech, the consideration of tonal intelligibility was investigated. Similar to the consonantal intelligibility case, experiment 4 was designed to examine the relationship between the quality of a set of processed Chinese speech (CDRT-Tone dataset) and tonal intelligibility where tonal intelligibility is assumed to be lower due to noise. Noise was excluded from experiment 5 to determine the same relationship where tonal intelligibility is reckoned to be higher. Assuming subjects would commit less errors in the quiet (noiseless) condition, more subjects partici-

pated in experiment 5 to obtain a substantial number of errors in tonal intelligibility.

5.5.1 Parameters and procedures of experiments 4 and 5

Similar to experiments 1 and 2, experiments 4 and 5 also consisted of two parts with the objective of each part being:

1. To evaluate the tonal intelligibility of Chinese speech processed with a speech codec (with or without noise) using the CDRT-Tone test.

2. To calculate correlation between subjective tonal intelligibility ratings from part 1 and objective quality scores.

All experimental parameters and procedures are similar to the two experiments in section 5.3, except that eight subjects instead of five were used in experiment 4 to obtain tonal intelligibility errors from 8 (subjects) $\times$ 80 (speech files) data points. The source files used in these two experiments were the 40 pairs of phonetically similar but tonally different Chinese syllables from the CDRT-Tone test [26]. A summary of the experimental parameters are shown in table 5.6. The tonal intelligibility of the 40 pairs of CDRT-Tone files coded with the *GSM* (plus noise) and *ITU-T G.728* speech coders were evaluated using the CDRT-Tone test in the first part of both experiments. Results of this first part for experiment 5 using 30 subjects were published, together with the results from experiment 2, in [21] (10 more subjects were later added to increase the confidence in this evaluation so that sufficient data points can be collected. This was not needed in experiment 4 as confidence level is sufficient from the 8 $\times$ 80 data points). In the second part, quality scores for all 80 (40 pairs) syllables were computed by PESQ and MNB. The Pearson's correlation coefficient was determined between tonal intelligibility and speech quality.

5.5.2 Results

At the CDRT-Tone Category level (where category 1 consists of the *Tone 1* - *Tone 2* pair, category 2 the *Tone 1* - *Tone 3* pair, category 3 *Tone 2* - *Tone 3*, and category 4 *Tone 3* - *Tone 4*), the averages of the quality scores from PESQ and MNB

Table 5.6: Experimental parameters for Experiments 4 and 5.

Experiment	**4**	**5**
No. of Subjects	8	40
Sampling Rate	8kHz	16kHz
Speech Coder	GSM	ITU-T G.728
Noise added to processed files	Yes	No
Amount of Noise	4% relative power	—

were plotted against the amount of intelligibility degradation for each category in figures 5.12, 5.13 and, 5.14, 5.15 for both noisy and noiseless conditions. This is to reflect the level of intelligibility of each tonal category and to depict their corresponding averaged quality rating (the average among all syllables in a particular category).

Since there are only four categories, the correlation trend cannot be determined at a category level. Only preliminary deductions can be made. From figure 5.12, PESQ seem to have taken tonal intelligibility into account in the noisy condition as the plot illustrated a good negative correlation for three of the points. Although category 3 (*Tone 2 - Tone 3*) with a negative amount of intelligibility degradation should have yielded a higher quality score, its low quality could be due to the lower ratings of other quality dimensions since intelligibility is considered to be high in this case. Due to this category being way off the good negative correlation trend from the three others, an undesired positive correlation of 0.657 was obtained for this case (line C on figures 5.12, 5.13, 5.14, and 5.15 showed the trend lines). Although the good negative correlation trend for the three points was not so straight for MNB as category 2 (*Tone 1 - Tone 3*) obtained a slightly lower quality score than category 1 (*Tone 1 - Tone 2*), a better (lower) correlation of 0.567 was obtained.

The plots for the noiseless case did not depict any good negative correlation for any of the OSQMs. Instead, signs of positive correlation were seen in MNB. This was confirmed from the calculated trend line having a positive gradient with correlation of 0.560. The gradient of the trend line for PESQ is negative and its correlation is -0.377. There are no resemblance between both plots for PESQ and

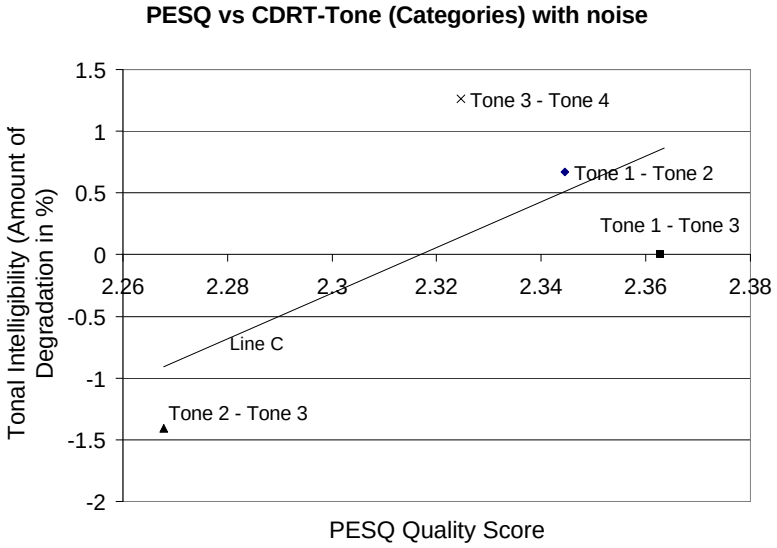

Figure 5.12: PESQ vs Tonal Intelligibility (with noise) in tonal categories (Experiment 4).

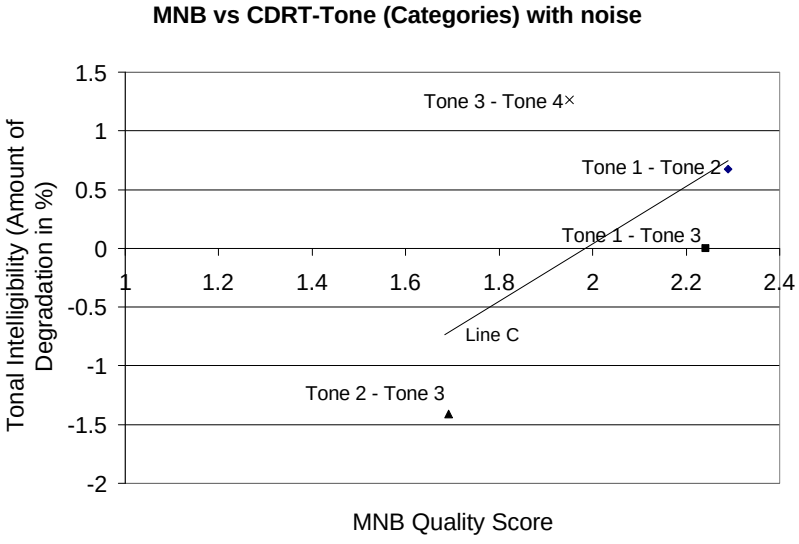

Figure 5.13: MNB vs Tonal Intelligibility (with noise) in tonal categories (Experiment 4).

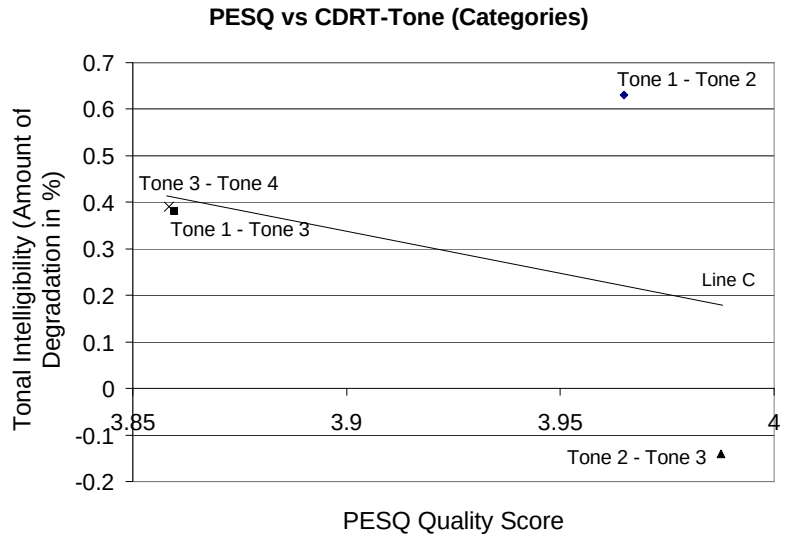

Figure 5.14: PESQ vs Tonal Intelligibility (without noise) in tonal categories (Experiment 5).

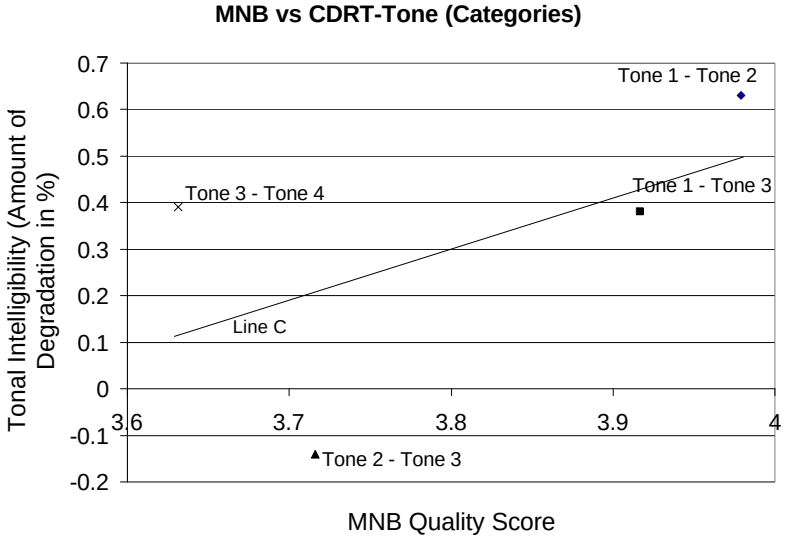

Figure 5.15: MNB vs Tonal Intelligibility (without noise) in tonal categories (Experiment 5).

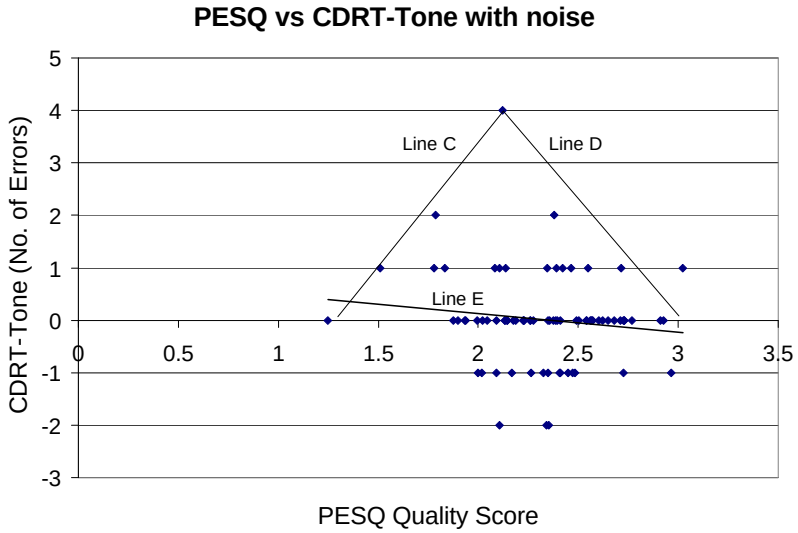

Figure 5.16: PESQ vs Tonal Intelligibility (with noise) for each individual syllable (Experiment 4).

MNB as there was in the noisy case. We can also see that the category with the highest amount of intelligibility degradation (*Tone 1 - Tone 2*) ranked second in quality scores by PESQ and first by MNB.

To illustrate the correlation between tonal intelligibility and objective speech quality more vividly, the number of intelligibility errors from the CDRT-Tone test for all 80 Chinese syllables were plotted against their quality scores from PESQ and MNB for both noisy and noiseless conditions. Figures 5.16 and 5.17 illustrated the plots for the noisy condition, and figures 5.18 and 5.19 for the noiseless.

For the condition with noise, figure 5.16 showed a nearly perfect equi- or bi-lateral triangular shape for PESQ in which we can see both a good positive correlation trend at the lower quality end (line C) and a negative one at the other end (line D) on the positive CDRT-Tone axis. The inverse of this can vaguely be seen on the negative side of the same axis. Correlation between PESQ and intelligibility is -0.127 for this case with the trend line shown as line E. The averaged PESQ score for syllables without errors (or negative) is 2.35 which is higher than those with errors (2.23). The MNB plot (figure 5.17) also resembles a triangle on the positive side of the CDRT-Tone axis. However, the gradient seems to be steeper

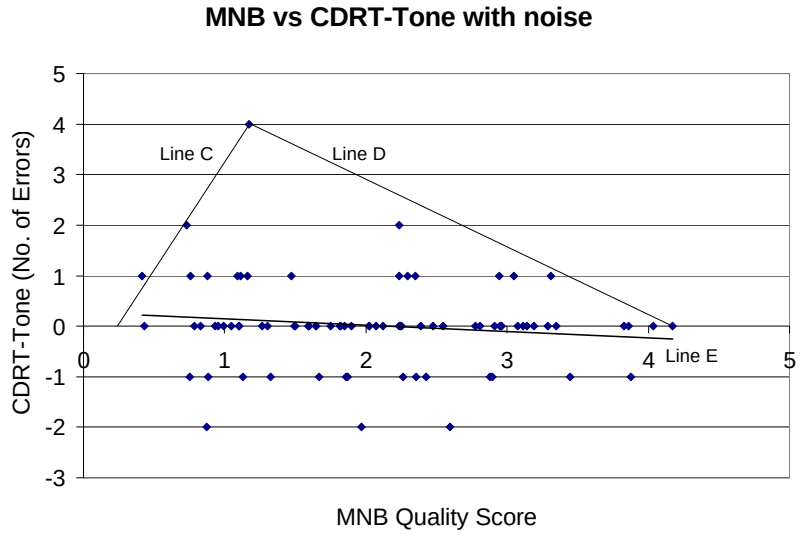

Figure 5.17: MNB vs Tonal Intelligibility (with noise) for each individual syllable (Experiment 4).

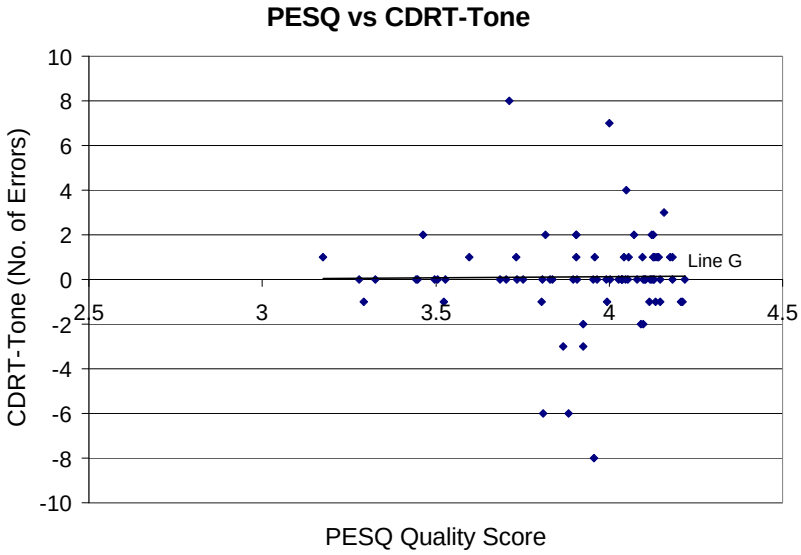

Figure 5.18: PESQ vs Tonal Intelligibility (without noise) for each individual syllable (Experiment 5).

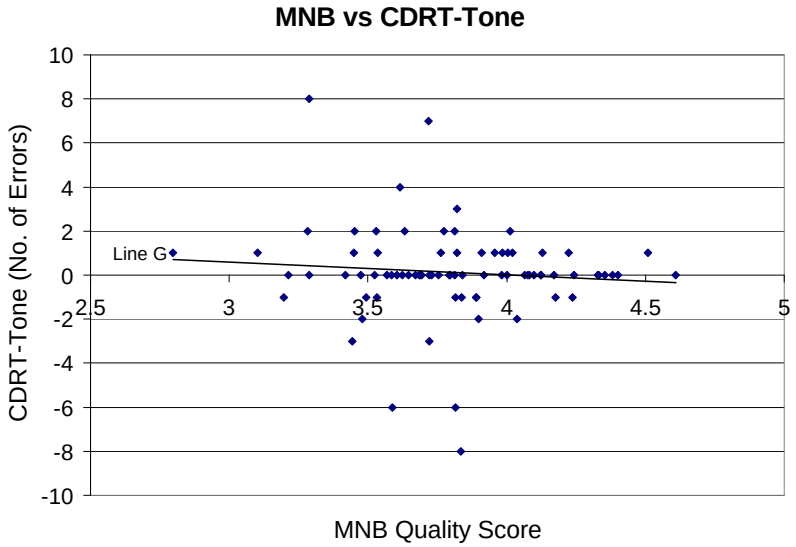

Figure 5.19: MNB vs Tonal Intelligibility (without noise) for each individual syllable (Experiment 5).

at the lower end of the quality scale (line C) and gentler at the higher end (line D). The trend line is again shown as line E with a better correlation of -0.133. The syllables without errors yielded a higher averaged MNB quality score of 2.113 than those with errors (1.701).

Looking at figure 5.18, we can see that the syllables with 3, 4, and 7 errors lying in the higher half of the quality scale respective to the other syllables. Their PESQ score is 4.159, 4.05, and 4 all of which are higher than the average (3.918). Although the decreasing PESQ score with increased number of intelligibility errors portrayed a good negative correlation among these three, their score could be lower to fit into the overall picture. The syllable with 8 errors should also have a lower quality score. As depicted in line G, correlation between speech quality and intelligibility is a positive 0.013 for PESQ in this condition which is not fitting to the supposedly negative correlation.

Figure 5.19 for MNB showed that the quality scores for the three syllables previously mentioned (syllables with 3, 4, and 7 errors) are more acceptable here. Line G showed a trend line with a negative gradient and a correlation of -0.096 was obtained in this case.

Table 5.7: Summary of Correlation coefficients between tonal intelligibility and speech quality.

Experiment	PESQ	MNB
4	-0.127	-0.133
5	0.013	-0.096

A summary of Pearson's correlation coefficients for PESQ and MNB in both conditions are given in table 5.7.

5.5.3 Discussions

The distinction between Chinese lexical tones can be analogous to that of the touch tones associated with each number on the telephone where these tones are not easily confused among themselves. Zhang mentioned in [102] that the four Chinese lexical tones exhibit a strong anti-interference property as it is a type of frequency modulation. Chinese tones can be recognised even in some extreme transmission conditions. From part one of experiment 4 from the same section where the CDRT-Tone test is applied to test the processed speech files, tonal intelligibility was over 93% for both original and processed speech files for all categories except category 3 (*Tone 2 - Tone 3*) in the noisy condition. Intelligibility was over 98% in experiment 5 besides the same category mentioned. Here we realised that tones 2 and 3 are the most easily confused ones even in the original speech files. Tone 2 has a mid-rising pitch contour while tone 3 is mid-falling-rising (please refer back to section 3.4.1). While the rising pitch from both tones exhibit some similarity, tone 3 would have distinguished itself by its initial falling pitch. The reason why Chinese speakers are confused over these tones might be due to their knowledge and recognition of them. Sometimes, tone 3's initial falling pitch was not completely articulated or exaggerated by a Chinese speaker. It would, therefore, sound like a constant followed by a rising pitch which strongly resembles tone 2. This misconception in the tones can be strongly related to the speaker's environment and the local dialect they speak. For example, a Taiwanese or Beijing speaker will pronounce the Chinese word for sister, "jie3 jie5" with the first syllable a third tone followed by a neutral one while an East Malaysian

speaker will pronounce as “jie2 jie3”. Hence there is a misconception between tones 2 and 3 in the first syllable.

Due to the strong anti-interference property of Chinese tones, it is unlikely that a Chinese speaker will misinterpret the tones unless he/she had a false knowledge regarding the pitch of the tones in the first place. Therefore, an effective OSQM would have distinguished any change of tone after a Chinese speech file was processed and this change of tone should cause a significant degradation in quality as it is in intelligibility. However, as this type of “distortion” is not common such as noise, loudness level, etc, that will occur on processed English and other non-tonal languages, OSQMs customised for these languages might result in erroneous quality scores for the distortion of tones. At the category level (figures 5.12, 5.13, 5.14, and 5.15), there are slight indications of a good negative correlation in the noisy situation except a point that lies far away from the other three. This point is category 3 (*Tone 2 - Tone 3*), which seemed the most intelligible among the categories. However, as we have previously mentioned, tones 2 and 3 may be easily confused by some Chinese speakers. This can be shown by the highest overall number of CDRT-Tone errors recorded for both the original and the processed speech files for this category compared to the rest. 19 errors were recorded for the original files and 17 for the processed files for this category while the second highest number of errors for original files were 9 and processed files were 10 both from category 1. A Pearson’s correlation was computed for the number of errors between the original and processed syllables in this category and a value of -0.316 was obtained. This negative correlation suggests that a syllable with the highest number of errors for the original speech files may have a slight chance that the number of errors from its processed counterpart would be the lowest (zero error) and vice versa. It is also highly unlikely that a syllable with no error in the original file will also obtain no errors in the processed. Therefore, there will be several syllables in this category where the intelligibility of the processed file would be worse than the original. This perhaps explains why this category is the best in intelligibility among the rest but also yielded the lowest quality score in the noisy condition if tonal intelligibility is taken into account. Figures 5.14 and 5.15, however, showed a different picture. The quality score ranking for this category (*Tone 2 - Tone 3*) is different between PESQ and MNB. Again, we can see the difference

between the "perception" of tonal intelligibility in both systems.

Although table 5.7 showed that MNB's correlations are better than PESQ's in both conditions, a one-tailed *t*-test showed that results were insignificant in both cases. It would be surprising to observe that MNB performs better than PESQ if results were significant. We can also see that correlation is better in the noisy case than the noiseless case for PESQ. However, this difference is significant at $0.01 \leq \alpha < 0.5$. The inconsistency of the sensitivity of tones can be seen here in the presence and absence of noise. The reason for both of these phenomenons may be due to either systems' incapability of handling tonal distortion. It was mentioned earlier in this section that the meaning of a Chinese syllable would be lost if its tone is distorted even though the syllable might sound perfectly clear (high fidelity) and noiseless. In the noiseless condition, if an OSQM did not take tonal intelligibility into account, then a high quality score might be given to a speech file that is tonally unintelligible. This quality score is of course inaccurate in terms of tonal intelligibility hence resulting in an erroneous correlation coefficient between quality and tonal intelligibility. This declination in correlation for the noiseless PESQ case may be the indication of an erroneous result.

5.5.4 Conclusions

In conclusion, neither of the plots and correlation values at the syllable level showed a sign of a good negative correlation between the amount of intelligibility degradation and speech quality. The unexpected worse correlation in the noiseless condition for PESQ compared to the noisy one could suggest that Chinese tones was not taken into due regards by PESQ. Further research is therefore required to investigate this case.

5.6 Conclusions on the evaluation of existing OSQMs

In this chapter, we have established two relationships relating speech quality and intelligibility. Basing on these relationships, in particular the second one, we evaluate an OSQM to investigate whether it accounts for speech intelligibility in its computation of speech quality by looking at the calculated Pearson's correlation coefficient. From the two experiments in section 5.3, we found out that there are

low correlations between the quality scores from the objective PESQ and MNB, and intelligibility ratings from the subjective CDRT. This shows that both PESQ and MNB did not give high regard to the consonantal intelligibility of Chinese speech. Regarding the third experiment mentioned in section 5.4, we cannot conclude that the loss of intelligibility due to temporal speech clipping to the initial consonant are taken into account by PESQ and MNB in their determination of speech quality. However, if this happens it causes the Chinese syllable to be unintelligible, and we know that both systems would give erroneous results. While this condition calls for further research, there are ways to improve the correlation for the conditions mentioned in section 5.3. These improvements will be discussed in the next chapter.

Two experiments were also conducted to investigate whether tonal intelligibility is taken into consideration by the OSQMs. The results from these experiments were that correlations between tonal intelligibility and speech quality are low.

Chapter VI

Improved Objective Speech Quality Measurement Systems

In chapter 5, we identified issues with the two objective speech quality measurement systems (OSQMs) mentioned. In summary, these issues are that neither system takes consonantal and tonal intelligibility into serious consideration when calculating an objective quality score for Chinese speech. Therefore there is room for improvement in these systems to work well for Chinese speech.

Since the nature of consonantal and tonal intelligibility are different and the case concerning tonal intelligibility is not as well established or understood, we shall only deal with the consonantal intelligibility issue in this research. Two methods to improve the correlation between speech quality and consonantal intelligibility shall be introduced in this chapter. We shall discuss the basis of the methods, show the results, and analyse the results for each method.

6.1 Basis for improvement

It was mentioned that the relative intensity or power of consonants and the duration of some consonants are lower than those of the vowels (section 3.3.2). Therefore, the recognition of consonants will be lower on average than that of vowels by a human. Since the OSQMs we evaluate work in the perceptual domain, it could also mean that this brief and low power portion of speech may be neglected by the systems in their measurement of speech quality, or at least that the important weighting given to these regions is low. Furthermore, these systems are not the exact replicas of the human auditory system. What is sensitive to the human ears might not necessarily be sensitive to these systems. Humans can determine whether or not a piece of speech is intelligible even if the consonant that bears the intelligibility content is a minute fraction of power or duration of the whole

Chinese syllable. It is also notable that even in humans, exposure to certain vocal characteristics during the formative years allows a listener to easily detect small nuances in speech that other listeners may miss [61]. This is particularly true in Chinese with the recognition of tonal differences that non-Chinese listeners would miss. In fact, adult learners of new languages may themselves be familiar with the occasions where an important speech feature of the language they are learning is not actually discernible to them. Considering that such issues trouble even human listeners and speakers, it is not surprising that computational models are also unable to discern such minute but critically important linguistic nuances. OSQMs, therefore may not be so sensitive to differentiate which portion of speech is most important to intelligibility unless it was programmed to do so (as in the case of an objective speech intelligibility measurement system). Considering the fact that none of the available OSQMs have been designed with Chinese speech in mind, and in some cases may not even have been tested with Chinese speech, there is ample possibility of poor performance by the OSQMs. In this regard, it was noted that a processed Chinese syllable which was distorted or whose power has been attenuated such that it is almost completely unintelligible to a Chinese speaker may not noticeably influence a good OSQM quality reading.

With regard to the attenuation of a signal, loss of intelligibility occurs if the attenuation causes the consonant portion to drop below the hearing threshold. To the system, it might perhaps be an insignificant loss of power that has caused a minor degradation to speech quality. It may also be just a small fraction in terms of time duration that the processed speech signal falls below the hearing threshold. Also in the event of noise, intelligibility of a Chinese syllable might be lost if the noise power is sufficiently large to cause a masking effect on the consonants. Since vowels generally have higher intensities, a similar intensity of noise will probably not cause a serious detrimental effect to the recognition of the vowels. Besides masking by noise, higher intensities in the vowels could also produce a backward masking effect (please refer back to section 2.3.3) on the consonants that causes a reduction to intelligibility within the syllable itself. Similarly the loss of intelligibility due to masking may be neglected by an OSQM in its determination of speech quality.

In order to allow a higher sensitivity for the low power or short duration con-

sonants in the OSQMs, some signal processing techniques can be incorporated into the systems or applied to the original and processed signals. Due to intellectual right issues related to the patent-protected OSQM code, we will not alter the OSQM systems themselves, but shall apply signal processing techniques to the input signals to hopefully improve the correlation between speech quality and intelligibility for PESQ and MNB. Since the low power or short duration of consonants affects consonantal intelligibility, modifications to the duration or power of the consonant portion of a signal may enhance the OSQMs' sensitivity to this part. Modifications to these two areas will be discussed in the next few paragraphs.

When the duration of speech is altered, it can either result in a change of pitch or a change in tempo when the natural prosody of speech changes. This might reduce intelligibility if it is not done correctly. In a study conducted by Vaughan *et al.* [85], they conjectured that when the duration of speech is altered, it may reduce the recognition of speech (intelligibility) when this effect has been combined with other types of distortion, for example noise. They also suggested that when this is only applied to a selected portion of the phonemes say, the consonants of speech, an inconsistency with regards to overall prosody of speech (pitch or tempo) will occur. Hence intelligibility might be degraded rather than enhanced. In another study by Nejime and Moore [62], they found out that slowing the speech rate did not improve intelligibility in both simulations of hearing loss. Rather, significant reduction in intelligibility resulted in one condition. Although this case did not directly imply that a slowed speech rate will not improve intelligibility of speeches for people with normal hearing, the results of their findings relates that of [85]. Therefore, an alteration to the duration of the consonant is not recommended as it may not improve intelligibility in our case but might reduce intelligibility which in turn reduces the OSQMs' sensitivity to this part.

An increase of the amplitude or intensity (consonant gain) of the consonant could be a possibility for improvement. In the 1970s, it was realised that increasing the power of the speech signal relative to the level of noise increases intelligibility. However, when this method is used in high noise situations, it might damage the ears of the hearer as the amplitude of speech gets too loud. Therefore the enhancement of speech intelligibility in noise by infinite amplitude clipping was investigated by Thomas and Niederjohn [82][83]. They found out that infinite

amplitude clipping increases the consonant-vowel power ratio and subsequently improves speech intelligibility in noise. This was because the consonants, which are weaker in intensity but are more important to intelligibility, are the first to be masked by noise [63]. Using the technique of infinite amplitude clipping, intelligibility of speech can be improved in noisy situations, yet protecting the ears of the hearer from being damaged due to exceedingly loud speech signals. It should be noted here that such techniques, although successful for improving intelligibility, generally significantly reduce quality. In another study by Gordon-Salant [37], where she measured the recognition rate of speech in four conditions: normal speech, speech with consonant duration increased by 100%, speech with consonant-vowel ratio (CVR) increased by 10 dB, and speech with both consonant duration and CVR increased. She found out that the recognition of nonsense syllables was improved the most when the CVR was increased by 10 dB under various conditions including quiet. When a lower gain factor of 2 is used in [85] by Vaughan *et al.*, improvement was only seen under noisy conditions at a normal speaking rate. This was because the intelligibility of the original speech with no consonant gain was approximately 100% and that with gain was about 99%. Thus no improvement in the quiet condition cannot determine that this method is not advantageous. Original speech files that are less intelligible should be used to investigate this area. Taking the investigations reported in [37], where the intelligibility of original speech was not perfect, significant improvements were yielded using this method. Therefore, it would be justifiable to conclude that increasing the CVR increases intelligibility.

The above findings indicate that increasing the consonant-vowel ratio improves intelligibility both in noise and in quiet (noiseless) conditions. If we perform the same technique on the speech signals before inputting into PESQ or MNB, it may be possible that the systems could provide a higher sensitivity for the lower powered consonants and henceforth place more emphasis on the consonantal intelligibility in their computation of speech quality. Two signal processing methods are thus proposed to increase CVR:

1. High Pass Filtering, and

2. Consonant Amplification (Gain).

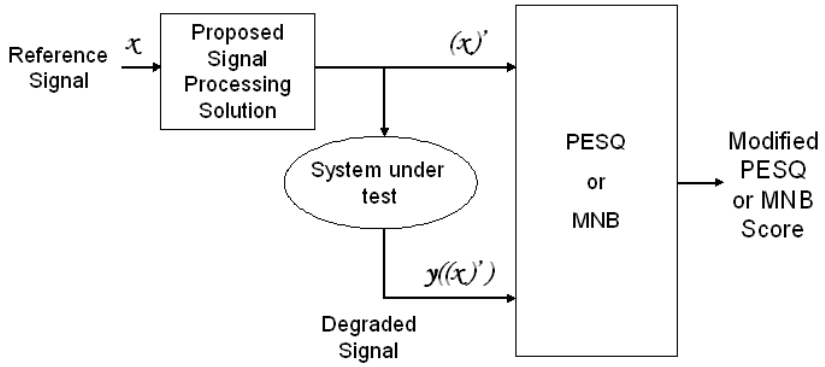

Figure 6.1: Structure of modified system where the proposed signal processing technique is applied to the original signal before being processed.

6.1.1 Point of application

Applying these two signal processing techniques to different points in the system would yield different effects. If they are applied to the original speech x before inputting into a speech processing system (figure 6.1), their effects will influence the processing of the speech by the system. The processed or coded signal would have been processed based on the original signal that was first manipulated by either signal processing technique resulting in $(y((x)'))$. The output in this case would not be similar to the intended one from the original speech, that is $y(x)$. If the techniques are applied individually to the original and processed signals, x and $y(x)$ (figure 6.2), the processed signal would be the intended one from the processing of the original signal. Hence, the signals to be input into the OSQM would be $(x)'$ and $(y(x))'$. In this case, signals that went through a similar signal processing technique $(\_)'$ would be compared in the OSQM. Hence, this method is adopted in our research.

6.2 Method 1 - High pass filtering

6.2.1 Introduction

Since most Chinese consonants are unvoiced and the frequencies of these unvoiced consonants are generally higher than those of the vowels which are voiced, a higher CVR can be obtained by attenuating some of the lower frequency energy

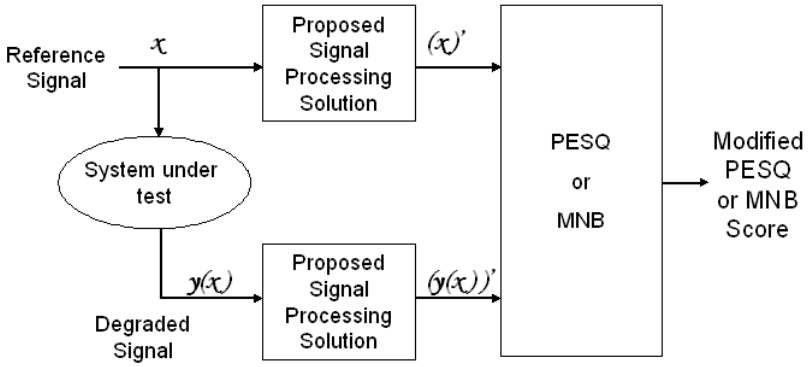

Figure 6.2: Structure of modified system where the proposed signal processing technique is applied to both the original and the processed signal individually. This is the method adopted in our research.

of the vowels. This can be done through high pass filtering above the predominant vowel frequencies. Section 3.3.3 mentioned that vowels can be determine by the first two formants with frequency ranges from approximately 270 Hz of the first formant (F1) to 2290 Hz of the second formant (F2). Therefore by attenuating frequencies below approximately 2 kHz, a better CVR can be achieved. Although vowels are not as susceptible as consonants to the reduction of intelligibility by attenuation since they have relatively higher amplitudes, the filtering process has to be done precisely as excessive attenuation will cause a reduction in vowel intelligibility. Therefore in our research, we used two cut-off frequencies for the high pass filters - 1 kHz (which is between F1 and F2) and 2 kHz (which is slightly lower than the highest of F2). For both cutoff frequencies, both 10 and 100 coefficient filters were tested so that the effects of the width of transition band in our proposed method can be noted. A *finite impulse response* (FIR) filter is used so that a linear phase response can be obtained as phase distortion is undesirable for the speech files. To enhance the effect of the finite response, the impulse response is set to zero after, say, *M* samples. When the rectangular method is incorporated to achieve this, undesired oscillations or peaks (Gibb's phenomenon) around the transition edge of the signal will arise due to discontinuities [40]. Therefore a *Bartlett window* with the simplest computation is used attenuate the signal gradually. The window is defined as follows:

Table 6.1: Correlation between amount of intelligibility degradation and quality scores for unfiltered/filtered Chinese syllables with noise and the percentage of improvement in correlation for filtered syllables over unfiltered. Averaged PESQ and MNB quality scores and the corresponding change in percentage.

Correlation Coefficients	**Unfiltered**	**1kHz HPF 10th order**	**1kHz HPF 100th order**	**2kHz HPF 10th order**	**2kHz HPF 100th order**
PESQ	-0.065	-0.083	-0.171	-0.134	-0.083
% improvement	—	28.1%[a]	162.6%	106.4%	28.1%
MNB	-0.068	-0.061	-0.038	-0.062	-0.011
% improvement	—	-10.5%[b]	-44.0%	-9.6%	-83.4%
Average PESQ or MNB scores					
PESQ	2.455	2.466	3.007	2.587	3.318
% change	—	0.4%	22.5%	5.4%	35.2%
MNB	1.976	2.070	2.787	2.481	2.940
% change	—	4.7%	41.1%	25.5%	48.8%

[a] All percentages were rounded to one decimal places and correlation coefficients listed in this table were rounded to three decimal places due to small figures.

[b] A negative means a degradation instead of an improvement.

$$w(n) = \begin{cases} \frac{2n}{N-1} & \text{, for } 0 \leq n < \frac{N-1}{2} \\ 2 - \frac{2n}{N-1} & \text{, for } \frac{N-1}{2} \leq n \leq N-1 \end{cases}$$

6.2.2 Results

The high pass filters (HPFs) were applied to the two sets of 192 CDRT speech files mentioned in section 5.3 for both noisy and quiet conditions. Sets of PESQ and MNB scores were computed for the filtered files and the Pearson's correlation between the amount of degradation from the subjective intelligibility test and objective quality scores were computed. These correlation coefficients are tabulated in tables 6.1 for noisy and 6.2 for quiet (noiseless) conditions together with the correlation coefficient for the unfiltered case.

Using equation 5.1 to perform a one-tailed *t*-test with $N-3$ degrees of free-

Table 6.2: Correlation between amount of intelligibility degradation and quality scores for unfiltered/filtered Chinese syllables without noise and the percentage of improvement in correlation for filtered syllables over unfiltered. Averaged PESQ and MNB quality scores and the corresponding change in percentage.

Correlation Coefficients	**Unfiltered**	**1kHz HPF 10th order**	**1kHz HPF 100th order**	**2kHz HPF 10th order**	**2kHz HPF 100th order**
PESQ	-0.087	-0.087	-0.103	-0.085	-0.049
% improvement	—	0.5%[a]	19.3%	-1.6%[b]	-43.1%
MNB	-0.150	-0.154	-0.132	-0.129	-0.061
% improvement	—	2.8%	-12.1%	-13.9%	-59.6%
Average PESQ or MNB scores					
PESQ	3.923	3.925	4.161	4.010	4.137
% change	—	0.1%	6.1%	2.2%	5.5%
MNB	3.846	3.798	3.697	3.653	3.712
% change	—	-1.2%[c]	-3.9%	-5.0%	-3.5%

[a] All percentages were rounded to one decimal places and correlation coefficients listed in this table were rounded to three decimal places due to small figures.

[b] A negative means a degradation instead of an improvement.

[c] A negative means a reduction in quality score instead of an improvement

dom to find out the significance of differences between the correlation for the unfiltered case r_{xy} and the filtered cases r_{zy}, none of the improvements by the HPFs were significant at $\alpha = 0.05$ from both conditions. However, improvements were significant at the $0.01 \leq \alpha < 0.05$ level from the 1 kHz 100th order and 2 kHz 10th order HPFs for PESQ in the noisy case.

6.2.3 Discussions

Table 6.1 shows that there were improvements in correlation for PESQ by all four HPFs when they were applied to the signals with noise, although significant improvements were only shown in the 1 kHz 100th order and 2 kHz 10th order cases at $0.01 \leq \alpha < 0.05$. For MNB, however, none of the HPFs improve the situation but degradations occurred. For the quiet condition, only two out of four HPFs contribute to an improvement in PESQ and one out of four HPFs in MNB. None of the improvements was significant at both $\alpha = 0.05$ and $0.01 \leq \alpha < 0.05$ levels. In the case of the noisy speech files, since the higher noise energies are usually below 2 kHz (chapter 9 of [94], and [18]), a HPF with a cutoff frequency below this would eliminate most noise power as well as attenuating some of the power of vowels. When this happens, the overall signal-to-noise ratio (SNR) for speech files used in our research will be increased taking into account the attenuated low frequencies with respect to the higher frequencies as well as the increased CVR. The improvements in correlation between intelligibility and PESQ may suggest this. The result obtained for MNB is, however, unexpected. This again indicates that MNB looks at the signal with a different psychoacoustic "perspective" to PESQ. Referring to table II.1 of the ITU-T recommendation P.861 [43] and table 1 of P.862 [45], we can see that the condition where environmental noise is included has demonstrated acceptable accuracy in PESQ while sufficient information has not been obtained regarding the accuracy of this in MNB. Thus, if we assume that MNB unreliably computes speech quality for this condition and exclude its results, we can safely assume that this High Pass filtering technique does improve the sensitivity of an OSQM with regards to speech intelligibility for Chinese speech from the results obtained for PESQ for this case.

Levitt mentioned in [55] that there will be some loss in intelligibility when a HPF eliminates frequencies in the region where the SNR is positive. Since power

of Chinese consonants are usually lower than that of vowels, SNR in the region of the vowels would be more positive. Therefore, the loss in intelligibility would mostly occur in vowels in our case. If significant amount of vowel energy is attenuated together with noise, the reduction in vowel intelligibility would also be significant. This could perhaps explain the case why the 2 kHz 100th order HPF causes degradations in three out of four cases in both conditions compared to the 10th order one which causes lower amounts of degradation, and a significant improvement in PESQ for the noisy condition. As the 100th order filter has a narrower transition band and hence a sharper impulse response, much of the energy below 2 kHz would have been attenuated. This would eliminate all the F1s and most of the F2s and would reduce vowel intelligibility greatly. Therefore correlation deteriorates when the resulting reduction in quality score (or worst an increase in quality that arises from the attenuated signals) did not match that of intelligibility. Since correlation deteriorates in most cases and the improvement was not significant for the 2 kHz 100th order high pass filtering, we shall not consider this HPF as one that can improve the situation, and will omit this filter from the rest of this discussion section. The reason for the insignificance in improvement by the 1 kHz 10th order HPF could be due to the gradualness of the filter impulse or transition response. This is a case of an "insufficient" filtering that results in an insignificant increase in the CVR. This filter should also be excluded as one that can improve CVR and hence correlation.

For the noiseless condition, the greatest improvement resulted from the 1 kHz 100th order high pass filtering process. However, this improvement was not significant. Since noise was not added, leaving the signal with ambient (background) noise, SNR is positive in this case. Referring back to the reduction in vowel intelligibility in regions where SNR is positive that was recently mentioned, we realised that there might be a trade-off for an improvement in CVR at the expense of vowel energy which might affect intelligibility for this method. Since vowel energies are generally higher than consonants', the intelligibility of vowels could still be preserved as long as its attenuation has not reached a certain threshold value (which we cannot conclude from our findings). If we assume that PESQ and MNB also cannot accurately determine this threshold value, and both systems have different allowances regarding the limit of attenuation of vowel energies, this explains why

improvements were insignificant, and also explains the degradation that occurred for this condition.

By looking at the changes in quality score for both PESQ and MNB in tables 6.1 and 6.2, it was noted that objective quality scores from PESQ and MNB increase when high pass filtering is applied to noisy files, while changes were not so prominent in the noiseless situation. There is no doubt that high pass filtering can remove some noise and therefore improve speech quality. From the improvements (more negative correlation in our case) in the correlation for PESQ in the noisy situation, we can further conclude that the high pass filtering method indeed improves PESQ's sensitivity towards intelligibility in this condition. The reason for this is when correlation improves (becomes more negative), this means that the decrease in the amount of degradation in intelligibility, which is the Y-axis in figures 5.2 to 5.5 (or the number of CDRT errors which is the Y-axis in figures 5.6 to 5.8), actually led to an improvement in speech quality. Thus, we can see that sensitivity towards consonantal intelligibility increases. The marginal increase in quality score in PESQ that led to an insignificant improvement (in some cases degradation) in correlation in the noiseless case (which is contrary to the trend for PESQ in the noisy situation) can again be explained by Levitt's point - that a reduction in (vowel) intelligibility arises when a HPF eliminates frequencies in the region where the SNR is positive. In the noiseless case, the SNR would indeed be higher than that of the noisy case. Therefore, when high pass filtering is applied, (vowel) intelligibility may be slightly reduced hence causing a slight degradation in correlation when quality increases instead of an improvement as in the noisy case. This reduction in intelligibility, however, is not severe since the quality or correlation only changes slightly. The decrease in quality from MNB may also suggest that the integrity of the vowel intelligibility is affected. From these, we realised that the high pass filtering method is not so effective when speech files were of a certain quality (high SNR) without the influence of noise.

Although significant improvements in correlation did not result from many of the tested HPFs, improvements can be inferred from the objective quality score which is not subjective (not subject to human error). Looking at the differences in quality scores for syllables with and without errors from table 6.3, we can see that PESQ scores for syllables without errors yielded a greater improvement

than those with errors in general (disregarding the 2 kHz 100th order case) for both conditions. Two out of three (again disregarding the 2 kHz 100th order case) HPFs caused a marginally greater decrease in MNB scores for the case with errors in the noiseless condition. These results generally show that the high pass filtering method exposes the discrepancy between the original and processed consonants which are sometimes neglected in the determination of quality. After exposing or magnifying the consonants, those consonants with discrepancies that led to an intelligibility error should result in a quality change that is of a lower quality than the change in those without or with less discrepancies that did not cause an error. Hence, syllables with intelligibility errors yielded a quality declination that was greater or an improvement that was smaller than those without intelligibility errors as the discrepancies in the consonants for the lower intelligible syllables are magnified.

The advantage of the high pass filtering method is that it is easy to implement. The whole speech file could be signal processed without having to select any part of the signal. Inaccuracies that arise from the selection process can also be eliminated. The effectiveness of this method however is also non-optimal since there is a trade-off in increase of CVR at the probable expense of reducing vowel intelligibility. Phase distortion might also occur if care is not taken in the design of filters. The degree, and cutoff frequency of the filters could probably be adapted to particular recordings of speech to derive a more optimal solution, although this is outside the scope of this thesis.

6.2.4 Conclusions

The high pass filtering method was proposed to improve the correlation between speech quality and consonantal intelligibility. The basis for this method is to improve the consonant-vowel ratio by attenuating the lower frequency vowel energies so that PESQ or MNB can be more sensitive to consonants with lower intensities and therefore pay more emphasis on them in the computation of speech quality. It was realised that MNB may not be sufficiently accurate to compute speech quality for processed speech files with noise as mentioned in the ITU recommendation [43]. Therefore, based on the proven accuracy of PESQ mentioned in its recommendation [45] for the same condition, the high pass filtering method was

Table 6.3: Increase in averaged PESQ and MNB scores (%) caused by the HPFs.

With Noise	**1kHz HPF 10th order**	**1kHz HPF 100th order**	**2kHz HPF 10th order**	**2kHz HPF 100th order**
PESQ (Error)[a]	0.4%	20.6%	4.6%	37.1%
PESQ (No Error)[b]	0.4%	23.0%	5.2%	34.4%
MNB (Error)	5.7%	50.8%	30.2%	61.6%
MNB (No Error)	4.8%	39.7%	25.3%	46.7%
Without Noise				
PESQ (Error)	0.1%	6.7%	2.6%	7.1%
PESQ (No Error)	0.1%	7.6%	3.7%	6.7%
MNB (Error)	-2.5%[c]	-4.4%	-6.3%	-3.5%
MNB (No Error)	-3.5%	-3.9%	-4.7%	-2.3%

[a] The averaged quality score for syllables with intelligibility errors were obtained by multiplying the respective quality scores with its number of errors and then obtained the average of the sum of these multiplied scores. This is to account for the weightage of the syllables according to their number of errors.

[b] The averaged quality score for syllables with no intelligibility errors were obtained by averaging the sum of their quality scores.

[c] A negative % means a decrease in quality score. All percentages were rounded to 1 decimal place.

shown with slightly over 90% confidence that it is effective in improving the correlation between speech quality and consonantal intelligibility in noise. However, careful selection of filter parameters is required as "insufficient" filtering would not result in any significant improvement while "excessive" filtering affects vowel intelligibility.

In the noiseless condition, improvements were insignificant and degradations occurred. This is again due to the reduction in vowel intelligibility when SNRs were positive and higher. Therefore, the high pass filtering method is not effective in this condition.

In both conditions, smaller improvements or greater declinations in quality scores for syllables with intelligibility errors after filtering showed that this method magnify the discrepancies in the consonants. Thus there is also some merit in this method although significant improvements was evident only in the noisy situation.

6.3 Method 2 - Consonant amplification

6.3.1 Introduction

To increase CVR, one can either attenuate the amplitude of vowels as in the high pass filtering method or increase the amplitude of consonants. The filtering method increases the CVR with a probable trade-off of reducing vowel intelligibility. In this second method, the consonants of the Chinese syllables were amplified without attenuating any parts of the signal with the aim that loss of intelligibility be avoided. However ample caution has to be taken during the amplification process to avoid distortion that arises from discontinuities in the signal (the Gibb's phenomenon previously mentioned in the last section). Therefore, the amplification is smoothed/graduated by half windowing preceding and proceeding the duration of the amplification section. Please refer to fig 6.3 for the windowed amplification process. Another factor to take note of is the degree of amplification. Too little would yield insignificant results while too much would degrade speech quality since it would introduce audible distortion and may even be damaging to human ears (chapter 4 of [61]) when the consonant becomes enormously loud. To avoid insufficient or excessive amplification for certain consonants, peak amplification factors of 1.5, 2, 4, and 8 times were determined to be appropriate

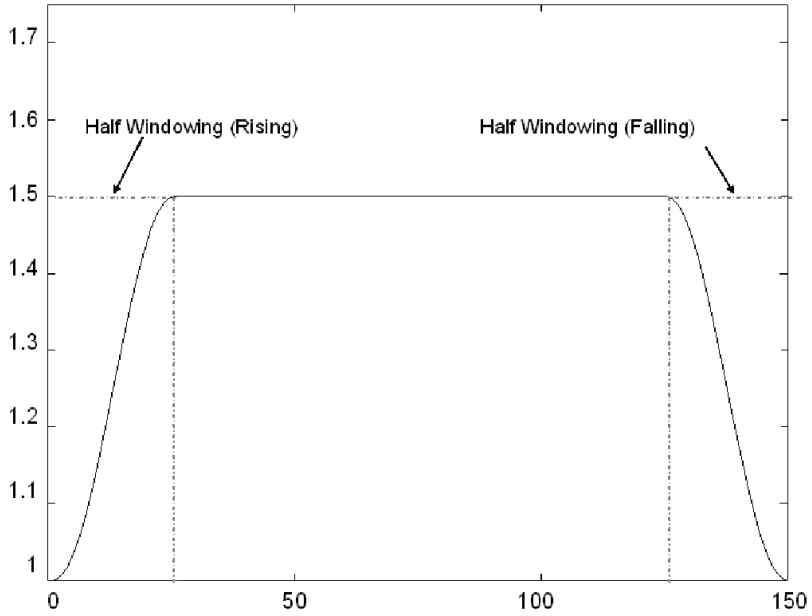

Figure 6.3: Windowed (smoothed) amplification process for factor 1.5 times.

after brief initial tests, and hence were adopted. To ensure accuracy, the start and end points were determined MANUALLY[1] from listening and previewing of the enlarged plotted signal.

6.3.2 Results

The initial consonants of the two sets (noisy and noiseless) of 192 CDRT speech files were amplified and their respective PESQ and MNB quality scores computed. Correlations between the amount of degradation from the subjective intelligibility test and objective quality scores were then calculated. Results were displayed in tables 6.4 and 6.5 for the noisy and noiseless condition respectively.

Again applying equation 5.1 mentioned in section 5.2.1 with $N-3$ degrees of freedom to find out the significance of differences between the correlation of the

[1] Without doubt, an eventual aim is to incorporate the methods into an automated process. However, to ensure that the accuracy of our result is independent of the accuracy of any automated consonant selection process, the manual listening and viewing process, despite being exceedingly tedious and long-winded for several hundred recordings, was performed. Note that reported consonant-vowel segmentation accuracy of up to 95.4% was achieved for Chinese syllables in [20].

Table 6.4: Correlation between amount of intelligibility degradation and quality scores for unamplified/amplified Chinese syllables with noise and the percentage of improvement in correlation for amplified syllables over unamplified. Averaged PESQ and MNB quality scores and the corresponding change in percentage.

Correlation Coefficients	**Unamplified**	**1.5 X C1**	**2 X C1**	**4 X C1**	**8 X C1**
PESQ	-0.065	-0.095	-0.142	-0.119	-0.118
% improvement	—	46.8%[a]	118.2%	83.2%	81.2%
MNB	-0.068	-0.082	-0.091	-0.089	-0.057
% improvement	—	20.6%	33.5%	30.4%	-17.1%[b]
Average PESQ or MNB scores					
PESQ	2.455	2.405	2.331	2.126	1.999
% change	—	-2.1%[c]	-5.0%	-13.4%	-18.6%
MNB	1.976	1.978	1.980	2.056	1.971
% change	—	0.1%	0.2%	4.1%	-0.3%

[a] All percentages were rounded to one decimal places and correlation coefficients listed in this table were rounded to three decimal places due to small figures.

[b] A negative in percentage means a degradation instead of an improvement.

[c] A negative means a reduction in quality score instead of an improvement.

Table 6.5: Correlation between amount of intelligibility degradation and quality scores for unamplified/amplified Chinese syllables without noise and the percentage of improvement in correlation for amplified syllables over unamplified. Averaged PESQ and MNB quality scores and the corresponding change in percentage.

Correlation Coefficients	**Unamplified**	**1.5 X C1**	**2 X C1**	**4 X C1**	**8 X C1**
PESQ	-0.087	-0.087	-0.098	-0.096	-0.115
% improvement	—	0.8%[a]	13.2%	11.2%	32.8%
MNB	-0.150	-0.177	-0.177	-0.139	-0.072
% improvement	—	17.9%	18.0%	-7.5%[b]	-51.7%
Average PESQ or MNB scores					
PESQ	3.923	3.890	3.862	3.817	3.781
% change	—	-0.8%[c]	-1.5%	-2.7%	-3.6%
MNB	3.846	3.805	3.785	3.715	3.637
% change	—	-1.05%	-1.6%	-3.4%	-5.4%

[a] All percentages were rounded to one decimal places and correlation coefficients listed in this table were rounded to three decimal places due to small figures.

[b] A negative in percentage means a degradation instead of an improvement.

[c] A negative means a reduction in quality score instead of an improvement.

original (unamplified) case r_{xy} and the amplified cases r_{zy}, the 2 X amplification factor causes a significant improvement at $\alpha = 0.05$ for PESQ in the noisy condition. An improvement was significant at the $0.01 \leq \alpha < 0.05$ level from the 1.5 X factor for MNB in the noiseless case.

6.3.3 Discussions

As shown in table 6.4, all four amplification factors improve correlation for PESQ in the noisy condition and three out of four caused improvements in MNB. However, only the 2 X factor which causes an improvement of 118.2% in PESQ was statistically significant. Similar to the high pass filtering method, MNB's accuracy is doubtable in this condition (please refer to section 6.2.3), therefore the results arising from signal processing using this method cannot be deemed accurate for MNB under this condition. Hence it is justifiable to state that this method is indeed effective in increasing the sensitivity of consonantal intelligibility in an OSQM based on the results obtained for PESQ. Although only the 2 X factor's improvement was significant, improvements in percentage seen in the 4 X and 8 X factors were remarkable at levels higher than 80%. We also noticed a huge step in improvement between 1.5 and 2 X. This could be due to inadequate amplification that did not fully illustrate the advantage of this method in the 1.5 X. This is specially so in noisy conditions when SNR is generally low. Hence the full advantage could only be manifested when amplification surpasses a certain level. In this case, it was noted that correlations between the 4 X and 8 X factors were very close. We have mentioned that excessive amplification can cause the sound to be annoying to the human ears and hence yields a lower quality score. When this happens, a trade-off between an improvement in CVR and decrease in speech quality due to excessive loudness appears. Although an improvement is still present in the case of 8 X, it might not be true when amplification factors are larger or in other conditions when SNR is high. Therefore, there also exists a loudness threshold in various conditions where exceeding it will cause a declination in speech quality due to loudness level to exceed the improvement in CVR. This means to say the improvement in speech intelligibility leads to a declination in quality. This is also true when the amplitude of the overall syllable is generally high. The loudness threshold for a human is about 140 dB SPL (section 2.3.1)

exceeding which would cause great annoyance to the ears. However for some people, annoyance already exists below this level. An example of this would be loudness levels of some rock bands (approximately 110 dB in section 3.3.5) which causes annoyance in the ears of some people.

From table 6.5, it was shown that amplification factors of 2 X and greater brought forth improvements to the correlations for PESQ in the noiseless situation and for MNB, improvements for the 1.5 X were significant at $0.01 \leq \alpha < 0.05$ and 2 X were very close to this significance level. 4 X and 8 X did not improve the situation. Improvements for PESQ in this condition were not as great as for the noisy one. This could be reasoned by the fact that besides the increase in CVR in the noisy condition, the signal (consonant) to noise ratio was also increased, reducing the masking effect of noise to consonantal intelligibility. This led to a double advantage. Results obtained for MNB, however, were the opposite of PESQ's. The 1.5 and 2 X brought forth better improvements in MNB while the larger improvement was seen in 8 X for PESQ. Once again, the difference in "perception" between PESQ and MNB was seen. This method was seen to produce the best results in MNB for this noiseless condition. Nevertheless, some improvements may have occurred in PESQ.

Although most of the factors did not result in significant improvements in correlation, improvements can be seen in the objective quality score which is not subjective (independent of human errors). Considering the quality scores in table 6.6, beside scores computed by MNB in the noisy condition, quality scores decrease with increasing amplification for both systems in both conditions. This change is more gradual in the noiseless condition. Since the amplitude of consonants were generally lower, differences between the consonant of the original and processed syllable would be minute. Hence, the OSQMs may not be that sensitive to detect this minute difference between the consonants. When the consonants were amplified, the discrepancy between the original and processed consonant would be more prominent. Speech quality will therefore be reduced considering this magnified discrepancy for Chinese syllables with intelligibility errors. The degradation of quality caused by discontinuities in the signal due to amplification will be minimal and hence disregarded in our case. The reason is firstly, the amplification process was smoothed with half windowing preceding and proceeding the

Table 6.6: Changes in averaged PESQ and MNB scores (%) caused by consonant amplification.

With Noise	**1.5 X C1**	**2 X C1**	**4 X C1**	**8 X C1**
PESQ (Error)[a]	-2.6%[b]	-7.4%	-17.5%	-22.3%
PESQ (No Error)[c]	-1.9%	-4.9%	-13.4%	-19.7%
MNB (Error)	0.1%	-0.5%	4.9%	5.8%
MNB (No Error)	0.2%	0.5%	4.2%	-2.0%
Without Noise				
PESQ (Error)	-1.0%	-2.2%	-4.7%	-7.3%
(PESQ No Error)	-0.1%	-0.6%	-1.8%	-4.6%
MNB (Error)	-2.3%	-3.1%	-4.3%	-5.0%
(MNB No Error)	-0.8%	-1.2%	-3.3%	-5.9%

[a] The averaged quality score for syllables with intelligibility errors were obtained by multiplying the respective quality scores with its number of errors and then obtained the average of the sum of these multiplied scores. This is to account for the weightage of the syllables according to their number of errors.

[b] A positive % means an improvement in quality score. All percentages were rounded to 1 decimal place.

[c] The averaged quality score for syllables with no intelligibility errors were obtained by averaging the sum of their quality scores.

consonant to be amplified. Secondly, since the correlation between intelligibility improves (more negative), this means to say that the decrease in speech quality is due to an increase in the amount of degradation in intelligibility of the processed signal. This point thereby confirms our reasoning of magnifying the discrepancies between the signals.

It was shown in figure 6.4 that the 2 X amplification factor improved correlation for both OSQMs in both conditions while the 1.5 X and 4 X factors improved three out of four situations. Considering statistical significance, the 2 X factor causes one significant improvement at the $\alpha = 0.05$ level for PESQ in noise and close to the $0.01 \leq \alpha < 0.05$ level for MNB in quiet. The 1.5 X also produces a

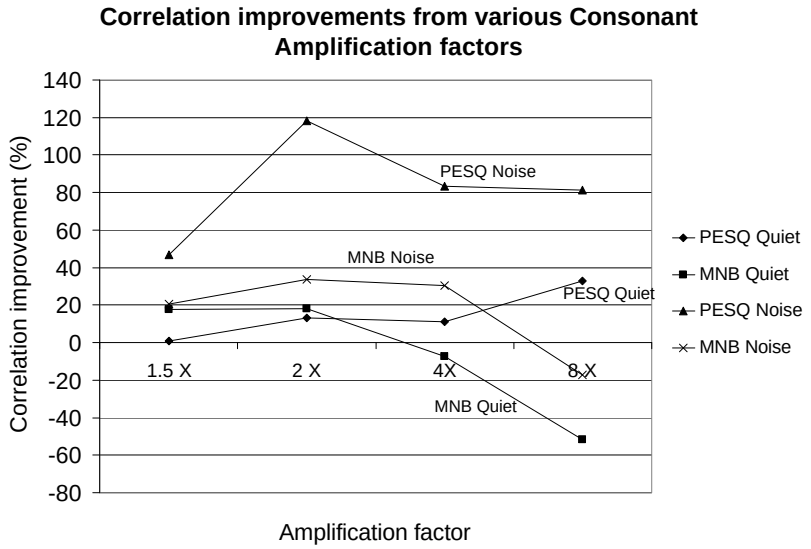

Figure 6.4: Correlation improvements from the 4 consonant amplification factors.

Table 6.7: Average improvements from both OSQMS in both conditions caused by consonant amplification.

1.5 X C1	**2 X C1**	**4 X C1**	**8 X C1**
21.5%	45.7%	29.3%	11.3%

significant improvement for MNB in quiet at the $0.01 \leq \alpha < 0.05$ level. Together with the average improvements shown in table 6.7, the 2 X amplification factor are recommended to improve correlation between speech quality and intelligibility. The 1.5 and 4 X factors could also be considered in specific conditions for a specific OSQM.

The advantages of this method lie in the simplicity of arithmetic calculations because only multiplications are required. However, processing time is compromised by the consonant selection process even if this is done automatically, as of course will the effectiveness of the technique. Phase distortion is unlikely to occur as in the filtering method because processing is performed linearly. No degradation of intelligibility is likely to arise because no parts of the signals are attenuated in this process and when amplification is not excessive. The disadvan-

tages firstly lie in the consonant selection process as the efficacy of this method relies on the accuracy of the selection when it is automated. Secondly, distortions due to discontinuities (Gibb's phenomenon) might arise if the amplification and deamplication processes are not done gradually at the start and end points of the consonants.

6.3.4 Conclusions

The second method of improving the correlation between speech quality and consonantal intelligibility was proposed in this section. The basis of improvement is that the CVR can be increased by increasing the amplitude of the consonants. No parts of the signals are attenuated in this method. However, extra caution has to be taken to eliminate discontinuities within the signal in the amplification process. It was also realised that while too little amplification is insufficient to manifest the effectiveness of this method, excessive amplification will lead to a degradation instead. In our research, four amplification factors of 1.5 X, 2 X, 4 X, and 8 X were used. Disregarding the MNB results in the noisy condition, this method was shown to be effective in increasing the sensitivity of an OSQM in general. Significant improvements were seen by the 2 X factor for PESQ in the noisy condition and by the 1.5 X factor for MNB in the noiseless. Averaged improvements also showed the 2 X factor produces the best overall improvement followed by 4 X and 1.5 X. The 2 X factor is therefore recommended as it yielded the highest overall improvements.

6.4 Conclusions on the improvements made to the consonantal intelligibility problem

To resolve the problem of a low correlation between consonantal intelligibility and quality of Chinese speech, two signal processing methods were proposed. The basis for the efficacy for these two methods is the increase in the consonant-vowel ratio because the enhancement in this ratio leads to a higher sensitivity for consonantal intelligibility which will increase its correlation with speech quality. Either one of the methods is to be applied individually on both the original and processed signal before they were input to an OSQM for the computation of an

objective quality score. In this way, the intended processed signal will be obtained from the sound processing system instead of the one where its original signal is first processed by our proposed methods.

The first method is the *high pass filtering* method whereby CVR is increased by attenuating the vowel energies. It was shown that significant improvements were yielded from this method in the noisy condition whereas its effect was minimal in the noiseless condition.

The *consonant amplification* method was proposed secondly. CVR is enhanced by amplifying the consonant that is relatively lower in energy. Comparing it to the first method, although the highest improvement of 162.6% (table 6.1) was obtained from the first method, the only significant improvement at the $\alpha = 0.05$ level appeared in the second method which was also more consistent in enhancing correlations in both conditions for both OSQMs. It was realised that while insufficient amplification cannot bring about the full efficacy of this method, excessive amplification will lead to a declination in correlation. Thus, the amplification factors of 2 X is recommended.

Chapter VII

Conclusions and Future work

7.1 Conclusions

Since the worldwide population of Chinese speakers is enormously significant, an objective speech quality measurement system (OSQM) suitable for this language to assess speech quality transmitted or processed through telephony, networks, and other speech communication systems is desirable. This is complicated by the fact that there are certain characteristics of Chinese speech that are not found in English and most other European languages which we have found affect the accuracy of existing OSQMs measuring quality of Chinese speech. In the previous chapters, we evaluated two OSQMs with regard to their assessment of the quality of Chinese speech output from example sound processing systems to demonstrate this claim.

Providing context for this research, the chapter on hearing gave an introduction to the human auditory process to aid understanding of how a perceptual model adopted by an advanced OSQM works. This chapter started off introducing the physiology of the human ear consisting of the peripheral and neural processing regions. Later, the psychological aspects of human hearing were expounded. The concepts of *loudness perception*, *critical band*, *masking*, and *pitch perception* were discussed and related to elements of speech processing and quality evaluation.

Chapter 3 discussed various issues regarding speech. Firstly, the process of speech production, then the characteristics of speech produced by the two essential processes of *initiation* and *articulation*, and the optional *phonation* in certain speech sounds. This was followed by discussing characteristics of English speech where the Latin alphabet is used to denote phonemes. The *International Phonetic Alphabet* was then introduced which represents most, if not, all speech sounds,

including those elements of Chinese speech investigated in this research. The production of English consonants and vowels was also described proceeded by the loudness and frequency range of intelligible speech. It was realised that although the general frequency range for human speech is from 50 to 10,000 Hz, telephony systems are usually band limited to between 300 and 3400 Hz to capture most speech energy. While this range is adequate to represent the intelligibility of vowels, this is not always true for the many consonants with intelligible frequency components higher than 4000 Hz. Regarding the loudness of intelligible speech, we found that an average signal-to-noise ratio of at least +6 dB must be achieved so that it can be readily heard. After this, the influence of speech context to intelligibility was mentioned. Although, context improves intelligibility, there are cases where intelligibility is independent of context. Therefore, speech intelligibility at the individual word or syllable level is crucial to effective communication, and intelligibility testing at such level is necessary. Lastly, the language of interest in our research, Chinese, was introduced in the same chapter. Its unique CVC phonetic structure which creates 39 confusing vocabulary sets, and the use of four lexical and one neutral tone were specifically mentioned.

The terms speech intelligibility and quality were formally defined in chapter 4. The description of various speech quality and intelligibility measurement systems or tests were also given. The approaches to measure or test speech intelligibility and quality was categorised into *subjective* and *objective* tests. Subjective tests involve a pool of human subjects to rate intelligibility or quality while objective tests involve computerised mathematical calculations of physical properties of speech to compute a rating score. Finally, the tests or systems involved in this research were discussed. They are the *CDRT* and *CDRT-Tone* subjective tests for testing the intelligibility of Chinese speech, and the PESQ and MNB objective speech quality measurement systems.

Underpinning the main part of our work, the relationships between speech quality and intelligibility were defined in chapter 5. They are:

1. When intelligibility is held constantly at a high level, speech quality cannot be predicted with confidence from a measure of intelligibility, i.e. speech quality can be high or low.

2. When intelligibility varies, speech quality tends to correlate with speech intelligibility in that:

   (a) high intelligibility generally yields a higher quality score, and

   (b) low intelligibility generally yields a lower quality score.

The two objective systems involved were then evaluated using particularly the second relationship. In the evaluation, two types of Chinese speech intelligibility were identified: *consonantal* and *tonal* intelligibility. From the evaluation, it was revealed that correlation between intelligibility and quality were low in both cases (consonantal and tonal). To resolve the low correlation between consonantal intelligibility and quality, two methods namely the *high pass filtering* method and the *consonant amplification* method were proposed and evaluated in chapter 6. The theory behind both methods was the improvement of the consonant-vowel ratio (CVR). Although CVRs were improved by both methods, the improvements were more evident in the latter. This was because the high pass filtering method improves CVR at a probable expense of reduction in vowel intelligibility while consonant amplification does not. Therefore, the finding in our research is that the consonant amplification method was evident to improve the sensitivity of the consonantal intelligibility in the computation of speech quality by the OSQMs.

7.2 Recommendations for Future Work

From this research, several issues were noted which prompt for related future work:

1. The issue of the low correlation between tonal intelligibility and Chinese speech quality was not resolved in this research since Chinese tones (or in fact any tonal intonation) were not considered in the design of current OSQMs. A new OSQM or model must therefore be developed to account for this aspect. Although the current work was specifically charged with analysis of existing OSQMs, we have concluded that, in order for a reliable high performance OSQM system to account for tone, there need to be additions to, and perhaps changes from, the existing psychoacoustic models.

Again, the worldwide economic and social importance of Chinese speech is growing rapidly; the proportion of world telecommunications users speaking Chinese is such that this has become an overdue research area.

2. Research can be performed to develop a *universal* or *multilingual* speech quality measurement system that works well for all languages. The reason behind this is that OSQMs have only been designed and tested in English and perhaps some European or Asian languages (generally French, German, Spanish and Japanese). These languages, however, do not exploit the full range of capability in the human speech production system. This means to say that, linguistically speaking, the complete range of speech features were not tested by these objective systems. This calls for extensive testing on these systems for the complete range of speech articulation features, and hence using the results to aid developing this multilingual system.

3. After conducting the subjective CDRT and CDRT-Tone tests, it was evident that some refinements are required to both tests to improve their effectiveness. An area to point out is the corpus of Chinese characters used. As some of the characters found in the test corpus were rarely used in common literature or speech, the visual recognition of these characters can be erroneous which might affect the credibility of the results from these tests. Similarly, the corpus can be printed according to the background of the subjects, for example, traditional Chinese script for Taiwanese and some Malaysian Chinese, and simplified Chinese script for Chinese from mainland China, Singapore, some South East Asian Chinese, and so on.

4. It was also evident that the condition of temporal clipping of speech mentioned in section 5.4.2 was not appropriately dealt with by MNB and PESQ. Since this condition is not uncommon in speech transmission or processing systems, it will be beneficial to evaluate OSQMs further under these conditions. This issue should probably also be considered in the design of new OSQMs.

References

[1] Full Chart of the International Phonetics Alphabet. http://www2.arts.gla.ac.uk/IPA/fullchart.html.

[2] International Phonetics Association. http://www2.arts.gla.ac.uk/IPA/index.html.

[3] The Taiwan Tongyong romanisation website. http://abc.iis.sinica.edu.tw/.

[4] *Oxford Advance Learner's English-Chinese Dictionary*, 4 ed. Oxford University Press, 1994.

[5] *Times New Chinese English Dictionary*. Federal Publications (Singapore) Pte Ltd, 1997.

[6] *Dictionary of Chinese Language*. Ministry of Education, the Mandarin Promotion Council of Taiwan, April 1998. http://140.111.34.46/dict/.

[7] *Cambridge Advance Learner's Dictionary*. Cambridge University Press, 2003.

[8] *Taiwan yearbook*. Government Information Office, Taiwan, 2004. http://www.gio.gov.tw/taiwan-website/5-gp/yearbook/P021.htm#3.

[9] AFIFI, A. A., AND AZEN, S. P. *Statical Analysis - A Computer Oriented Approach*, 2 ed. Academic Press, 1979, p. 140.

[10] AGRAWAL, A., AND LIN, W. C. An online speech intelligibility measurement system. *IEEE Transactions on Acoustics, Speech, and Signal Processing (ASSP-22) 22*, 3 (1974), 203–206.

[11] AMERICAN NATIONAL STANDARDS INSTITUTE. *ANSI S3.2-1960: Method for measurement of monosyllabic word intelligibility*, 1960.

[12] AMERICAN NATIONAL STANDARDS INSTITUTE. *ANSI S2.3-1989: American national standard method for measuring intelligibility of speech over communication systems*. Acoustical Society of America, 1990.

[13] AMERICAN NATIONAL STANDARDS INSTITUTE. *ANSI S3.5-1997: Methods for the calculation of the speech intelligibility index*, 1997.

[14] AMERICAN STANDARDS ASSOCIATION. *Acoustical Terminology SI, 1-1960*, 1960.

[15] BEERENDS, J. G. Improvement of the P.861 perceptual speech quality measure. Tech. rep., International Telecommunicaion Union, ITU-T SG12 COM-20E, 1997.

[16] BEERENDS, J. G., HEKSTRA, A. P., RIX, A. W., AND HOLLIER, M. P. PESQ - The new ITU standard for end-to-end speech quality assessment part II - Psychoacoustic model. *The Journal of the Audio Engineering Society 50*, 10 (October 2002).

[17] BLALOCK JR, H. M. *Social Statistics*, 2 ed. McGraw-Hill Book Company, 1972.

[18] BUNA, B. *Transportation Noise Reference Book*. Butterworth and Co. (Publishers) Ltd., 1987, ch. 6.

[19] CATFORD, J. C. *A Practical Introduction to Phonetics*. Oxford University Press, 1994.

[20] CHEN, S. H., AND WANG, J. F. Application of wavelet transforms for C/V segmentation on Mandarin speech signals. In *IEE Proceedings - Vision, Image and Signal Processing* (April 2001), vol. 148, IEE, pp. 133–139.

[21] CHONG, F., PAWLIKOWSKI, K., AND MCLOUGHLIN, I. Evaluation of ITU-T G.728 as a voice over IP codec for Chinese speech. In *Proceedings of the Australian Telecommunications Networks and Applications Conference 2003 (ATNAC 03')* (8-10 December 2003).

[22] CONWAY, A. E. A passive method for monitoring voice-over-IP call quality with ITU-T objective speech quality measurement methods. In *Proceedings of the 2002 IEEE International Conference on Communications (ICC 2002)*, vol. 4, pp. 2583–2586.

[23] CORBISIER, C. Living with noise. In *Public Roads Magazine*, vol. 67. U.S Department of Transportation. Federal Highway Administration, Jul-Aug 2003. `http://www.tfhrc.gov/pubrds/03jul/06.htm`.

[24] DAI, M., YU, K., XU, B., AND YU, C. Low SNR robust Chinese tone extraction based human auditory model. In *Proceedings of the 5th International Conference on Signal Processing (WCCC-ICSP 2000)* (21-25 August 2000), vol. 2, IEEE, pp. 752–755.

[25] DING, Z., MCLOUGHLIN, I. V., AND TAN, E. C. Intelligibility evaluation of GSM coder for Mandarin speech using CDRT. *Speech Communication 38* (2002), 161–165.

[26] DING, Z. Q., MCLOUGHLIN, I. V., AND TAN, E. C. Extension of proposal of standards for intelligibility tests of Chinese speech: CDRT-tone. In *IEE Proceedings - Vision, Image Signal Processing* (February 2003), vol. 150, IEE, pp. 1–5.

[27] DYNASTAT. Summary of speech intelligibility testing methods. `http://www.dynastat.com/SpeechIntelligibility.htm`.

[28] DYNASTAT. Summary of speech quality testing methods. `http://www.dynastat.com/SpeechQuality.htm`.

[29] EGAN, J. P. *Articulation testing methods*, vol. 11 of *Benchmark Papers in Acoustics - Speech Intelligibility and Speaker Recognition*. Dowden,

Hutchinson and Ross, Inc., 1977, ch. 15, pp. 175–202. Reprinted from pp. 955-981 of American Laryngological, Rhinological and Otological Society 58(9), 1948.

[30] ENCYCLOPEDIA BRITANNICA FROM ENCYCLOPEDIA BRITANNICA PREMIUM SERVICE. Sino-Tibetan languages. `http://www.britannica.com/eb/article?tocId=9109793`.

[31] ERSKINE, C., AND AMMON, S. Implementation of the floating point ITU-T G.728 speech coding algorithm on a 24 bit fixed point DSP.

[32] EURESCOM STAFF. Project p603 - quality of service: Measurement method selection (deliverable 2). Tech. rep., EURESCOM, 1997.

[33] EUROPEAN TELECOMMUNICATIONS STANDARDS INSTITUTE. *ETS 300 580-2: Digital cellular telecommunications system (Phase 2); Full rate speech; Part 2: Transcoding (GSM 06.10 version 4.2.1)*, Dec 2000.

[34] FRENCH, N. R., AND STEINBERG, J. C. *Factors governing the intelligibility of speech sounds*, vol. 11 of *Benchmark Papers in Acoustics - Speech Intelligibility and Speaker Recognition*. Dowden, Hutchinson and Ross, Inc., 1977, ch. 12, pp. 128–152. Reprinted from pp. 90-119 of Journal of Acoustical Society of America, 19(1), 1947.

[35] GABRIELSSON, A., AND SJOGREN, H. Perceived sound quality of sound-reproducing systems. *Journal of the Acoustical Society of America 65*, 4 (April 1979), 1019–1033.

[36] GOLD, B., AND MORGAN, N. *Speech and Audio Signal Processing*. John Wiley and Sons, Inc., 2000, ch. 14.

[37] GORDON-SALANT, S. Recognition of natural and time/intensity altered cvs by young and elderly subjects with normal hearing. *Journal of Acoustical Society of America 80*, 6 (1986), 1599–1607.

[38] HANSEN, J. H. L., AND NANDKUMAR, S. Objective speech quality assessment and the RPE-LTP coding algorithm in different noise and language conditions. *Journal of the Acoustical Society of America 97*, 1 (January 1995), 609–627.

[39] HARRINGTON, J., AND CASSIDY, S. *Techniques in Speech Acoustics*. Kluwer Academic Publishers, 1999, ch. 4.

[40] IFEACHOR, E. C., AND JERVIS, B. W. *Digital Signal Processing - A Practical Approach*. Addison-Wesley Publishers Ltd., 1993, ch. 6.

[41] ITU-T RECOMMENDATION. *G.728:Coding of speech at 16 kbit/s using Low-delay Code Excited Linear Prediction*, September 1992.

[42] ITU-T RECOMMENDATION. *P.800: Methods for subjective determination of transmission quality*, August 1996.

[43] ITU-T RECOMMENDATION. *P.861: Objective quality measurement of telephone-band (300-3400Hz) speech codecs*, February 1998.

[44] ITU-T RECOMMENDATION. *P.310: Transmission characteristics for telephone band (300-3400 Hz) digital telephones*, May 2000.

[45] ITU-T RECOMMENDATION. *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*, February 2001.

[46] JANSSEN, J. H. *A method for the calcuation of the speech intelligibility under conditions of reverberation and noise*, vol. 11 of *Benchmark Papers in Acoustics - Speech Intelligibility and Speaker Recognition*. Dowden, Hutchinson and Ross, Inc., 1977, ch. 14, pp. 158–163. Reprinted from pp. 305-310 of Acustica 7(5), 1957.

[47] KIANG, N. Y. S., AND MOXON, E. C. Tails of tuning curves of auditory nerve fibres. *Journal of the Acoustical Society of America 55* (1974), 620–630.

[48] KITAWAKI, N., ITOH, K., HONDA, M., AND KAKEHI, K. Comparison of objective speech quality measures for voiceband codecs. In *IEEE International Conference on Acoustics, Speech, and Signal Processing* (1982), vol. 2, IEEE, pp. 1000–1003.

[49] KITAWAKI, N., NAGABUCHI, H., AND ITOH, K. Objective quality evaluation for low bit-rate speech coding systems. *IEEE journal on selected areas in communications 6*, 2 (February 1988), 242–248.

[50] KLEINBAUM, D. G., AND KUPPER, L. L. *Applied regression analysis and other multivariable methods*. Duxbury Press, A division of Wadsworth Publishing Company, Inc., 1978, ch. 6.

[51] KRYTER, K. D. *Validation of the articulation index*, vol. 11 of *Benchmark Papers in Acoustics - Speech Intelligibility and Speaker Recognition*. Dowden, Hutchinson and Ross, Inc., 1977, ch. 13, pp. 153–157. Reprinted from pp. 1698-1702 of Journal of Acoustical Society of America, 34(11), 1962.

[52] LADEFOGED, P. *A Course in Phonetics*, 2 ed. Harcourt Brace Jovanovich, Inc., 1982.

[53] LADEFOGED, P. *Vowels and Consonants: An Introduction to the Sounds of Languages*. Blackwell Publishers, 2001.

[54] LEE, L.-S. Voice dictation of Mandarin Chinese. In *IEEE Signal Processing Magazine*. July 1997, pp. 63–101.

[55] LEVITT, H. Noise reduction in hearing aids: An overview. *Journal of Rehabilitation Research and Development* (2001).

[56] LI, Z., TAN, E. C., MCLOUGHLIN, I., AND TAN, T. T. Proposal of standards for intelligibility tests of Chinese speech. In *IEE Proceedings - Vision, Image Signal Processing* (June 2000), vol. 147, IEE, pp. 254–260.

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

[58] MCLOUGHLIN, I., AND DING, Z.-Q. Mandarin speech coding using a modified RPE-LTP technique. In *The 2000 IEEE Asia-Pacific Conference on Circuits and Systems (APCCAS 2000)* (14-16 December 2000), IEEE, pp. 748–751.

[59] MILLER, G. A. *Language and Communication*, 1 ed. McGraw-Hill Book Company, Inc., 1951.

[60] MILLER, G. A., AND NICELY, P. E. *An analysis of perceptual confusions among some english consonants*, vol. 11 of *Benchmark Papers in Acoustics - Speech Intelligibility and Speaker Recognition*. Dowden, Hutchinson and Ross, Inc., 1977, ch. 30, pp. 332–346. Reprinted from Acoust. Soc. Am. J. 27(2):338-352 (1955).

[61] MOORE, B. C. J. *An Introduction to the Psychology of Hearing*, 5 ed. Academic Press, 2003.

[62] NEJIME, Y., AND MOORE, B. Evaluation of the effect of speech-rate slowing on speech intelligibility in noise using a simulation of cochlear hearing loss. *Journal of the Acoustical Society of America 103* (1998), 572–576.

[63] NIEDERJOHN, R. J., AND GROTELUESCHEN, J. H. The enhancement of speech intelligibility in high noise levels by high-pass filtering followed by rapid amplitude compression. *IEEE transactions on Acoustics, Speech, and signal processing 24*, 4 (August 1976).

[64] NORMAN, J. *Chinese*. Cambridge University Press, 1997.

[65] PICKETT, J. M. *The Acoustics of Speech Communication*. Allyn and Bacon, 1999, ch. 2.

[66] PLOMP, R., AND MIMPEN, A. M. Improving the reliability of testing the speech reception threshold for sentences. In *Audiology* (Jan-Feb 1979), vol. 18, pp. 43–52.

[67] PREMINGER, J. E., AND VAN TASELL, D. J. Quantifying the relation between speech quality and speech intelligibility. *Journal of Speech and Hearing Research 38* (June 1995), 714–725.

[68] RABINER, L. R., AND SCAHFER, R. W. *Digital Processing of Speech Signals*. Prentice-Hall, Inc., 1978, ch. 3.

[69] RIX, A. W. Advances in objective quality assessment of speech over analogue and packet-based networks. In *IEE Data Compression Colloquium* (23 November 1999), IEE.

[70] RIX, A. W. Comparison between subjective listening quality and P.862 PESQ score. Tech. rep., Psytechnics Limited, September 2003.

[71] RIX, A. W., BEERENDS, J. G., HOLLIER, M. P., AND HEKSTRA, A. P. PESQ - The new ITU standard for end-to-end speech quality assessment. *109th AES Convention* (September 22-25 2000).

[72] RIX, A. W., BEERENDS, J. G., HOLLIER, M. P., AND HEKSTRA, A. P. Perceptual evaluation of speech quality (PESQ) - A new method for speech quality assessment of telephone networks and codecs. In *Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '01)* (2001), vol. 2, IEEE, pp. 749–752.

[73] RIX, A. W., AND HOLLIER, M. P. The perceptual analysis measurement system for robust end-to-end speech quality assessment. In *Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '00)* (5-9 June 2000), vol. 3, IEEE, pp. 1515 – 1518.

[74] RODMAN, J. The effect of bandwidth on speech intelligibility. White paper, Polycom inc., 2003.

[75] SEN, D. Determining the dimensions of speech quality from PCA and MDS analysis of the diagnostic acceptability measure. In *Measurement of Speech and Audio Quality in Networks (Online Workshop)* (2002). http://wireless.feld.cvut.cz/mesaqin2002/full10.pdf.

[76] STEENEKEN, H. J. M. The measurement of speech intelligibility. In *Proceedings of the Institute of Acoustics* (2001), vol. 23, Stratford-upon-Avon, UK.

[77] STEENEKEN, H. J. M., AND HOUTGAST, T. Subjective and objective speech intelligibility measures. In *Proceedings of the Institute of Acoustics* (1994), vol. 16, pp. 95–121.

[78] STEVENS, S., AND VOLKMAN, J. The relation of pitch to frequency. *American Journal of Psychology 53* (1940), 329–353.

[79] SUN, H., SHUE, L., AND CHEN, J. Investigations into the relationship between measurable speech quality and speech recognition rate for telephony speech. In *Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04)* (May 2004), vol. 1, IEEE.

[80] THE COLUMBIA ENCYCLOPEDIA, SIXTH EDITION. Pinyin, 2001. `http://www.bartleby.com/65/pi/Pinyin.html`.

[81] THE COLUMBIA ENCYCLOPEDIA, SIXTH EDITION. Voice, sound produced by living beings, 2001. `http://www.bartleby.com/65/vo/voice1.html`.

[82] THOMAS, I. B., AND NIEDERJOHN, R. J. Enhancement of speech intelligibility at high noise levels by filtering and clipping. *The Journal of the Audio Engineering Society 16* (1968), 412–415.

[83] THOMAS, I. B., AND NIEDERJOHN, R. J. The intelligibility of filtered-clipped speech in noise. *The Journal of the Audio Engineering Society 18*, 3 (June 1970), 299–303.

[84] THORPE, L., AND YANG, W. Performance of current perceptual objective speech quality measures. In *Speech Coding Proceedings* (1999), IEEE, pp. 144–146.

[85] VAUGHAN, N. E., FURUKAWA, I., BALASINGAM, N., MORTZ, M., AND FAUSTI, S. A. Time-expanded speech and speech recognition in older adults. *Journal of Rehabilitation Research and Development 39*, 5 (September/October 2002), 559–566.

[86] VOIERS, W. D. Diagnostic acceptability measure for speech communication systems. In *Proceedings of the 1977 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '77)* (May 1977), vol. 2, IEEE, pp. 204–207.

[87] VOIERS, W. D. *Diagnostic evaluation of speech intelligibility*, vol. 11 of *Benchmark Papers in Acoustics - Speech Intelligibility and Speaker Recognition*. Dowden, Hutchinson and Ross, Inc., 1977, ch. 34, pp. 374–387.

[88] VOIERS, W. D. Interdependencies among measures of speech intelligility and speech "Quality". In *Proceedings of the 1980 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '80)* (April 1980), vol. 5, IEEE, pp. 703–705.

[89] VOIERS, W. D. Evaluating processed speech using the diagnostic rhyme test. *Speech Technology* (1983), 30–39.

[90] VORAN, S. Objective estimation of perceived speech quality - Part I: Development of the measuring normalizing block technique. *IEEE Transactions on Speech and Audio Processing 7*, 4 (July 1999), 371–382.

[91] VORAN, S. Objective estimation of perceived speech quality - Part II: Development of the measuring normalizing block technique. *IEEE Transactions on Speech and Audio Processing 7*, 4 (July 1999), 383–390.

[92] VORAN, S., AND SHOLL, C. Perception-based objective estimators of speech quality. In *Proceedings of the 1995 IEEE Workshop on Speech Coding* (September 1995), IEEE.

[93] WANG, S., SEKEY, A., AND GERSHO, A. An objective measure for predictive subjective quality of speech coders. *IEEE journal on selected areas in communications 10*, 5 (June 1992), 819–829.

[94] WHITE, F. A. *Our Acoustic Environment*. John Wiley and Sons, Inc., 1975.

[95] YANG, M. Low bit rate speech coding. IEEE Potentials. http://ieeexplore.ieee.org/iel5/45/29584/01343228.pdf, 2004.

[96] YANG, W. *Enhanced Modified Bark Spectral Distortion (EMBSD): An Objective Speech Quality Measure Based On Audible Distortion and Cognition Model*. PhD thesis, Temple University, May 1999.

[97] YANG, W., BENBOUCHTA, M., AND YANTORNO, R. Performance of the modified bark spectral distortion as an objective speech quality measure. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '98* (12-15 May 1998), vol. 1, IEEE, pp. 541–544.

[98] YANG, W., DIXON, M., AND YANTORNO, R. A modified bark spectral distortion measure which uses noise masking threshold. In *IEEE Workshop on Speech Coding For Telecommunications Proceeding* (7-10 September 1997), IEEE, pp. 55–56.

[99] YANG, W., AND YANTORNO, R. Comparison of two objective speech quality measures: MBSD and ITU-T recommendation P.861. In *IEEE Second Workshop on Multimedia Signal Processing* (7-9 December 1998), IEEE, pp. 426–431.

[100] YIP, M. *Tone*. Cambrigde University Press, 2002, ch. 7.

[101] ZHANG, G., ZHENG, F., AND WU, W. Tone recognition of chinese continuous speech. In *International Symposium on Chinese Spoken Language Processing* (13-15 October 2000), IEEE, pp. 207–210.

[102] ZHANG, J. Phonetic and linguistic features of spoken Chinese. In *Proceedings of the 1994 International Symposium on Speech, Image Processing and Neural Networks (ISSIPNN '94)* (13-16 April 1994), vol. 1, IEEE, pp. 117–121.

[103] ZHANG, J. On the syllable structures of Chinese relating to speech recognition. In *Proceedings of the Fourth International Conference on Spoken Language (ICSLP 96)* (3-6 October 1996), vol. 4, IEEE, pp. 2450–2453.

[104] ZWICKER, E., AND FASTL, H. *Psychoacoustics: Facts and Models*. Springer-Verlag Berlin Heidelberg, 1999.

Appendix A

Evaluation of ITU-T G.728 as a Voice over IP codec for Chinese Speech

This section include the details of the evaluation part for experiments 2 and 5 in sections 5.3.1 and 5.5.1 conducted with 30 subjects that was published in the proceedings of the Australian Telecommunications Networks and Applications Conference (ATNAC) 2003 [21].

A.1 Abstract

Voice-over-IP is expected to become a popular service offered by the Internet. Thus, it is important to ensure high quality of service. In this paper, we look at two standards proposed for evaluating the intelligibility of Chinese speech. Adopting the philosophy and methodology of the Diagnostic Rhyme Test (DRT) for testing English speech, the Chinese Diagnostic Rhyme Test (CDRT) evaluates the six elementary phonemic attributes of Chinese words. Since Chinese is a tonal language, an extension of CDRT called CDRT-Tone evaluates the tonal attributes of Chinese speech. These two tests were used to evaluate the ITU-T G.728 speech coder as a VoIP codec for Chinese speech. Results are compared to the previous evaluations on a GSM 06.10 coder.

A.2 Introduction

Voice over IP systems use speech codecs to optimise the usage of transmission bandwidth as well as storage. Due to the fact that some speech information is lost in speech coding, the original speech might not be recoverable after transmission. This loss of information might affect both intelligibility and quality of the

output speech, where intelligibility means how well one can understand what is being said, and quality means the degree of goodness in the perception of speech. Although these are two different attributes, they are not totally exclusive of each other. Having good quality will mean that intelligibility is of a high standard but this relationship is not reciprocal. Various intelligibility and quality tests were introduced to these two attributes on IP networks or speech codecs. Such tests can be categorised as the *subjective* and *objective* tests, where subjective tests involve a group of human listeners to rate either of the two attributes, and objective tests involve some mathematical expressions used to determine speech quality.

Some well known subjective intelligibility tests include the *Diagnostic Rhyme Test* (DRT), *Modified Rhyme Test* (MRT), and *Phonetically Balanced Word Lists* (PB) [27]. These are the ones listed as the ANSI standards for speech intelligibility testing. The more popular subjective quality tests are the *Diagnostic Acceptability Measure* (DAM) and the *Mean Opinion Score* (MOS) [28]. Various objective quality measures include the *Perceptual Speech Quality Measure* (PSQM) [43], *Perceptual Evaluation of Speech Quality* (PESQ) [45], and *Deustche Telekom Speech Quality Estimation* (DT-SQE) [32].

In this paper, the issue of intelligibility is dealt with, in particular, intelligibility of Chinese Speech. Taking into account that Mandarin Chinese is a language spoken by more than one billion people throughout the world, to provide a better quality of service for the Voice over IP environment, two sets of standards for testing the intelligibility of Chinese speech namely the *Chinese Diagnostic Rhyme Test* (CDRT) [56] and its extension, CDRT-Tone [26] were proposed. The testing methods of CDRT and CDRT-Tone are being reviewed, and applied to test the ITU-T G.728 speech coder. The results were used to compare to those from previous evaluations on a GSM 06.10 coder [25][26].

A.2.1 Chinese Diagnostic Rhyme Test (CDRT)

Adopting the philosophy and methodology of the DRT, the CDRT was proposed to evaluate the intelligibility of Chinese speech transmitted through communication systems. It is effectively the DRT applied to Chinese. It uses a corpus of 192 words in 96 rhyming pairs. From this 96 rhyming pairs, six elementary phonemic attributes are tested. They are *airflow*, *nasality*, *sustention*, *sibilation*, *graveness*,

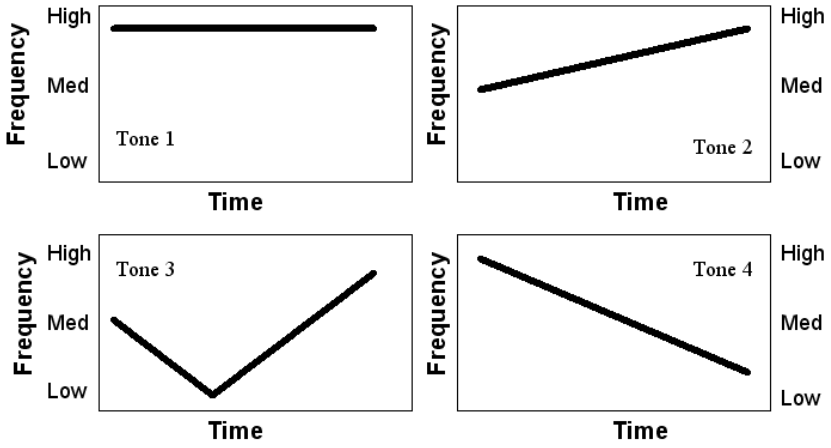

Figure A.1: Frequency characteristics of the Four Chinese Tones

and *compactness*. By obtaining results on which attribute fails, a system's flaw could be easier identified and therefore corrected. Although the DRT is rather extensive in testing important attributes of English speech, the CDRT does not test all the characteristics of Chinese speech because Chinese, differing from English, is a tonal language. Since CDRT only discriminates consonants, vowels and tones are not tested. Hence one cannot form a concrete conclusion of the intelligibility of Chinese speech in a particular system solely based on CDRT. The corpus of Chinese characters is given in [56].

A.2.2 CDRT-Tone

In the Chinese language, most syllables/phonemes can be pronounced with one of four different tones [100]. Tone 1 is a high-level tone, tone 2 is a mid-rising tone, tone 3 is a low-falling-rising tone, and tone 4 is a high-falling tone. Figure A.1 shows the frequency characteristics of the four tones. Pronouncing a syllable with different tones gives different meanings. For example the Chinese syllable "ma" will mean "mother" with the first tone, "numb" with the second, "horse" with the third, and "scold" with the forth. As an extension to CDRT, CDRT-Tone tests the tonal intelligibility of Chinese syllables. It consists of 40 pairs of Chinese syllables divided into four categories according to the similarity of pitch height and contour among the four tones. The categories are: *(tone 1-tone 2)*, *(tone 1-*

tone 3), *(tone 2-tone 3)*, and *(tone 3-tone 4)*. Categories like *(tone 1-tone 4)* and *(tone 2-tone 4)* are omitted because their pitch heights and contours are different. With the CDRT-Tone, the intelligibility of Chinese speech transmitted through a particular system can be more confidently concluded on top of using CDRT alone. The 40 pairs of Chinese characters are given in [26].

A.3 Evaluation of G.728 using CDRT and CDRT-Tone

ITU-T G.728 [41] LD-CELP is a 16 kbit/s low delay speech coder standard based on the principle of Low Delay Code Excited Linear Prediction. It is commonly used for transporting audio in VoIP systems. The CDRT and CDRT-Tone tests are applied to G.728 and the results later compared to those for GSM. In this evaluation, the source files of the 96 rhyming pairs of Chinese words in CDRT and the 40 pairs in CDRT-Tone, spoken by a native Chinese speaker, were recorded in an Anechoic chamber, with a sampling rate of 16kHz. This is to provide a better quality source with a reduction of background noise and a higher sampling rate. Using an almost similar methodology used in the previous evaluation, this evaluation is done using a Laptop computer with a high quality Philips HS900 Headphone. The source files (original datasets) were recorded and stored in the computer. A set of processed files (processed datasets) were obtained by coding and then decoding the original datasets using the G.728 coder. For CDRT, 192 original plus 192 coded-decoded files were played in random sequence. 80 original plus 80 coded-decoded for CDRT-Tone. 30 native Chinese speakers with no hearing impairments participated in this evaluation and all of them took part in both the CDRT test and the CDRT-Tone test.

In both tests, a word pair is presented to listeners with one of the words played through the headphone. To ensure recognition of the Chinese characters, the Hanyu Pinyin (Pronunciation of the Chinese words written using English alphabets) is displayed next to each character. The subject is asked to select which of the presented word is being played using the numerical keyboard. The subjects are allowed to listen to the word again if they did not hear the first one correctly and they were also allowed to make corrections if they pressed a wrong key. A trial session was given to the subjects before the actual test to help them familiarise with the test procedures and to adjust the loudness of the headphones. After

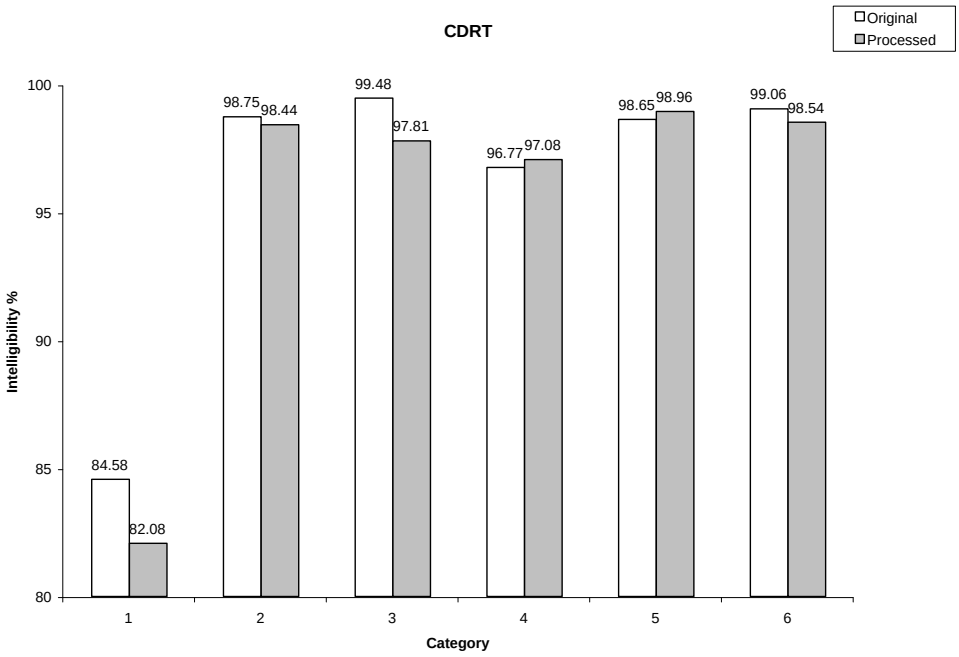

Figure A.2: CDRT test results

Category 1: *(Sibilated vs Unsibilated)*; Category 2: *(Compact vs Diffuse)*; Category 3: *(Grave vs Acute)*; Category 4: *(Nasal vs Oral)*; Category 5: *(Airflow vs No Airflow)*; Category 6: *(Sustained vs Interrupted)*

listening to every 32 words for the CDRT (20 for CDRT-Tone), they were allowed to take a two minute break to reduce the effects of fatigue.

A.4 Results

Results of the CDRT and CDRT-Tone tests are presented in Figures A.2 and A.3 respectively.

Figure A.2 shows that for the CDRT test, the intelligibility of the original versus the processed speech is on average 0.73 % higher. Compared to the results obtained in [25], the level of intelligibility of both original and processed speech

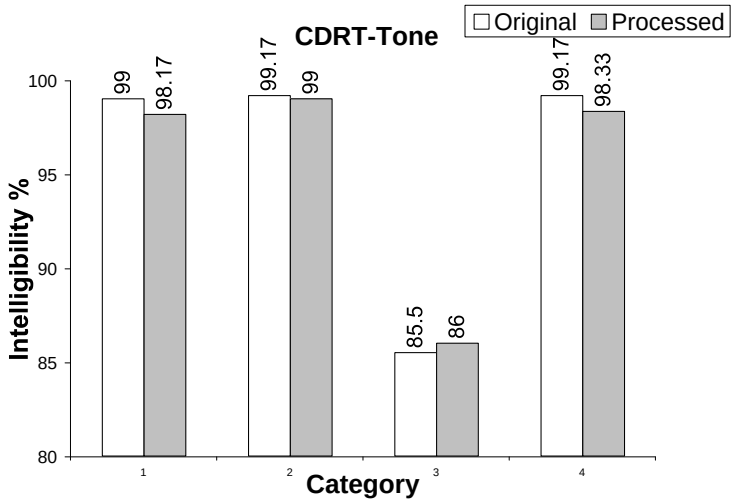

Figure A.3: CDRT-Tone test results
Category 1: *(Tone 1-Tone2)*; Category 2: *(Tone 1-Tone3)*; Category 3: *(Tone 2-Tone3)*; Category 4: *(Tone 3-Tone4)*

Table A.1: Comparison of degradation between G.728 and GSM for CDRT.

Category	G.728	GSM
1	3.05%	16%
2	0.31%	0.14%
3	1.71%	0.14%
4	-0.32%[a]	0%
5	-0.31%	3.49%
6	0.53%	1.41%
Average[b]	0.93%	3.53%

a Negative results denote improvement instead of degradation in tables A.1 and A.2.
b Negative results are regarded as 0% when calculating averages in tables A.1 and A.2.

files are higher in all six categories. This could be due to the fact that the original sound files were sampled at 16kHz here rather than 8kHz as used in the previous evaluation. None of the six categories has a degradation of intelligibility higher than 3%. From this fact, we can see that all six elementary phonemic attributes were well preserved by the G.728 coder. Refering to table A.1, besides categories 2 and 3, the G.728 coder yields a higher intelligibility than GSM, especially in category 1 where there is a significant difference (3.05%(G.728) vs 16%(GSM)) in the amount of degradation. The average degradation of the G.728 coder is 0.93% while the GSM is 3.53%. We cannot, however, conclude that G.728 performs better than the GSM due to the difference in sampling rate used and it is not our intention to in this paper to compare the performance between the two coders.

When tested using CDRT-Tone, the G.728 is shown to have preserved tonal intelligibility excellently. The degradation of intelligibility for all categories is lower than 1% with an average of 0.47%. This shows a significant difference from the 8.05% for the GSM (See table A.2).

A.5 Discussions

Analysing the results, the good performance by the G.728 coder is somewhat expected. It uses a high-order (50th order) linear predictor which is used for exploiting both pitch and formant redundancies. Furthermore, the filter coefficients and gain information are unquantised since they are calculated using robust adaptation

Table A.2: Comparison of degradation between G.728 and GSM for CDRT-Tone.

Category	G.728	GSM
1	0.85%	11.63%
2	0.17%	3.33%
3	-0.58%	11.63%
4	0.85%	5.62%
Average	0.47%	8.05%

algorithms in both the analyser and synthesiser, hence resulting in its good output quality (in this case, intelligibility) [31]. Note that the level of intelligibility of the processed speech for categories 4 and 5 of CDRT and category 3 of CDRT-Tone is higher than the original. This phenomena is explained by the good performance of the G.728 coder. It is also interesting to note that although the scores are generally higher than previous evaluations done using CDRT and CDRT-Tone, the third category *(Tone 2-Tone 3)* of CDRT-Tone yields a lower overall intelligibility (85.5%(original), 86%(processed) vs 96%, 86%). This phenomena is not likely to be caused by the difference in sampling rate as a higher sampling rate should improve intelligibility. It might also not be caused by the coder since the difference in intelligibility occurs even for the original unprocessed speech. Judging from feedback given by some of the subjects, some of them find it difficult differentiating between tone 2 and 3. Some even commented that the word they hear is neither of the two displayed on the monitor. From this phenomena and the feedbacks, two possible reasons can be deduced:

1. The characteristics or background of subjects

2. The limited recognition of the words used in the tests

The characteristics or background of subjects can refer to where the subject comes from or is brought up. Even though native Chinese speakers, Chinese from mainland China, Taiwan, Southeast Asia and other countries speak Mandarin Chinese with different accents. Accents are different even between people from different parts of China or Taiwan. This is similar to the difference in English accents spoken by Americans, English, New Zealanders, etc as well as the northern part

of the UK and the south. Having different backgrounds, the way of pronouncing Chinese words and/or their tones differ. This would meant that perception of the word or tone will be different. Among the subjects, 8 were from mainland China, 18 from Taiwan, and 4 from Southeast Asia (Singapore and Malaysia). The background of subjects can also refer to trained and untrained(naive) listeners, Chinese language teachers or someone more exposed to the phonetics of Chinese speech and common native Chinese speakers. The subjects who participated were from such backgrounds. For this reason, differences in background not only affect tones, but other aspects like vowels and consonant attributes as well. This may describe the reason for the lower intelligibility scores obtained in category 1 *(Sibilated vs Unsibilated)* in evaluations done for both GSM and G.728 coders, and category 3 *(Grave vs Acute)* for GSM, when compared to other categories. To solve the problem for this issue, it is suggested that subjects used should be trained listeners who are familiar with the pronunciation and tones of the words in the tests. If untrained listeners are used, people with similar backgrounds should be considered.

Another reason is the recognition of the words used in the tests. Some subjects commented that several words used in the tests are not commonly seen or used. Although the Hanyu Pinyin of each Chinese character is displayed next to each Chinese character, this only assists subjects from mainland China and some parts of Southeast Asia since the Taiwanese have not been exposed to the Hanyu Pinyin system. Hence for the Taiwanese, their only way is to recognise the Chinese characters. A word or character displayed during the test that is not so commonly used might yield an erroneous result by them. To minimise such errors, words within the test that are less commonly used can be replaced by words with similar pronunciation that are more commonly used. Another issue regarding recognition is that some characters have more than one pronunciation. One example in CDRT is a character that can be pronounced either as /zan4/ or /zhan4/. This might cause an error if a subject who recognises this character as /zhan4/ chooses this word, ignoring the other /zhan4/ in the same pair. This also happens in CDRT-Tone where a word could be pronounced in several tones and will also cause some confusion. It is therefore suggested that words that have only one pronunciation are used in the tests.

A.6 Conclusions

The CDRT and CDRT-Tone tests have been used to evaluate intelligibility of transmitted Chinese speech so as to predict quality of service in a Voice over IP environment. In this evaluation, the ITU-T G.728 LD-CELP coder is shown to accurately preserve the six elementary phonemic as well as tonal attributes of Chinese words. The results are compared to previously published data on GSM to show that G.728 more faithfully preserves attributes important to the intelligibility of Chinese speech. Although both tests can adequately test the intelligibility of systems, they could also be further improved to obtain a higher level of accuracy.

Appendix B

A study on the influence of subjective background on speech intelligibility tests

B.1 Introduction and results

To increase the confidence and to investigate the effect the background of subjects has on the results of the evaluation mentioned in section A.5, 10 more subjects from mainland China participated in the evaluation. This is to balance the number of mainland Chinese and Taiwanese subjects each to 18. Table B.1 and B.2 showed the averaged consonantal and tonal intelligibility (in percentage) of the original and processed files, and amount of degradation for the two groups.

B.2 Discussions

From table B.1, we observed that difference in averaged intelligibility between mainland Chinese and Taiwanese were very small (less than 3.6%) for categories 2 to 6 of CDRT. For category 1 (sibilation), however, the Taiwanese yielded a lower averaged consonantal intelligiblity with a difference of 12.6%. This difference is significant at $\alpha = 0.05$ when a one-tailed *t*-test is performed. The case for the CDRT-Tone test is rather similar with the Taiwanese yielding a lower intelligibility in category 3 (Tone 2 - Tone 3). This difference, however, is not statistically significant. The few background factors mentioned in section A.5 were that:

1. there are differences in the pronunciation and recognition of some Chinese words by Chinese from different regions,

2. some words listed in both tests are less commonly used,

Table B.1: Average consonantal intelligibility of original and processed files, and amount of degradation from mainland Chinese and Taiwanese.

CDRT	Averaged intelligibility (%)		Amount of degradation (%)	
Phonemic Category	Mainland Chinese	Taiwanese	Mainland Chinese	Taiwanese
1	92.8	80.2	0.6	2.2
2	98.8	98.5	0	0.2
3	99.3	98.5	1.4	2.3
4	94.1	97.6	-1.5[a]	-1.1
5	99.6	98.8	-0.5	0
6	99.9	98.1	-0.2	1.1

a Negative results denote improvement instead of degradation in tables B.1 and B.2.

Table B.2: Average tonal intelligibility of original and processed files, and amount of degradation from mainland Chinese and Taiwanese.

CDRT-Tone	Averaged intelligibility (%)		Amount of degradation (%)	
Phonemic Category	Mainland Chinese	Taiwanese	Mainland Chinese	Taiwanese
1	99.2	98.3	0	0.6
2	99.0	99.2	0.3	0.6
3	92.8	86.5	0.6	-1.6
4	98.9	99.2	-0.6	0.6

3. some words listed in both tests have more than one phonemic or tonal pronunciation, and

4. the difference in the romanisation process (different alphabetic representations for mainland Chinese and Taiwanese).

During the experiments, a considerable fraction of subjects that participated in both tests reflected that some words used in CDRT or CDRT-Tone is uncommon or seldom used in daily lives. This, however, can be compensated if a subject can recognise one word from each pair that could enable him/her to select or reject either of each word presented. From our results, where both types of intelligibility is generally high and a significant difference only occurred in one category from CDRT, we conjecture that this factor does not heavily influence the results of the tests unless all uncommon words occurred in the same category (sibilation).

We also observed that average intelligibility of words from the same category in CDRT, where a significant difference occurred between the mainland Chinese and Taiwanese, and category 3 in CDRT-Tone were lower than other categories in their respective tests by both groups of subjects. It was previously mentioned in section 5.3.3 that sibilation is one of the most commonly confused consonants among all Chinese words and this was again shown here. The distinction between tone 2 and tone 3 were also confused by some Chinese (see section 5.5.3). These results can be said to be heavily influenced by subjective background as shown by the significant difference in both groups for the sibilation category. Even among the same country, recognition or pronunciation of words also differ by people from different states. This phenomenon is similar to people from different parts of England or different English speaking countries having different pronunciations or accents. Therefore, subjects used should be carefully chosen, for example, choosing subjects having the same background.

There are a few words in CDRT and CDRT-Tone that can be pronounced in either way within the same pair, for example the word /zan4/ in CDRT is pronounced as /zan4/ by the mainland Chinese, Singapore, and some other countries that adopts Mandarin Chinese from mainland China [5], while both /zan4/ or /zhan4/ can be pronounced by the Taiwanese [6], and the words /fa3/ and /hua2/ can also be pronounced as their counterpart in the same pair [6]. In some cases,

some words may have multiple pronunciations like the word /zen3/ in CDRT can also be pronounced as /ze3/ or the word /tiao3/ in CDRT-Tone as /tiao1/. Some of these examples will directly affect the recognition of words and therefore the results of the tests, and the probability is even greater when their counterparts in the same pair is uncommon (not easily recognised). Although recognition was improved by the corresponding Hanyu Pinyin alphabets given next to the Chinese characters in the test, these alphabets could only be recognised by the mainland Chinese as Taiwanese have their own set of alphabetic representations (see section 3.4.1). Therefore, a concise research is required for the list of Chinese words used in Chinese intelligibility tests to avoid such confusions. Another solution would be just to list the specific Hanyu Pinyin or Tongyong Pinyin alphabets for respective subjective groups without displaying the Chinese characters.

B.3 Conclusions

The influence of subjective background on speech intelligibility tests mentioned in section A.5 ([21]) was investigated. From this investigation, it was realised that the main influence on the CDRT and CDRT-Tone tests was the differences in the pronunciation and recognition of some Chinese words by Chinese from different regions as seen in the significant difference between both groups for the sibilation category. The choice of words used in the lists was also influential despite the presentation of Hanyu Pinyin alphabets which only assisted the mainland Chinese. This, however, can be compensated if ambiguity is not too great.

Appendix C

Correlation between subjective DMOS test and OSQMs

C.1 Introduction and results

As a side track from our research, a subjective *Degradation Mean Opinion Score* (DMOS) test, derived from the degradation category rating (DCR) test [42], was performed to obtain a set of subjective scores. The *DMOS* test was chosen because both PESQ and MNB estimate the quality scores based on comparison of the original speech signal and the processed (degraded) one, and this comparison approach parallels the DMOS test. This experiment was conducted using 30 subjects giving their opinions on the quality of 60 processed Chinese speech files based on their original files as references. The following conditions were applied on the processed files to possibly degrade their quality:

- Simulated vehicular noise
- Filtering (Low, High, or Band pass)
- Speech coding
- Temporal clipping of speech
- Signal distortion
- Signal amplification
- No changes made (original signal)

The objective quality scores for the 60 processed Chinese speech files were also computed by PESQ and MNB. A Pearson's correlation coefficient was computed from the results of the DMOS test and outputs from each OSQM. The correlation between DMOS and PESQ is 0.885, and MNB is 0.464 (Please refer to section 4.2.2 for more details on the DMOS test).

C.2 Discussions and conclusion

Although it is irrelevant to compare our results to those published for both OSQMs due to different languages and test conditions, it is interesting to note that both PESQ and MNB yielded higher correlations of 0.935 [16] and a range from 0.910 to 0.986 [91]. A possible reason could be the differing test conditions. Among the conditions, temporal clipping of speech (see section 5.4.2) and signal amplification were not validated by ITU [43][45] and hence PESQ and MNB's accuracy is doubtable. The lack of accuracy can be seen by the poor correlation of 0.464 for MNB. PESQ's correlation of 0.885 is acceptable although it did not match that which was published. From PESQ's fair correlation, the doubt of the cause for this lower correlation arises: was this due to the inaccuracy of unvalidated conditions, difference in test language, or by chance? Taking MNB's clear distinction in the difference of correlations as example, the deterioration in PESQ's correlation should be lower if it is due to inaccuracies of unvalidated conditions. This makes the possibility of the difference in test language, or chance higher. Referring to the consonantal and tonal intelligibility issues mentioned in chapter 5, there is a high possibility that language is a factor in this deteriorated result. These results, on top of our main research findings, again showed us that OSQMs developed for certain languages may not adequately be used for other languages such as Chinese. This also prompted us for further research into a universal or multilingual speech quality measurement system.