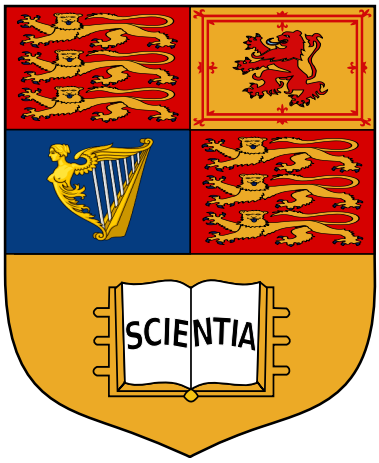
Imperial College London

Department of Electrical and Electronic Engineering

Final Year Project Report 2019-20



Project Title:	**Hearables: Recreating standard speech from the speech recorded inside the ear-canal**
Student:	**Anirudh Yadav**
CID:	**01219167**
Course:	**EE4**
Project Supervisor:	**Prof. Danilo P. Mandic**
Second Marker:	**Prof. Patrick A. Naylor**

Plagiarism Statement

Final Report Plagiarism Statement

I affirm that I have submitted, or will submit, electronic copies of my final year project report to both Blackboard and the EEE coursework submission system.

I affirm that the two copies of the report are identical.

I affirm that I have provided explicit references for all material in my Final Report which is not authored by me and represented as my own work.

Acknowledgements

I would firstly like to thanks Prof. Danilo Mandic, Dr. Iam Williams and Metin Yarici for their active involvement in helping me direct the project objectives. I am also grateful for their valued time in having technical discussions about the project, and providing me guidance during the course of this project.

In addition to this, I would like thanks my family and friends for their support in helping me navigate through trying times and issues whilst completing this project.

Abstract

With advancements in wearable technology, sensors can be embedded into inconspicuous devices to monitor and capture bio-signals. Although smart watches currently dominate the wearable market, wearable earpieces (termed hearables) have become popular. Hearables are electronic in-ear devices which can be used for wireless transmission to communication, and more interestingly to monitor neural and physiological function.

Along with the ability to capture neural and physiological function, Hearables can be used to capture inner-ear speech, through microphones facing the ear-canal. Since Hearables provide an acoustic seal between the external noise and in-ear speech, the sound quality of in-ear speech is often superior to that of external microphones. However, in-ear speech suffers from dampening of high-frequencies through the skull and other tissues in the head, which can affect its intelligibility.

The project aims to utilise in-ear speech superior quality and recover the high-frequency components of speech by establishing a transfer function between in-ear and externally recorded speech.

Contents

List of Abbreviations

AEM	Acoustic Microphone
ALSD	Average Log-Spectral Distortion
AR	Autoregressive Modelling
BC	bone-conducted speech
BWE	Bandwidth Extension
ECM	Electret Condenser Microphone
EQ	Equalisation
FFT	Fast-Fourier Transform
IEM	Inner Ear Microphone
LPC	Linear Predictive Coding
LPCC	LPC Cepstral coefficients
MMS	Multi-Modal sensor
MMSE	Minimum Mean Squared Error
PSD	Power Spectral Density
SG	Spectral Gain
SNR	Signal-to-Noise Ratio
STFT	Short-Time Fourier Transform
STSA-MMSE	Short-Time Spectral Amplitude-MMSE
VAD	Voice Activity Detection

Chapter 1

Introduction

1.1 Bone-conducted vs Aerial Speech

Traditionally, speech is acquired using an aerial/acoustic microphone, placed near the mouth. Although, still the most prevalent method of speech capture; aerial microphones often suffer from a low signal-to-noise ratio (SNR) in noisy environments. Therefore, post processing techniques to enhance speech intelligibility are required. This task becomes excessively difficult when multiple noise sources are present, and especially when background noise is non-stationary in nature. Microphones are available which cancel noise in either passive or active (adaptive filtering) manner [1], but require multiple inputs microphones at specific positions.

A proposed solution to capturing speech free of environmental noise is through principle of bone conduction. Bone conducted speech can be acquired by placing microphones inside the occluded ear [2], called in-ear microphone, or through bone conduction sensors on the cranium [3]. This project specifically investigates the relationship between speech acquired from an in-ear microphone and external microphone. IEMs operate by capturing speech in the ear canal that propagates thorough the vibrations in the skull tissue and bone. Since, bone-conducted speech avoids any aerial pathways, IEM recorded speech generally has a greater signal-to-noise ratio (SNR) [4] compared to speech inputted using aerial microphones. However, bone-conducted speech signal has a limited bandwidth, which cuts-off at approximately 2kHz [5]. Furthermore, in-ear speech sounds boomy and causes ringing due to resonance of low-frequency formants, hence further degrading the perpetual quality of in-ear speech. Hence, to leverage the high SNR of in-ear speech the undesired characteristics must be filtered, and this is attempted in the project by establishing a transfer function, to convert in-ear speech to externally recorded speech. Next the applications of in-ear speech are explored as a motivation for conducting research on this topic.

1.2 Motivation

The study of the project is also motivated by the potential applications of Hearables in capturing intelligible speech. Typically, Hearables have been widely researched in monitoring cardiac, neural and respiratory function over large temporal intervals[6]. Similarly, the project can be utilised for capturing intelligible speech over larger intervals. The need for extensive speech data has been highlighted for applications is detecting neurological disorders, such as PTSD, and depression[7], using voice analysis technology.

In addition, this study can prove beneficial in extending studies to improve communication in excessively noisy environment. According to Occupational Safety and Health Administration, [8] excessively noisy levels, which corresponds to levels larger than 90dB for greater than 8 hours, require Hearing Protection Aids. While hearing aids are highly effective in protecting hearing, they limit communication between subject under such noise environments. In addition, limited communication has been registered one of the primary complains by users of Hearing protecting devices [9]. Hence the noise-free property of in-ear microphones can be utilised to

ensure communication between subjects under extremely noisy situation. The proposal for this problem is also studied by Voix. et.al, (2015).

Furthermore, given the current global health situation under Covid-19; the use of face masks is becoming increasingly common. Although the face mask does provide protection from spreading the viral infection; face masks also deteriorate the speech quality. Deterioration is in form of muffling the speech, making it unintelligible, hence hindering communication. This obstacle in communication in hospitals is a cause of serious concern. Therefore, utilising hearables for such application in used speech from inside ear-canal can improve speech quality, especially in form of mobile communication.

1.3 Aim

The project aims to establish a relationship between in-ear and externally recorded speech, under quiet conditions, through an invertible transfer function. The transfer function must be such that the it can then be used in absence of corresponding aerial microphone speech for enhancement of in-ear microphone recordings to improve its intelligibility. Finally, the implementation will be delivered in form of a software simulation in MATLAB, and this report which documents the development process of the system.

1.4 Objectives

In order to compute a transfer function between the media of speech capture, the following approaches are considered:

- **Development of In-ear Speech database**
  Databases with simultaneously record In-ear and external microphone speech are not readily available for open development. This is mainly due to two reasons, firstly, the research on hearables is a relatively novel field, and secondly; transfer function characteristics vary significantly depending on speaker identity, and the material of the hearables as well. An objective of the project is to develop an In-ear speech database, which can be utilised for further experiments on In-ear speech using the proposed hearable device.

- **Linear Spectral Gain**
  This approach requires calculating the transfer function as a frequency equalisation gain between the in-ear and aerial speech using short-term Fast Fourier Transform (FFT).

- **Artificial Bandwidth Extension**
  In contrast to linear spectral gain method, artificial bandwidth extension model is based on modelling the biophysics, specifically the Source-filter model, which generates higher order frequency harmonics based on mutual information between lower (0-2kHz) and higher (2-4kHz) frequency bands.

- **Phoneme-based mapping** This is an extension of the Linear Spectral Gain method. While the Linear Spectral Gain, applies a constant transfer function to full audio utterances, the Phoneme-based mapping assign each frame to a particular transfer function. This aims method model the bone-conduction channel's ability to attenuate sounds differently.

1.5 Report Organisation

The content mentioned above is arranged in the following manner:

- **Chapter 2: Technical Background**
  This section explores and outlines necessary analytical tools and techniques which are used in the process of implementing the approaches outlined in section 1.4. In addition, the section contains a survey of literature in exploring methods which have been used in the research community.

- **Chapter 3 & 4 (part 1): Analysis and Testing**
  Prior to the implementation section, analysis and design provides description of the proposed approaches from a system-level perspective. The section shows the development process of the approaches, and highlights decisions made during the design process. Analysis and design are divided into 2 chapters, as the chapters corresponds to Bandwidth Extension and Spectral-domain approach.

- **Chapter 3 & 4 (part 2): Implementation**
  The implementation section delves into the details of the approaches discussed in the Analysis and Design section. In this section, the high-level design is proposed in form of an algorithm. The section, therefore, also discusses several sections of code in implementation of the approaches mentioned in section 1.4. Implementation is divided into 2 chapter, as each chapter corresponds to Bandwidth Extension and Spectral-domain approach.

- **Chapter 5: Results and Testing**
  This section gathers findings from the proposed methods to develop the transfer function. In the Testing subsection, all model parameters are identified, and experimented to obtain optimal model performance.

- **Chapter 6: Conclusion and Evaluation**
  Conclusion and Evaluation review the success of the project. The section also identifies the strengths, and the pitfall of the proposed approaches and builds upon them by suggesting future work.

1.6 Chapter Summary

In this chapter, the problem of recording speech from an in-ear microphone is introduced. It is discussed firstly the advantages of using IEMs over AEMs, especially in noisy environment; where the SNR of AEM speech is low. Next the challenges of IEM speech mentioned in form of limited bandwidth; which in turns causes decline in speech intelligibility. With motivation being provided for using IEMs, along with the bandwidth problems, the project problem is introduced. The problem being able to establish a transfer function between IEM and AEM recorded speech.

The project objective of developing a transfer function is proposed to be approached from both a time and frequency domain perspective. Deliverable will be in form of a MATLAB implementation and this report providing documentation. Lastly, the report organisation is mentioned. In the next chapter the background information on basics of bone-conduction, along with the preliminary experiments on IEM recorded speech are discussed.

Chapter 2

Background

2.1 Biophysics of Bone-conduction

Before proceeding to method used to acquire bone-conducted speech, a brief explanation is provided for the conduction pathways in the human ear.

There are two transmission pathways for acoustic energy to propagate to the ear, and be perceived as sound. These are air-conducted and bone-conducted speech. Air-conduction is the typical pathway where sound energy from the vocal tract propagates to the ear through vibrating air molecules. These vibrations enter the ear-canal, and consequently cause the tympanic membrane (ear drum) to vibrate. These vibrations are then passed to the cochlea through middle ear bones or ossicles. Cochlea is a coiled structure, and constitutes a core component in human hearing. Inside the cochlea, are present hair cells which vibrate at different frequencies along the coil, as seen in Fig.2.1. Hair cells also transform vibrations into electrical impulses which are finally transported to the brain through the Cochlear nerve.

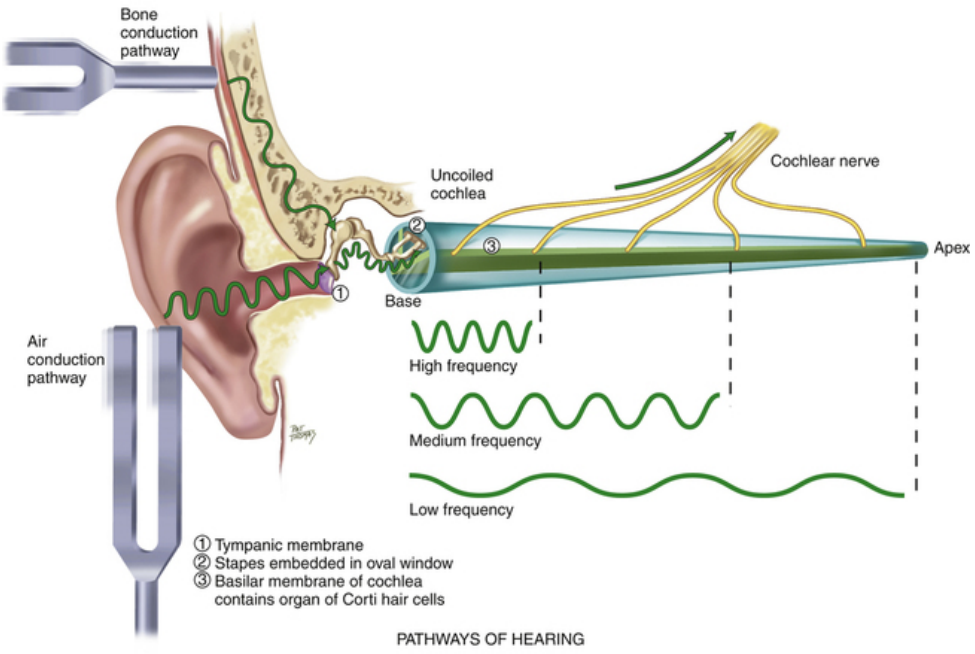


Figure 2.1: Sound Conduction Pathways

As introduced in the section above, skull bone provides an alternative pathway for acoustic energy to propagate from the vocal tract to the cochlea, or the inner ear. This is evident when the ear canal is simply blocked; one can not only hear their own voice, but the signal power also increases. Sound energy propagated through the skull stimulates the bone and cartilage components in the ear, which consequently produces sound pressure in the ear [10]. The pressure then stimulates the ear drum and the energy is passed to the cochlea, where the sound is perceived.

Differences between AC and BC speech primarily arise due to the differences in conduction pathways, specifically from the transfer function of the human skull. In the BC pathway, the skull bone acts as a low-pass filter which attenuates frequencies greater than 2kHz [5]. Such attenuation causes noticeable differences in terms of speech quality and intelligibility between AC and BC speech. In the next section, the outline and construction of a sensor to acquire BC speech is discussed.

2.2 Sensor Outline and Dataset

2.2.1 Multi-modal Sensor and Speech Acquisition

Bone-conducted speech can be acquired using two main methods: bone-conduction sensors or inner-ear microphones facing the ear canal and occluding the ear cavity from external environment. The hearable device used for speech acquisition is an inner-microphone. In-ear microphone was chosen as they are known to have a wider bandwidth than BC microphones and are known to provide more intelligible speech as well. The in-ear microphone used for the project is detailed in [6]. The proposed device is a multi-modal (MMS) in-ear sensor. The following are the key components of the hearable sensor: two in-ear microphones, two EEG electrodes and a foam-substrate. Along with providing structural integrity to the device, the foam substrate also serves to occlude the inner ear from environmental noise conditions, and faithfully capturing speech from the inner ear canal. Fig.2.4 shows the multi-modal sensor used for capturing in-ear speech.

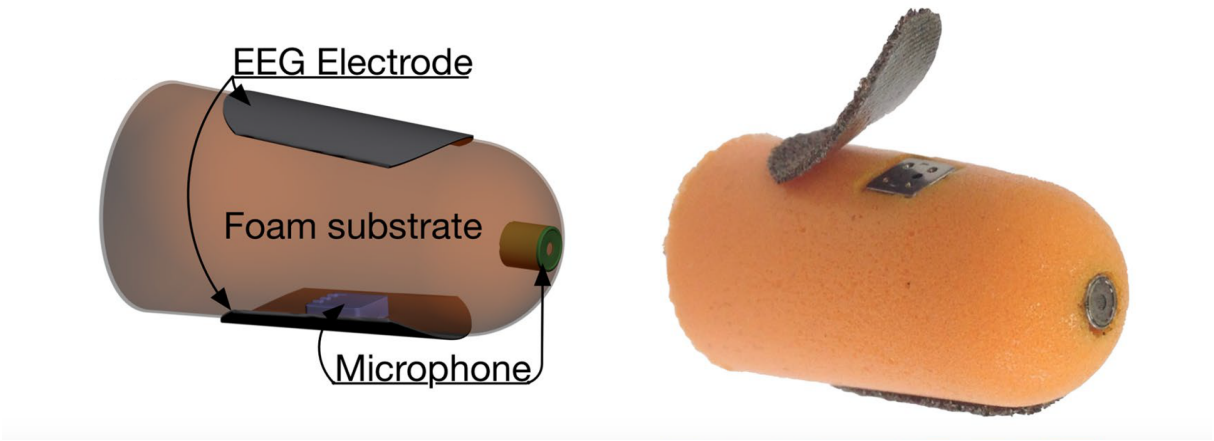


Figure 2.2: Multi-modal Sensor used for in-ear speech acquisition. Speech is recorded from the microphone place at the tip. (Adapted from: Goverdovsky et. al. 2017)

The microphones placed on the earpiece, as shown in Fig.2.4 are electret condenser microphones (ECM), with one placed at the base of the foam substrate, while other being placed at the tip, facing the ear canal. It is due to these microphones the hearables retain the ability to record speech from inner the ear canal. The ECM placed at the base of the multi-modal sensor, uses the principle of bone conduction to capture speech. More specifically the microphone measures the speech signal travelling from the vocal tract to the ear canal, through dense tissue and skull in the human head. In contrast to the base microphone, the ECM facing the inner ear canal directly measures the speech signal evolving from the vocal cords and travelling to the ear canal through the auditory tube in the ear. Experiments done in [6] demonstrate the ECM facing the

inner ear canal outperforms the bone-conducted signal. Therefore, the latter microphone setup is used for this project.

2.2.2 Occlusion Effect in IEMs

For the purpose of this project, only the inner-ear microphone from the MMS is used. Therefore, now the speech capture mechanism in Inner-ear microphone (IEM) will be discussed. IEMs lie between Aerial Ear Microphones (AEM) and Bone-conducted Microphones (BCM). While IEM speech has a wider bandwidth than BCM, IEMs speech has a narrower bandwidth compared to AEMs. Speech is captured using the IEM through the occlusion effect, which blocking the ear cavity and is detailed in [11]. On occlusion of ear canal, the sound pressure increases to the level where frequency structure of BC speech is further affected [12]. Multiple reasons are validated for this phenomenon. [13] suggests that while open ear is a high-pass filter, occluding the ear dominates low-frequency formants. In [14] low-pass filtering effects of occlusion are explained as a consequence of resonance in the ear canal due to blocking.

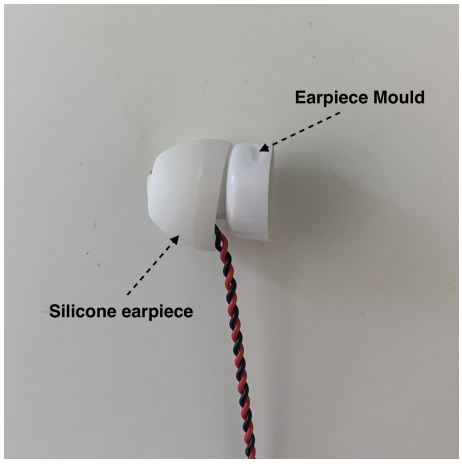
Insulating or occluding the ear from external environment, as done using the proposed MMS sensor, causes the speech signal to resonate inside the occluded ear-canal and be captured by the IEM. Using such approach offers two primary gains. Firstly, the foam substrate attenuates external noise, hence providing larger SNR for in-ear speech. In addition, the IEM signal shares significant frequency information between lower frequency bands ranging from 0-2kHz and high frequency bands (2-4kHz), which is not the case for bone-conducted speech [15]. It is due to this mutual information; the band limited signal can be expanded while preserving its high SNR.

Although occlusion is necessary for capturing in-ear speech, on surveying the literature it was established that the nature of occlusion effect is largely unknown and is heavily dependent various factors such as: speaker voice characteristics, material used for occlusion. Therefore, it was decided to compile a custom-made dataset for external and in-ear microphone speech.

2.2.3 Custom IEM sensor construction and Recording Setup

Following discussing the construction of the original hearable earpiece and its functionality in capturing inner-ear speech, it was decided to construct a custom IEM for this project, which is inspired from the design of the original MMS. This decision was taken as the added functionalities of the MMS, such as the EEG electrodes were of no use to the project. Furthermore, the EEG electrodes required to be attached behind the ear, making the use of MMS complicated for this application. Lastly, by creating an hearable setup which directly records the acquired in-ear speech into the database.

The proposed in-ear microphone is a simpler design, directly tailored towards recording in-ear speech. The device is constructed from a mould of a conventional in-ear headphone. As discussed in section 2.2.2, occlusion of the ear is essential in capturing bone-conducted speech; the mould is fitted with a silicone gel earpiece to occlude the ear. The silicone gel earpiece serves as the foam substrate equivalent in Fig.2.4 in providing solidifying the microphone's position inside the ear and more importantly, occluding the ear from the external environment. Using the structure described above, the earpiece is then inserted with an electret conduction microphone (ECM). The ECM is placed such that it is tightly held into place by the silicon ear gel, and directly faces into the inner-ear canal to record BC speech. The finished hearable is displayed in Fig.2.3.



(a) Side View: displaying earphone mould and silicone seal for occlusion

(b) Top View: displaying the ECM facing the inner-ear canal

Figure 2.3: Construction of custom hearable device for capturing in-ear speech

Once the in-ear microphone was constructed, it was used to record in-ear speech, while the same ECM was placed near the speaker's mouth to record AC speech for simultaneously recording BC and AC speech to establish a speech database to be used for calculating a transfer function between in-ear and external microphone recordings. In order record external and in-ear microphone speech simultaneously, speech data in form of 120 phonetically balanced sentences is recorded from a male speaker in an anechoic chamber, so that clean or reference signal (REF) can be acquired. The sentences were extracted from the TIMIT[16] and Harvard speech databases[17] . Speech is recorded using a late-2013 MacBook Pro at sample rate of 44,100 Hz at 16-bit quantization, but then downsampled to 16kHz. To ensure synchronisation of AC and BC speech, an external USB Dual-channel Sound Card is relayed between the microphones and the computer. The in-ear microphone and the external ECM are both connected to each channel, and the stereo sound card is turn connected to the laptop where speech is stored. The following diagram illustrates the recording setup for the experiment.

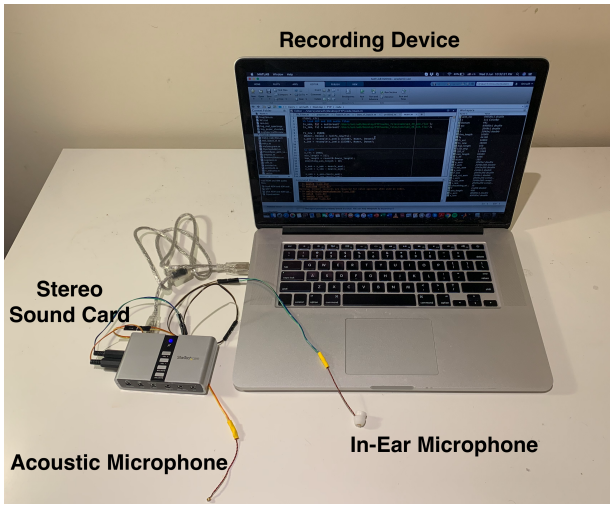


Figure 2.4: Recording setup used for simultaneously recording In-ear and external speech

2.3 Experiments in IEM vs AC speech

2.3.1 Time-Domain Representation

On recording the IEM-AEM pair speech database, this section plots time series signals and spectrogram in order to draw comparisons between the two acquisition methods. Fig.2.5 shows the mean and amplitude normalised plots for a speech utterance chosen from the database. The utterance records ”George seldom watches daytime movies”. It should be noted that the signal was normalised so that both IEM and AEM signal have equal loudness. However, before normalisation is was noted that IEM speech sounded louder than the corresponding AEM recording. More specifically, IEM amplitude ranged from [0.146, -0.129], while AEM amplitude from [0.083,-0644]. Therefore, suggesting a minor amplification in IEM signal. In addition, after normalisation it is seen in the silent regions that amplitude of environmental and sensor noise is lower for IEM speech. However, qualitatively it can be noticed that some high-frequency regions in the IEM speech have been damped.

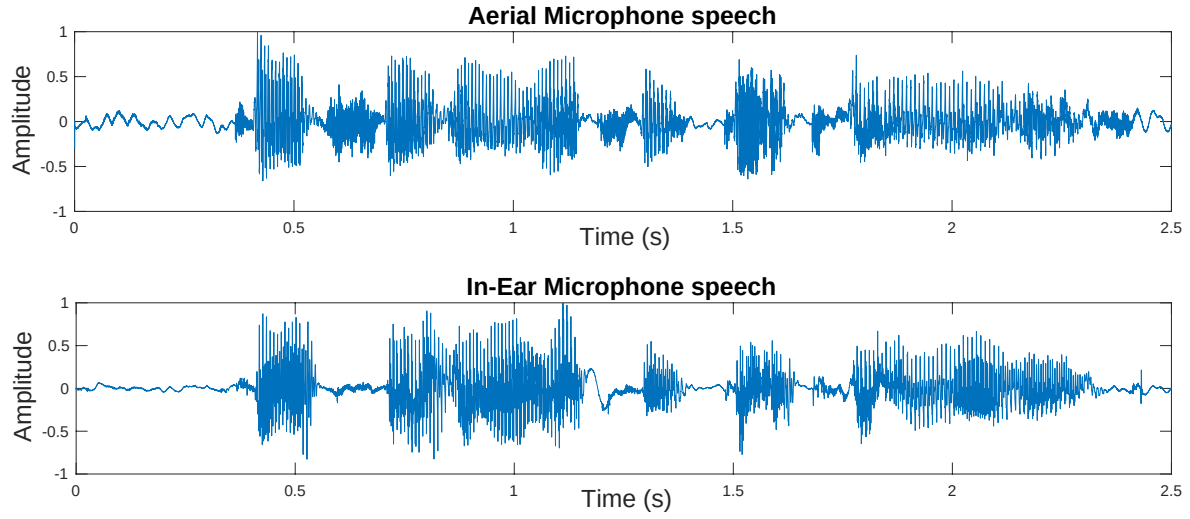


Figure 2.5: Mean and Amplitude-normalised time-domain representation of External and In-ear microphone speech signal

2.3.2 Power Spectral Density

Time series representation, however; does not inform of the frequency difference between the microphones. The Power Spectral Density (PSD) is plotted. PSD is plotted using the Welch overlapping window method. Speech sampled at 16kHz is framed into overlapping windows of 25ms (400 samples), with 6.25ms (160 samples) of overlap between consecutive windows. Window length is chosen such that each windowed section displays statistically stationary properties in the interval. In regards, to overlap, overlapping window help eliminate any transients at window edges by averaging then between multiple windows. Once the speech is split into windows, a Fast-Fourier Transform (FFT) is performed on each frame. On performing the FFT, there will be multiple PSD estimate. These PSD estimates are averaged to obtain the final PSD

plot. This process can be represented by the following equation:

$$\hat{P}(\omega_n) = \frac{1}{KLU} \sum_{i=0}^{K-1} \sum_{k=1}^{N-1} w[k]x[k+iD]e^{-j\omega_m k} \tag{2.1}$$

Where $\hat{P}(\omega_n)$ represents the PSD at a given frequency, K and L denoting the number of windows and window length respectively. Using eq.2.1 PSD estimates for both in-ear and external microphone speech are plotted and shown in Fig.2.6. Before plotting the PSD, a log-transformation is applied to the PSD vector. Applying the log-transformation places emphasis on smaller variations in frequency. In addition, it also corresponds with human auditory perception of frequencies in sound, as human ear perceives sound on the logarithm-scale as well. PSD is plotted to the Nyquist frequency (8kHz).

As discussed in section 2.1, PSD plot shows IEM causes attenuation in the AEM spectra, with most prominent changes in the high-frequency regions. The cut-off frequency is approximately at 1.5kHz. Although high-frequencies are attenuated, frequencies lower than the cut-off frequencies are amplified. This amplification of lower frequencies can be a possible cause for IEM recorded speech's dampened and boomy sounding characteristic. It should be noted, however; that the harmonic structure of two methods is very similar. This is evident as most peaks are on the plot are aligned, especially between 2-5kHz. Furthermore, attenuation is largest between 2-5kHz and improves for frequencies greater than 5kHz.

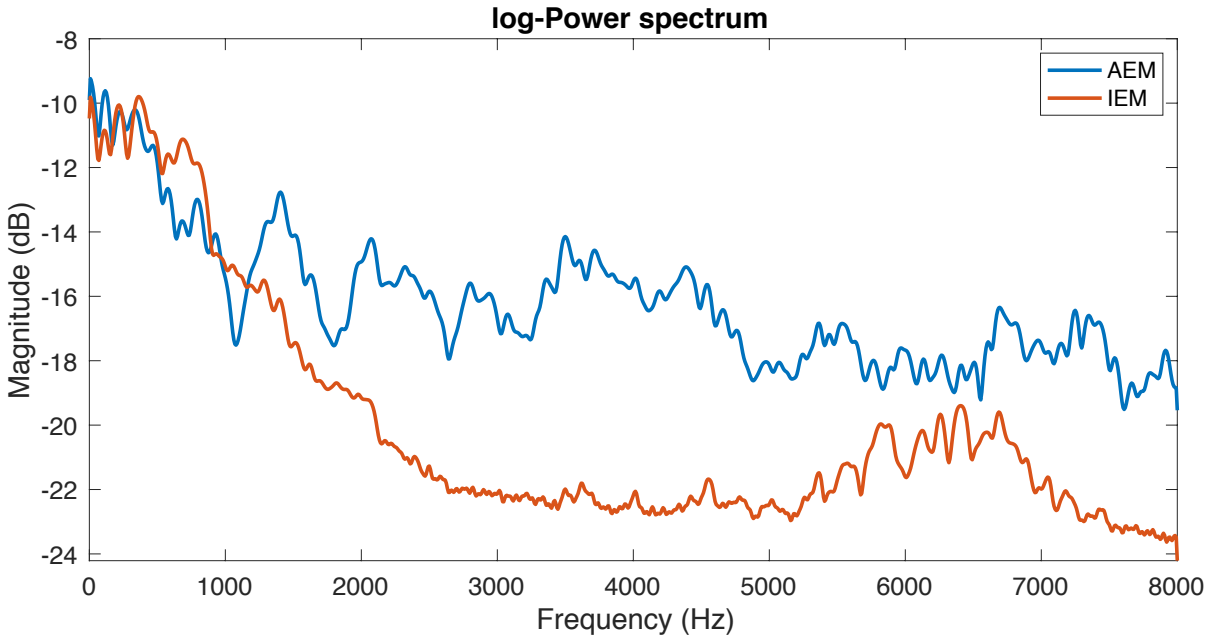


Figure 2.6: Power Spectral Density representation of External and In-ear microphone speech signal

2.3.3 Spectrogram Representation

Since PSD is a static representation of the frequency content of the speech signal, no information on the temporal variation in frequency and energy can be retrieved from PSD plots. Therefore, spectrograms are commonly used in displaying speech signals. A spectrogram is fundamentally a

collection of PSD plots of a signal, at each time instance. In order to produce a spectrogram, the short-time FFT (STFT) is used, where short-time denotes computing FFT for individual frames. Time is denoted on the x-axis, and frequency on the y-axis. In addition, signal power at a particular instant is encoded in form of colour intensity.

Fig.2.7 shows the spectrogram representation of AEM and IEM audio pair. Firstly, it should be noted that the spectrogram is plotted between frequency ranging from [0, 3400Hz]. The decision was taken in accordance to when experiments conducted showed no significant differences in quality by low-pass filtering the 8kHz bandwidth to 4kHz. From the IEM spectrogram(Fig.2.7 bottom) the cut-off frequency can clearly be inferred to be around 1.8kHz. Secondly, the spectrogram explicitly shows the harmonic structure of speech. It is evident that there are significant similarities in the low-frequency range (0-2kHz) between the two microphones.

It should be noted that energy is concentrated in the low-frequency region of speech. These energy concentrated bands (dark on the spectrograms) in frequency ranging 500-1000Hz are known as formants. Formants are consequence of resonance of acoustic energy in the vocal tract, and each formant corresponds to an individual fundamental frequency [18]. On a PSD plot formants will can be identified as sharp peaks. Formants lead to equally spaced energy levels on frequency which are called harmonics. Inspecting the in-ear microphone spectrogram, it is evident that IEM has caused higher-order harmonics to dampen. Furthermore, impulsive vertical lines are observed in IEM captured speech. These sharp impulses were found to be clicking/ popping sounding artefacts, and were another source of distortion in the IEM spectrogram.

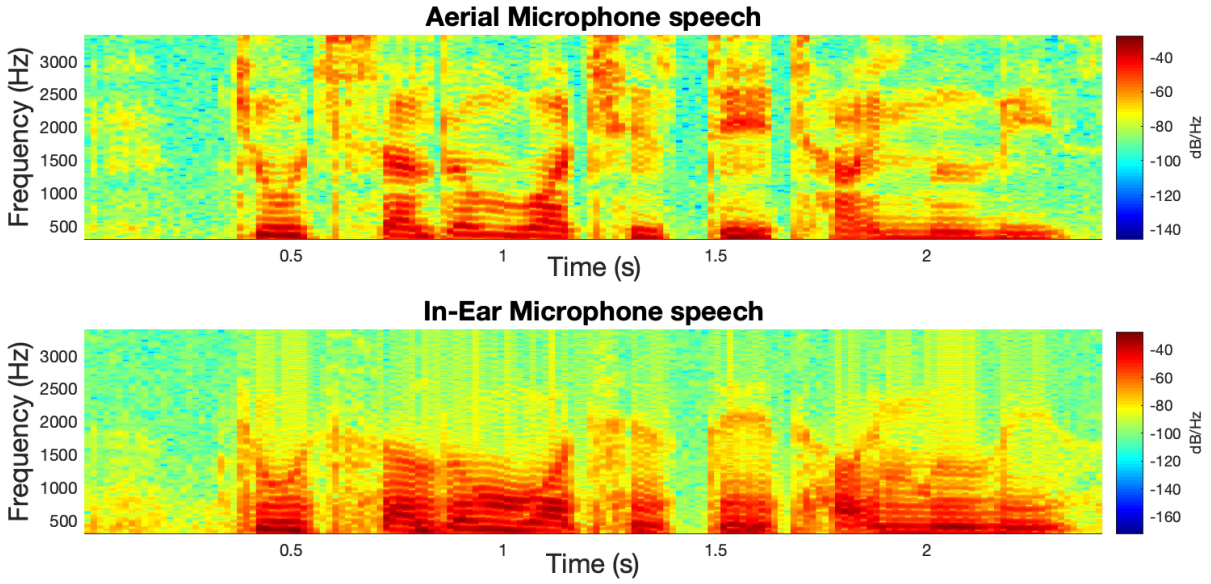


Figure 2.7: Spectrogram representation of External and In-ear microphone speech signal

2.3.4 PSD and Spectral analysis of vowel sounds and other signals

In this section, AEM-IEM audio pair is recorded for fundamental sounds during interactions and the effect of bone-conduction is investigated. These fundamental sounds include vowel sounds in the English language, breathing sound - inspiration and expiration, and finally coughing

noises. The experiment is conducted to investigate if bone-conduction channel affects these different sounds in similar or different manner. Firstly, the vowel sounds are examined. Spectrum of vowel sounds are shown in Fig.2.8. The spectrum also plots the error between their respective AEM-IEM spectra. There trend in spectral error is roughly similar, for both phonemes. Although there are variations in the error. In this case, along the 1.5-2kHz region. Variations in log-spectra error between phonemes can be attributed to the variation in difference phonemes, which are produced by varying shapes of the mouth, tongue and jaws. Such variation in sound production can lead different transmission route to the cochlea, hence having a different transfer function.

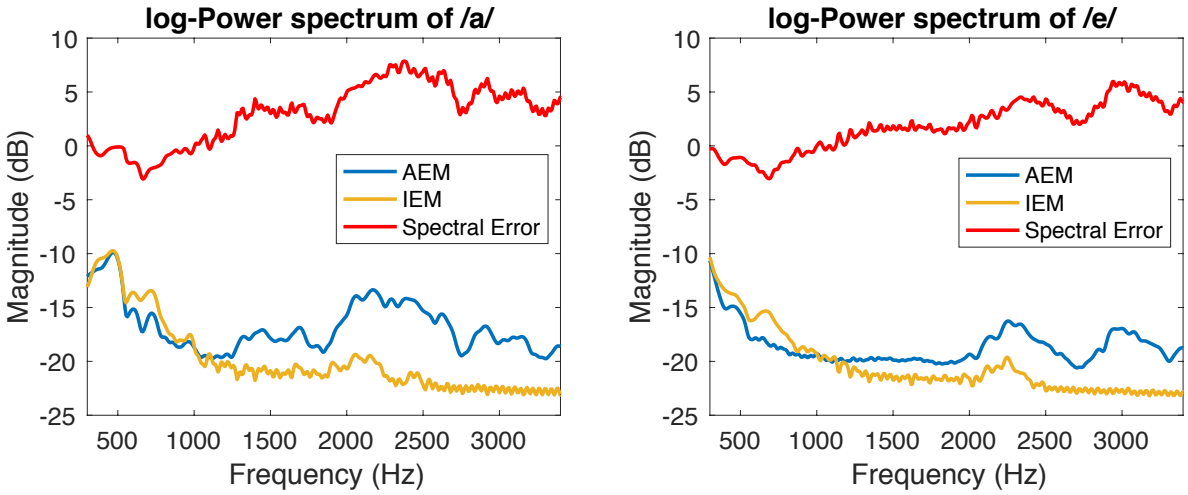


Figure 2.8: PSD of vowel sounds /a/ and /e/ for AEM (blue) and IEM (chrome) microphones. Spectral error between AEM and IEM shown in red.

While vowels are an essential aspect of English speech, breathing and coughing; on the other hand, are known to reduce speech intelligibility and hence considered as noise. To capture breathing, the subject was asked to take deep breaths at a constraint rate. Fig.2.9 shows that there is significant difference between breathing sounds. Firstly, as expected, the IEM recorded breathing is band limited at 2kHz. In addition, the breathing sounds are significantly more prominent, with greater signal power, in IEM recorded signal. However, due to the distortion causes from the bone-conduction pathways harmonic structure in breathing sounds is not observable in IEM. In its AEM counterpart, the harmonics are visible with fundamental peaks around 1kHz. In regard to the coughing spectrogram, there are minor differences between the two microphones. The bandwidth of IEM signal is 2.5kHz wider than conventionally seen in prior experiments in speech and breathing signals.

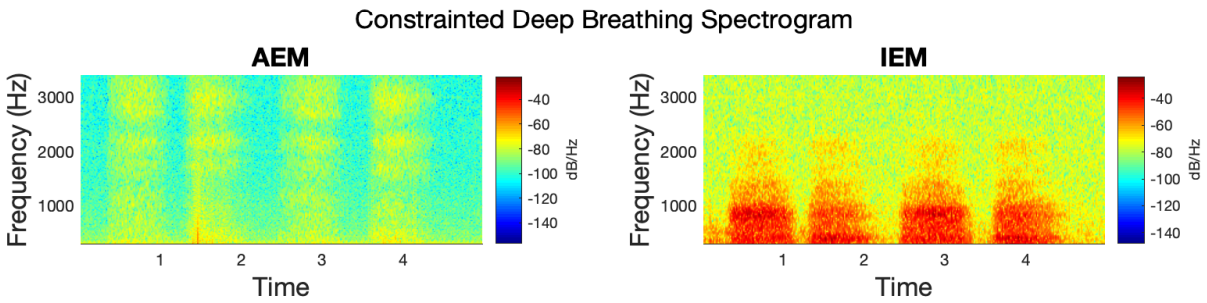


Figure 2.9: Spectrogram representation of External and In-ear microphone Breathing

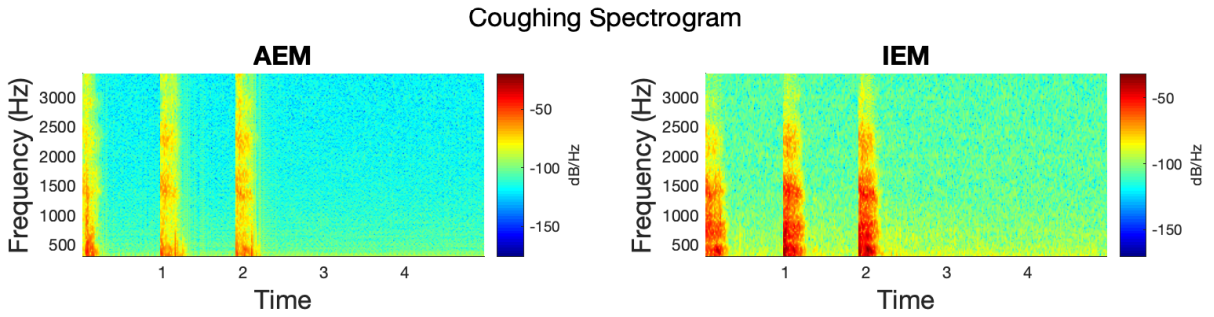


Figure 2.10: Spectrogram representation of External and In-ear microphone Coughing

2.4 Clustering Methods

In this section, a brief background for K-means clustering is provided. K-means clustering is used for the Phoneme-based transfer function mapping approach, hence requires to be discussed. K-means clustering is an unsupervised clustering method that partitions observations into clusters by computing cluster-centroids. Given a set of observations $\mathbf{x} = \{\mathbf{x_1}, \mathbf{x_2}, \ldots \mathbf{x_n}\}$, where each observation spans the feature space R^D, the algorithm aims to compute K mean centres which are cluster centres. The parameter K, is a *a priori* hyperparameter, which determines the number of cluster the data is to be segmented in. Partitioning is done in such manner that within-in cluster between point is minimised. The high-level algorithm for the approach is as follows[19]:

- **Step 1**: Set the number of cluster, K.
- **Step 2**: Initialise cluster centres $\{\mu_1, \mu_2, \ldots \mu_\mathbf{K}\}$.
- **Step 3**: Classify each sample $\mathbf{x_n}$ in the dataset, to the nearest cluster centre $\mu_\mathbf{K}$.
- **Step 4**: Re-compute cluster centres with new entries and update centres.
- **Step 5**: Terminate process when no significant change in $\{\mu_1, \mu_2, \ldots \mu_\mathbf{K}\}$.

Conventionally cluster centre initialisation is performed by selecting random samples the dataset. However, this method causes the possibility of the cluster centres being in close proximity together. Therefore, the implementation used in the project is the k-means++ algorithm[20]. The aforementioned algorithm initialises the cluster centres by setting random seeds. Following cluster initialisation, a distance metric is required to classify novel observations to existing cluster. In this project, L2-norm or Euclidean distance, is utilised as the distance metric.

2.5 Literature Survey

Recreation of acoustic microphone (AM) speech from bone-conducted (BC) speech is a difficult task. This is primarily due to the variation of the transfer function between speakers and sounds as well. Furthermore, the transfer function is also shown to be dependent on the method of occlusion and quality of the in-ear microphone used in the study [21]. In the existing literature, there is more research directed towards BC sensors, in contrast to IEMs. However, since both methods suffer from bandwidth attenuation problems; similar techniques can be used on both speech capturing methods. These methods are mainly divided into the following approaches: **(1)** Filter-based, **(2)** Bandwidth-Extension (BWE) and **(3)** Machine-learning techniques. The literature review details these approaches and the datasets utilised in recreating AM speech.

2.5.1 Databases

On surveying the literature, it was found that most methods have used custom-made datasets. For instance, [22] recorded 50 phonetically diverse sentences in Japanese by 3 male and female speakers in quiet conditions at 16kHz. The neural network based approaches proposed in [23] recorded 12 mins by 10 speakers in 600 speech utterances. In addition, [24] proposes a bilingual (English and French) database of IEM and external microphone. The speech corpus is also recorded under quiet conditions by a female speaker at 8kHz. The spoken speech are 100 sentences from the Harvard phonetically balanced sentences database [17]. Most of the datasets mentioned here and elsewhere in the literature are not made publicly available. Therefore, a custom dataset was used for the project, as outlined in section 2.2.3.

2.5.2 Evaluation Metrics

The evaluation plan methods which were reviewed for evaluation of the transfer function or method used for mapping IEM speech to air-conducted speech. Numerous methods are highlighted in [25]. These methods are largely divided into subjective and objective quality assessment methods. For the purpose of this project only objective quality assessment methods are considered due to the limited time frame and current conditions to conduct subjective testing. The following are the performance metrics utilised in the literature.

- **Perceptual Evaluation of Speech Quality (PESQ)**: PESQ is most widely used measure across the speech enhancement research community. The measure provides a rating ranging between -0.5 to 4.5. This is the wide-band estimator recommended in ITU-T P.862.2. [26]. PESQ estimator requires both, the clean and the enhanced signal, hence making it an intrusive measure. In addition, for this reason, it can only operate on end-to-end systems.
- **Cepstral Distance(CD)**[27]: The distance metric represents cepstral mismatch between spectrums, and is measured in dB. Initial experiments done on difference between in-ear and externally recorded speech in [6] uses the CD as evaluation metric.
- **Log-Spectral Distortion (LSD)** LSD is used for comparing spectral representation of speech, hence beneficial in comparing spectral-based method . LSD is a distance metric which measures the spectral mismatch between two power spectrum. Given the PSD of two signals, the LSD metric is represented as follows:

$$LSD(dB) = \sqrt{\frac{1}{N}\sum_{\omega} 10\log\frac{P(\omega)}{\hat{P}(\omega)}} \tag{2.2}$$

2.5.3 Analytical Methods

In the past various methods have been used to enhance bone-conducted speech quality. In [28] a reconstruction filter is derived using the long-term spectrum of the bone-conducted (BC) speech. It is assumed that BC speech can be obtained by passing the AM speech through a linear filter. A linear reconstruction filter is constructed using the log-spectra ratio. Results show improvement in speaker recognition rates. However, minimal consistency was showed between test phrases and speakers.

One approach to compute a transfer function is to recover the spectral modification, namely the attenuation in high-frequency bands, introduced as a consequence of recording from an IEM. [22] recovers the aerial microphone signal by applying an equalisation gain for each frequency bin. The gain is calculated as the ratio between the magnitude of FFT spectrum of air-conducted and IEM speech. Since FFT segments speech into various frames, multiple equalisation gain will be available. Therefore, a moving average is taken to get the average gain for each utterance. The approach notes improvements in CD. Authors in [29] use a similar method as well. More importantly, both [22],[29] demonstrate that the transfer function is speaker and microphone dependent.

[30] proposes BWE approach as solution for improving the speech quality in the high-frequency region. BWE is a non-linear approach which aims to utilise the mutual information between low and high frequency bands in generating harmonics in the attenuate frequency bands. The BWE approach in [30] is performed under quiet conditions. This is because applying BWE under noisy conditions, will also amplify noise. BWE approach under noisy condition has proven to be computationally expensive and slow as evident in [31], [32]. However, according to [23] bandwidth extension approach is criticised for its inability to enhance the degraded low-frequency bands.

2.5.4 Machine-learnt Methods

As seen above, the transfer function between IEM and aerial microphone is non-linear and complex to generalise using linear filtering methods. Therefore, several approaches have used non-linearity of neural networks in computing the transfer function. [23] uses a hybrid approach in recovering the spectral envelope of air-conducted speech. This is done by first passing the time-domain IEM speech through a high-band booster. Mel-frequency cepstral coefficients (MFCC) of boosted IEM speech are inputted into a two-hidden layer neural network. The neural network outputs MFCCs of the estimate air-conducted speech. However, band-energy conversion is done in time-domain to achieve low complexity, by avoiding calculating inverse DFT and using overlap-and-add. Results show the proposed method outperforms the linear transfer spectral Gain approach in terms of Average Log-spectral distortion.

[33] propose a Deep Neural Network (DNN) approach for restoring the low-order cepstrum of bone-conducted speech. Restoring low-order cepstrum is motivated by [34], as authors demonstrate the spectral envelope of bone-conducted speech significantly deteriorates the quality of bone-conducted speech. More specifically, [33] assumes both AM and bone-conducted speech have the same excitation source (vocal cords); however due to different propagation media, both pathways have different transfer functions. Using this, DNN aims to transform the log-amplitude spectrum of the bone-conducted speech to the low-order cepstrum of AM speech. Following this the low-order cepstrum of AM speech is concatenated to the cepstrum of input bone-conducted

speech, to provide the estimate AM restoration. Results show the proposed method outperforms the linear filter approach in [28] in terms of Mean Opinion Score (MOS). Furthermore, authors show cepstrum estimation is better in estimating high-band energy rather than the DNN directly estimating the log-amplitude spectrum of the estimated AM speech restoration.

In addition, [35] proposes a multi-channel approach by using 2 IEMs and a fully convolutional network (FCN) to map IEM speech to AEM speech. Performance is evaluated using Perceptual Evaluation of Speech Quality (PESQ). Results show improvement in PESQ compared to unprocessed multi-channel speech. Furthermore, authors utilise dilated convolutions to further improve on the previous results. The proposed method has a high computational complexity and is heavily data-intensive, as authors used 7138 speech utterances. It should be noted that this approach could be intricate mainly due the lack of speech data.

2.6 Chapter Summary

In this chapter, firstly the Background content behind IEM recorded speech is discussed. Firstly, in-ear speech conduction pathways are defined, and the process behind capturing IEM speech is outlined. In this discussion, the occlusion effect is highlighted. The occlusion effect occurs by blocking the ear-canal successfully record speech. However, occlusion is also the cause of the boomy and ringing sounding characteristics of IEM speech. Following this the construction for the custom-made IEM is described. The sensor constructed for the project is inspired from [6]. On constructing the sensor, the simultaneously recorded database of AEM-IEM speech is discussed in section 2.2.3.

In section 2.4, various methods of speech representation are discussed. It was shown through PSD plots and spectrogram that cut-off in IEM signal bandwidth is approximately 2kHz. It is was also established that the relationship between AEM-IEM transfer function is dependent on individual sounds (phonemes).

In the last section of the chapter, a literature review on IEM speech enhancement is conducted. In this context enhancement refers to improving bandwidth of IEM speech. It is was found to be achieved in 3 main methods: A transfer function approach, bandwidth extension, and machine learning based methods. Along with methods, datasets are also reviewed. It was found that datasets are highly customised and dependent on the equipment used for recording, hence justifying the need to produce a custom-made dataset. In the next chapters, Analysis and Design and Implementation for the proposed transfer function computation techniques are outlined.

Chapter 3

Artificial Bandwidth Extension

This chapter proposes an artificial bandwidth extension scheme, for mapping IEM speech to AEM counterpart. It is a time-domain blind techniques which aims to extend the bandwidth of signals by leveraging information in the lower bands.

3.1 Analysis and Design

Artificial bandwidth extension method (ABWE) is a time-domain technique, which is originally used in telephony to extend the bandwidth of telephone signals from 4kHz to 8kHz, over the telecommunications channel. The principle behind BWE is to leverage the information in the lower frequency bands to estimate the attenuated frequency content. Since the IEM microphones also suffers from the problem of bandwidth attenuation, bandwidth extension can prove to improve intelligibility of IEM speech. A flowchart of bandwidth extension framework in Fig.3.1. The method operates on individual speech frames in the time-domain. The use of this method is adapted from work done on bandwidth extension in [30]

Firstly, the speech frames are up sampled by a factor of 2. Upsampling causes spectral folding, hence generating high-frequency harmonics. Following this the speech signal is passed through a whitening filter. The whitening filter is an LPC filter approximation of the current speech frame, s_i. Subsequently, the whitening filter output is passed through a non-linear function. It is the non-linear function which helps to extend the bandwidth in the high-frequency region using the information in low-frequency bands, ranging from 0-2kHz. The output of the excitation signal is then combined with the whitened signal. Lastly, the band-extended signal is filtered to extract the relevant frequency band information, which in this case, ranges from 300-8000Hz. Next the design choices and the requirements for the main modules, shown in Fig.3.1 are discussed.

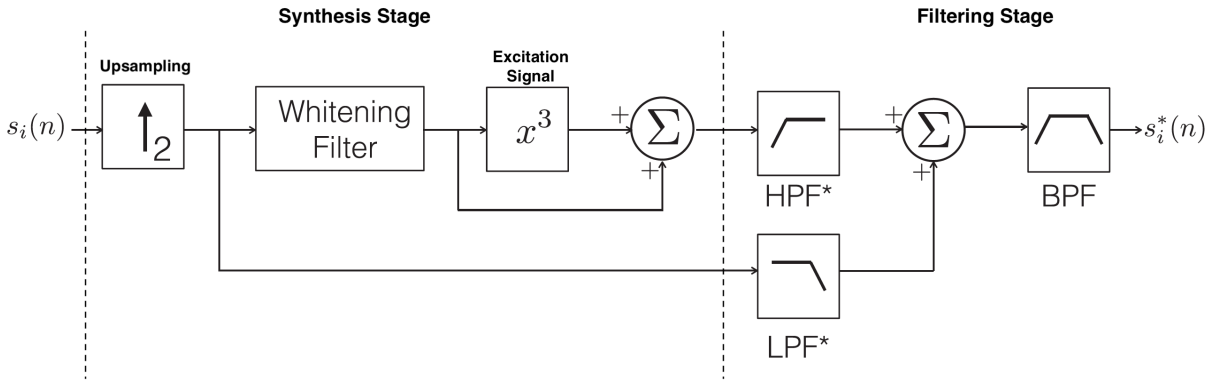


Figure 3.1: Artificial Bandwidth Extension algorithm system-level flowchart. Adapted from Voix. et.al. 2015

Upsampling

Upsampling is the first module which extends the narrow band signal. This is achieved by adding zeros in between the samples in the time-domain speech signal. Consequently, in the frequency domain, upsampling causes spectral folding, which is seen as mirroring of the spectrogram along the Nyquist frequency, in this case 8kHz. Fig.3.2 shows the resulted upsampled spectrogram. A particular advantage of upsampling is the mirroring property helps to recover high-frequency utterances or unvoiced utterances, ranging from 4-8kHz,which are completely attenuated by the bone-conduction channel. In addition, it has been shown the temporal characteristics of unvoiced utterances below the bone-conduction channel cut-off frequency are significantly correlated to the frequency content above cut-off [36].

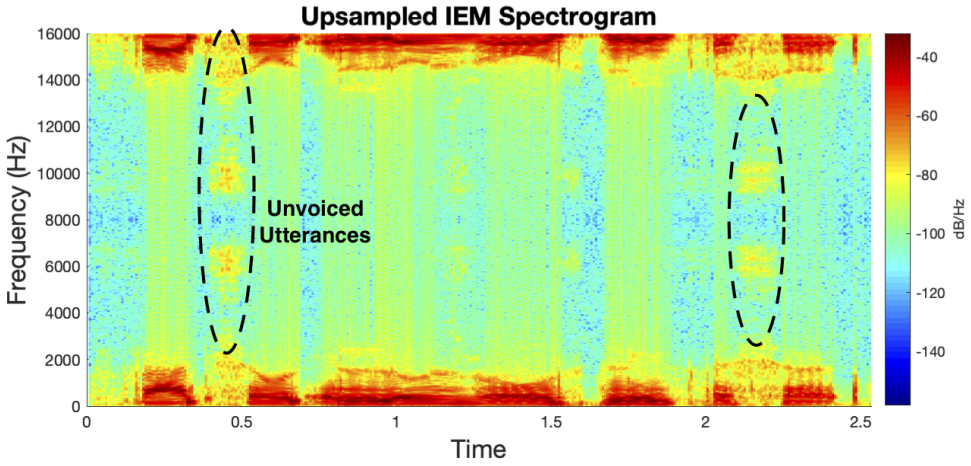


Figure 3.2: Artificial Bandwidth Extension algorithm spectrogram

Excitation Signal Generation

The excitation filter can mainly be divided into stages - the Whitening Filtering and the Excitation signal. The whitening filter is Finite Impulse Response (FIR) filter, which aims to generate the excitation signal from the narrowband speech frame. The basics of the excitation lie in the Source-Filter model for generation of speech, which now be discussed briefly.

Whitening Filter

The Source-Filter Model is based on anatomical analysis conducted on the human speech production mechanism. The source aspect is a model of the sound-production apparatus, being the lungs and the vocal cords. The filter aspect, on the other hand, aims to model cavities and transmission pathways from which the produced speech is filtered. These include the Nasal, Oral and the Pharynx cavities. The air pressed from the lungs exert pressure on the vocal cords, which leads them to behave in the two following ways. First case being the vocal cords are loose, and the pressure causes turbulent air-flow. Second being the vocal cords are contracted and closed. Increase in pressure causes the vocal cords to open, and consequently release the pressure. On releasing the pressure vocal cords shut once again. Hence, the outlined process takes place in a periodic manner [36]. The signal exciting the vocal cord activity is called the excitation signal. The source-filter theory aims to model this phenomenon using the excitation signal generators.

These excitation signals are of two types: (1) the unvoiced utterances, generated due to the turbulent air flow in the vocal cords, and modelled using a noise generator. (2) Voiced utterances, which are formant sounds, and is modelled using a periodic signal.

For estimating the effect of filtering cavities on the produced sound, an autoregressive model (AR) is adopted. The impulse response on the AR process is denoted by $H(z, n)$, and is given as follows:

$$H(z, n) = \frac{G(n)}{1 - \sum_{i=1}^{M} a_i(n) z^{-i}} \tag{3.1}$$

Here $H(z, n)$ is called the vocal tract function, which aims to model the modulation of filtering cavities in the vocal tract. Therefore, $H(z, n)$ is essentially the impulse response of an acoustic filter, which amplifies and attenuates certain frequency characteristics. Such modelling is called Linear Predictive Coding (LPC). The frequency characteristics of the vocal tract are dependent on the $G(n)$ which is the input from the source part, and $a_i(n)$ are coefficients of the LPC model. Lastly, M being the LPC filter order. In literature, M ranged between 18-25, therefore $M = 18$ is used in this implementation.

Hence, the whitening filter can be interpreted as an 18-th order LPC FIR filter, which is applied to the upsampled signal in order to attain the excitation of the narrow band signal. Since the LPC approximation is linear in nature, it requires the input data to be stationary. Stationarity of speech was maintained by framing speech into 25ms segments, and subsequently extracting the LPC coefficients of that particular segment. The whitened output of the whitening filter in time-domain, $s_w(n)$, can be written as follows:

$$s_w(n) = \sum_{i=1}^{N} s_i(n - i) a_i(n) \tag{3.2}$$

Wideband Excitation Signal

As discussed above, the whitening filter generates the narrowband level excitation signal. The narrowband whitened signal is extended in bandwidth by passing it through a function with non-linear characteristics. The non-linear characteristics can be applied due to the periodic nature of the excitation function, as non-linearities have the property to generate harmonics when applied to periodic signals[36]. These functions extend the narrowband excitation signal by extending the harmonics to the high frequency range, which ranges from above 2kHz to the Nyquist frequency, 8kHz. It should be noted that while upsampling helps in extending the unvoiced utterances, the non-linear excitation function extends voiced utterances in speech by generating higher-order harmonics. Finally, the band-extended speech, $s_i(n)$ is generated by summation of the whitening filter and wideband excitation signal. This is represented as follows:

$$\begin{aligned} s_i(n) &= e_{wb} + s_w(n) \\ &= e_{wb} + \sum_{i=1}^{N} s_i(n - i) a_i(n) \end{aligned} \tag{3.3}$$

Where the e_{wb} is the wideband excitation signal, and is represented in terms of the narrow-band excitation, e_{nb}:

$$e_{wb} = e_{nb}^3 \tag{3.4}$$

In generation of the excitation function, approach apart of non-linear functions were also considered from the literature. This was the modulation function technique. The modulation function is a time-domain multiplication of the narrow-band excitation e_{nb} with a cosine signal of frequency of the cut-off frequency. Time-domain multiplication with cosine leads to the two-side narrowband spectra to be centred around the cut-off frequency, hence extending the voiced utterances. However, the modulation approach requires additional information on pitch tracking, which is information on fundamental frequency.

Filtering Stage

On extending the bandwidth of the upsampled signal there is significant overlap between the high and low frequency components, hence filtering is required to extract the wide and narrow band content, and combine them. Filters were designed such that the cut-off frequency causes little overlap between the high and low frequency components. The optimal cut-off frequency then will be where the IEM cut-off frequency lies. In order to determine the cut-off frequency, the LPC spectral envelope of sounds /a/, /e/ and /o/ are plotted (Fig.3.3). The LPC spectral envelope show the cut-off frequency is at cut-off frequency is at 2100Hz. Therefore, 2100Hz is used as cut-off frequency for the filters.

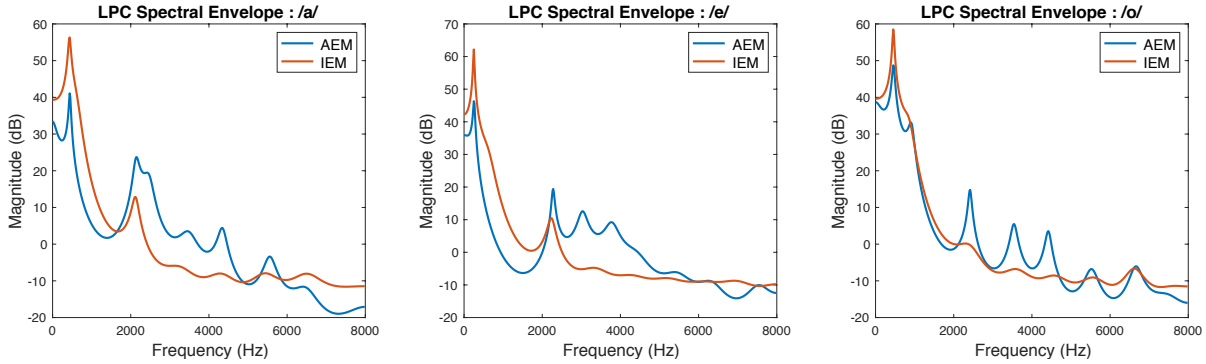


Figure 3.3: LPC Spectral Envelope of AEM and IEM speech for vowels sounds: /a/, /e/, /o/

In regard to the filter design, the estimated wideband excitation signal is passed through Butterworth high-pass filter with a cut-off frequency of 2.1kHz. High-pass filtering isolates the wideband information in the region 2.1-8kHz. On the other hand, the upsampled speech frame (pre-whitening filter) is filtered using a Butterworth low-pass filter with cut-off frequency of 2.1kHz, as determined by the LPC spectral envelope plots. Next the narrow and wide band estimates are added. Following this, the upsampled signal is post-processed using a bandpass filter. The bandpass filter is constructed by cascading a lowpass and highpass Butterworth filter. The need for bandpass filtering arises as bandwidth extension creates harmonics past the Nyquist frequency, and in the lower bands adds low frequency noise. Hence the bandpass filter's passband ranges from 160-8000Hz.

3.2 Implementation

In the implementation of the artificial bandwidth extension, the major system-level modules described in the section above are discussed in finer detail. This is done by discussing the software implementation of these modules. The implementation also aims to present the design and analysis in form of an algorithm, which is then used in the Results and Testing section.

3.2.1 Bandwidth Extension Algorithm

The bandwidth extension algorithm function BWE is shown in Fig.7.1 in the Appendix. The function inputs arguments are the time-domain input speech samples, the sampling frequency, and the frame length over which bandwidth extension is conducted. Along with this, function inputs tuning parameters, these include whitening filter order, and filter cut-off frequency to avoid cross-band information overlap.

The bandwidth extension function firstly iterates over the time framed speech. The frame size in the implementation is set to 400 samples. For each 400 sample utterance, the speech is firstly upsampled to provoke spectral folding. Next the upsampled time frame is passed through the whitening filter. As discussed in the previous section, the whitening filter is an FIR filter, with the filter coefficients being the LPC coefficients of the present speech frame. The order of the LPC filter used was $M = 18$. Hence, the whitening filter is an all-zeros filter of filter 19 coefficients. The output of the filter is whitened spectrally flat speech, with the excitation signal of the wideband speech. The estimated wideband excitation signal is passed through the non-linearity function. Non-linear function use in this implementation is a cubic non-linearity. The applying the cubic non-linearity to the whiten speech frames generate odd-harmonics along the higher frequency range.

Next is the implementation of the filter stages. As aforementioned firstly, a combination of power normalised high-pass and low-pass filter is implemented to isolate the relevant information from the frequency band. The high-pass filter is a 3rd-order Butterworth filter with cut-off frequency at 2kHz, as determined from LPC spectral analysis. Similarly, the low-pass filter is 3rd-order Butterworth filter with cut-off frequency at 2kHz as well. The low-pass filter retrieves information from the lower bands. Next the filter outputs are summed to obtain the BWE speech. The bandwidth extended signal ranges from 0-8kHz, however; the region of interest is the voiced frequency bands which ranges from 300-4000Hz. Therefore, 4-th order Butterworth bandpass filtering is used to extract the relevant frequency regions.

3.3 Chapter Summary

This chapter presents the analysis and design, and the implementation of a time-domain method, called artificial bandwidth extension, for bandwidth extending the IEM speech to improve speech quality. This is a filter-based approach where the first the excitation signal of narrowband speech (IEM) is estimated using whitening filter. Subsequently the wideband excitation is generated using a non-linear function, x^3 to extend bandwidth by generating harmonics. Lastly, there is the filtering stage to extract band relevant information. Next the implementation is presented, which discusses the approach in algorithm format. Code is provided in appendix (Fig.7.1).

Chapter 4

Spectral Equalisation Approach

In the previous chapter, Artificial Bandwidth Extension, a time-domain bandwidth extension scheme is implemented. One of the pitfalls of an estimation scheme such as BWE is its inability to learn the dependency between the IEM and AEM transfer function. In this chapter, the problem of mapping in-ear speech is approached from a frequency-domain transfer function approach. The chapter proposes the development of an equalising transfer function by investigating potential factors which determine the variability in the transfer function - specifically speaker identity and phoneme dependence.

4.1 Analysis and Design

Utilising the established speech database, a speaker-dependent equalising transfer function approach is adopted in enhancing the IEM speech. Equalisation is done by computing the short-term FFT ratio of air-conducted and bone-conducted speech frame on a training set. Once transfer function is learnt, it is applied to unseen speech. A system overview is shown in Fig.4.1. The overview of the model is as follows:

1. Pre-processing: Perform Voice Activity Detection (VAD) and prune silent regions of speech. This is then followed by the subsequent enhancement of speech using the MMSE algorithm.

2. Feature Extraction: Obtain STFT spectral features of the corresponding AEM-IEM speech pair

3. Compute transfer function: The spectral gain is computed over the training set and mean spectral gain is stored.

4. Model Validation: Next open training is performed on unseen speech utterances.

4.1.1 Pre-processing

Speech from both microphones is pre-processed in the following steps. Firstly, speech is normalised in form of peak normalisation, where the DC-component is removed from then signal by subtracting the mean, and dividing signal samples by the highest value. Next step is to prune silent sections from the speech signal, in order to implement this a spectral energy Voice Activity Detector (VAD) is used. Since the design of the VAD is outside the scope of the project, and will require considerable time implementing one, an implementation from [37] is used.

While recording, there were several manifestations of noise. First is the noise from the microphone, when recording the microphone adds mains noise at 50-100Hz in the recordings. Such noise can be heard in form of a buzzing sound. Furthermore, while speaking multiple artefacts were noticed especially in the In-ear microphones. These noises can be attributed to bodily movements of jaws, ears and tongue which further corrupt signal quality. Hence, justifying the need for speech enhancement. Speech quality enhancement is done using the Short-Time Spectral Amplitude (STSA)-MMSE [38] approach. This method was chosen as it is

a statistically optimal method in enhancing speech quality in an MMSE sense. Similar to VAD, speech enhancement module is outside the defined project scope. STSA-MMSE implementation was used from the VOICEBOX speech processing toolbox [39]. Furthermore, literature shows the STSA-MMSE method does not pose any assumptions on data linearity, unlike Wiener Filter. STSA-MMSE methods compared to other enhancement schemes do not introduce artefacts (musical noise) as well.

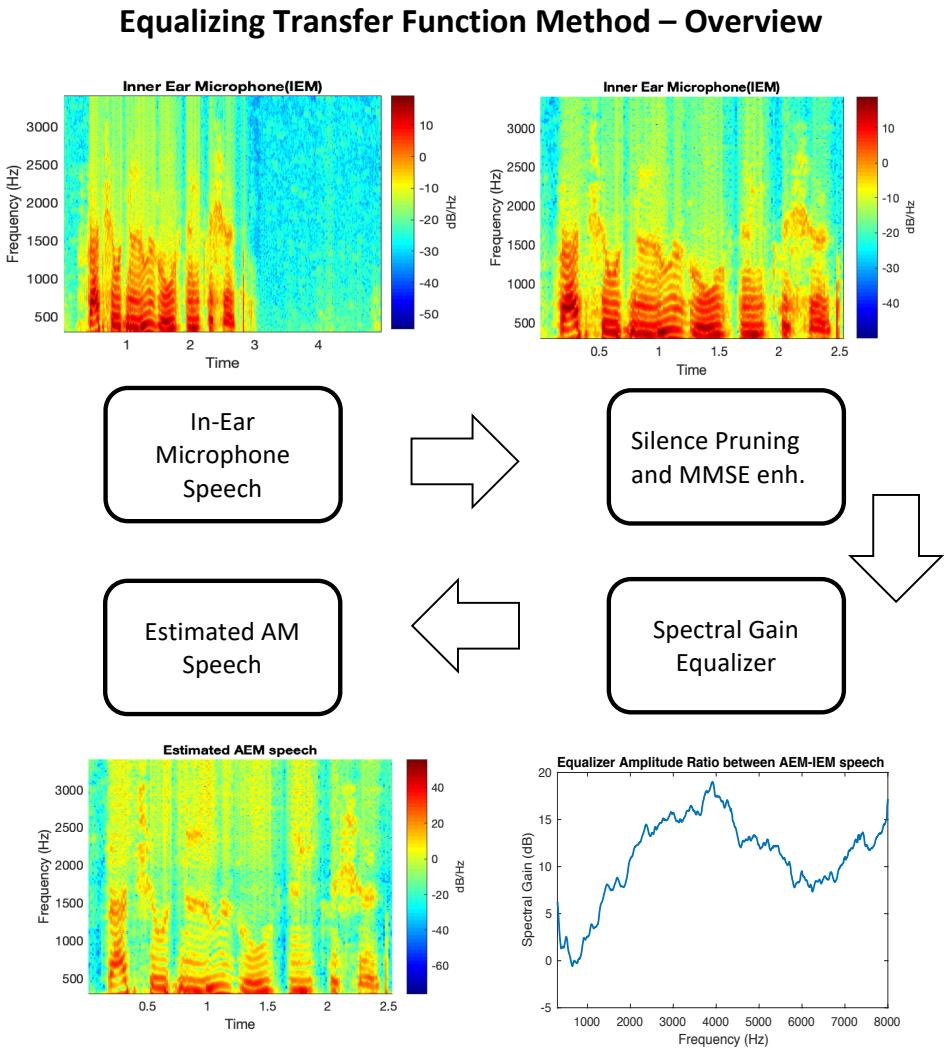


Figure 4.1: Spectral Equaliser Transfer function system-level flowchart

4.1.2 Feature Extraction and Spectral Gain

Following the pre-processing of the data, the time domain speech signals are prepared for estimating the in-ear microphone transfer function. In this approach, a spectral gain or spectral equalising (EQ) function is to be estimated. In order to estimate the EQ-function, STFT of AEM-IEM speech pair is computed. Specifically, the EQ is a frequency dependent function, which is determined by the gain between the magnitude of the STFT of AEM and IEM signal. This can be represented as follows:

$$|\mathbf{H_{EQ}}| = \frac{|\mathbf{H_{aem}}|}{|\mathbf{H_{iem}}|} \tag{4.1}$$

Where $|\mathbf{H_{EQ}}|$ represents the magnitude spectrum of the equalising function, and $|\mathbf{H_{aem}}|$ and $|\mathbf{H_{iem}}|$ being the magnitude spectrum of AEM and IEM speech respectively. It should be noted that $|\mathbf{H_{eq}}|$ is column vector of size the number of frequency bins (k) in the spectra. Where k is determined by the size of the ST-FFT, which is denoted as N-FFT. The frequency spectrum ranging from 0-f_s Hz, f_s being the sampling frequency is divided into N-FFT frequency bins. However, according to Nyquist's theorem; the FFT spectra ranges from 0-$\frac{f_s}{2}$Hz. Therefore, the size of each STFT magnitude is column vector of $\frac{N-FFT+1}{2}$. In addition, the STFT segments the speech signal into overlapping frames. Therefore, such an $|\mathbf{H_{EQ}}|$ estimate will be calculated for each frame. In regard to this, each equalisation function estimate can be written in form of indices of particular frame m and frequency bin k.

$$|\mathbf{H_{EQ}}(m,k)| = \frac{|\mathbf{H_{aem}}(m,k)|}{|\mathbf{H_{iem}}(m,k)|} \tag{4.2}$$

On obtaining m estimates of the equalisation function, an average EQ-function is computed; where the average is computed along the m frames. This is represented as $|\hat{\mathbf{H}}_{\mathbf{EQ}}(k)|$. The calculated average transfer function is shown in Fig.4.2(left). Significant transient activity is present on the averaged transfer function, therefore; an moving-average window is applied to smooth the transfer function magnitude response, as seen in Fig.4.2(right). It is evident that using a moving-average does average some harmonic information along with transient characteristics in the transfer function. However, since the human ear is not apt in distinguishing frequencies close together, this should not affect intelligibility significantly.

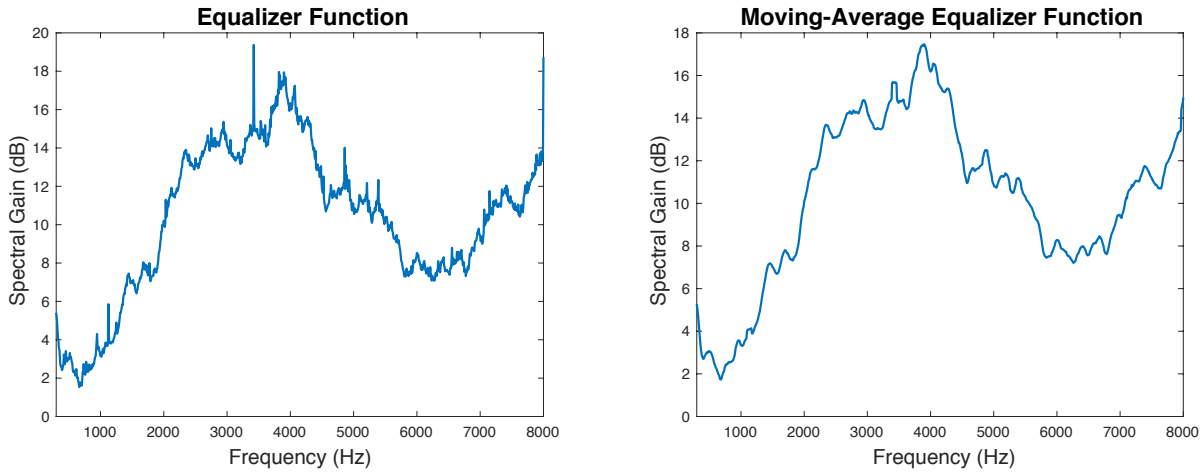


Figure 4.2: Spectral Equaliser Transfer function with (right) and without (left) moving-average smoothing window.

On obtaining the EQ-function the test set IEM speech can be transformed into AEM speech. This requires computing the STFT of test IEM speech, followed by multiplication of the transfer function with each framed segment of test IEM speech, as shown in eq.4.3 .

$$|\hat{\mathbf{H}}_{\mathbf{aem}}(m,k)| = |\hat{\mathbf{H}}_{\mathbf{EQ}}(k)||\mathbf{H_{iem}}(m,k)| \tag{4.3}$$

Experiments are conducted using the linear Equalising transfer function approach suggested above. On using this approach it was seen that the effect of averaging causes loss of information between different sounds, or phonemes in speech. This results the algorithm providing a constant gain in the 2000-4000Hz region, hence a sound dependent proposed next.

4.2 Phoneme-based Clustering Spectral Equaliser

Since the attenuation caused by the bone-conduction channel is speaker and phoneme dependent, a phoneme mapping is suggested in this section. The phoneme based mapping, is obtained by adopting a clustering-based approach. The idea is train a cluster mapping segmenting utterances into clusters of phoneme sounds and calculate a transfer function for each cluster. Then the trained mapping can be inferred to fit new points (speech utterances) and transform unseen speech utterances. A flowchart of system is shown in Fig.4.3

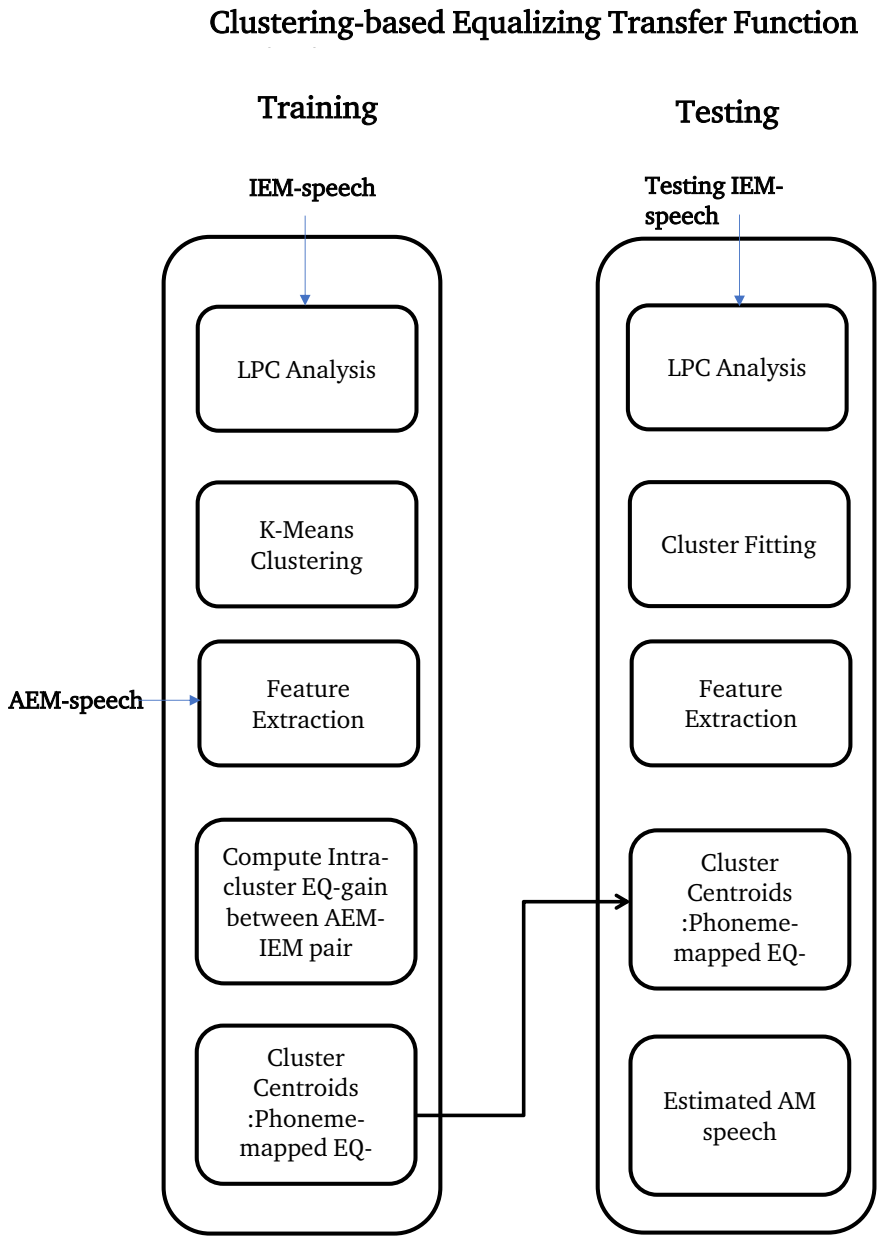


Figure 4.3: Phoneme-based clustering spectral equaliser approach flowchart

4.2.1 Linear Predictive Coding Feature Representation

As established the relationship between AEM and IEM speech varies with phonemes, therefore; there is variation in per frame relationship between AEM and IEM speech. In order to model this LPC coefficients are chosen. LPC coefficients were discussed in section 4.2.1 when used for the whitening filter in Artificial BWE approach. Here LPC coefficients are used as speech feature encoding technique to store the harmonic information of speech utterances, which can subsequently be clustered. Along with providing harmonic information on speech utterances, LPC coefficients encode speech, using an AR model, to reduce the dimensionality of problem.

Per the source-filter model, the filter part is an all-pole function $A(\omega)$. The vocal tract function filter $A(\omega, m)$, is multiplied with the excitation signal $E(\omega)$ to produce the bone-conducted speech $B(\omega, m)$, which is recorded by the ECM by the IEM. This can be represented as follows:

$$B(\omega, m) = A(\omega, m)E(\omega, m) \tag{4.4}$$

Where $A(\omega, m)$ is as shown in eq.4.5. With $a_i(m)$ representing the LPC coefficients of frame m, M being the filter order of the LPC filter, and $G(m)$ is the residual. Since all harmonic information is encoded in the LPC coefficient, only they are used for clustering.

$$A(\omega, m) = \frac{G(m)}{1 + \sum_{i=1}^{M} a_i(m) z^{-i}} \tag{4.5}$$

Although LPC coefficients, provide the benefit of encoding harmonic information and reducing the dimensionality of the clustering problem, LPC coefficients are not suitable in providing suitable clustering performance. The reason behind this is explained the next section, in context to K-means clustering method. However, here to improve the clustering performance LPC Cepstral coefficients (LPCC) are used. Cepstral Coefficients are defined as the magnitude inverse STFT of the log-power STFT coefficients of speech frame. Cepstral coefficients for AR processes are represented as follows [40]:

$$\begin{aligned} c_0 &= \log(P) \\ c_1 &= -a_0 \\ c_k &= \begin{cases} a_i + \sum_{m=1}^{k-1} (1 - \frac{m}{n}) a_m c_{k-m}, & \text{if } 1 < k \leq M \\ \sum_{m=1}^{n-1} (1 - \frac{m}{n}) a_m c_{n-m}, & \text{if } M < k \end{cases} \end{aligned} \tag{4.6}$$

Here c_k denotes the LPC cepstral coefficients, while a_i is the LPC coefficient and M being the filter order of the LPC process. As seen LPCC can exceed the size of the filter order, however; the LPCC have a faster decay, therefore, after a few coefficients LPCCs tend to zero. Next the K-means clustering is discussed and, the justification for using LPCCs instead of LPC is provided.

4.2.2 K-means Clustering

In the training phase of clustering, the LPCCs of IEM speech are computed on a per-frame basis. Frames are created using a periodic Hamming window, which is then propagated with a 75% overlap across the training AEM and IEM speech. Once the training matrix is created, LPCC analysis is conducted on the training IEM matrix on a per-frame manner. This results in a matrix of dimensions - number of frames by filter order, M; and is then provided to the k-means algorithm for clustering. The number of clusters in the K-means algorithm is a hyper-parameter, and is

required to be optimised. The output of the K-means algorithm are the LPCC coefficients cluster centres and a vector of size of the number of training examples (IEM training frames) whose entries indicate the cluster membership of each IEM training frames. On calculating the grouping of similar phonemes in the LPC sense, their corresponding cluster spectral gain function needs to be computed. For each cluster, a corresponding AEM-IEM training pair EQ-function is computed (eq.4.2), which is then averaged over the cluster to obtain EQ-function. This process is iterated over the total number of clusters.

In the testing phase, test IEM speech is enframed in a similar manner described above. LPC coefficients are then computed, and fit to the K-means clustering. Model fitting in K-means is done using a dissimilarity distance metric, such as Manhattan or Euclidean distance. The dissimilarity measure computes the distance of the point from each cluster and assigns to the nearest one. This is where the need for LPCC over LPC is justified. LPC are AR coefficients which are as follows:

$$\mathbf{a} = [a_0, a_1, \ldots, a_{M-1}] \tag{4.7}$$

In time-domain, the filter response is the weighted summation of historic values of signal. Therefore, with increasing order, higher-order values are of decreasing significance compared to a_0, a_1, a_2. A dissimilarity measure like Euclidean distance cannot distinguish between this, can therefore penalises the difference in both lower and higher equally. An alternative will be to use a weighted Euclidean measure, however; that will introduce extra parameters in the problem. Cepstral coefficients provide a solution to this problem, and hence used with standard Euclidean distance.

Once cluster membership for a frame is established the corresponding EQ-function is applied (eq.4.3). Iteration over the frames, provides a spectral estimate of AEM speech.

4.3 Implementation

In the implementation of the phoneme-based transfer function, the major system-level modules described in the section above are discussed in finer detail. This is done by discussing the software implementation of these modules. The implementation also aims to present the design and analysis in form of an algorithm, which is then used in the Results and Testing section in section 4.4.

Data Preparation : Pre-processing and Training Data

The pre-processing and the training module deals in manipulation raw speech waveform to prepare the AEM and IEM speech pair for LPC feature extraction. The primary step is the pre-processing module. Input parameters to the pre-processing module are the individual IEM and AEM speech file in sampled format, the original sampling frequency of the audio files, which is 44.1kHz, and finally the desired down-sampling frequency. Output to the pre-processing module are the silence-pruned and STSA-MMSE enhanced IEM and AEM signal, downsampled at 16kHz. The pre-processing module code is shown in Fig.7.2 in the Appendix.

Silence removal is performed using a spectral energy VAD, which prunes frames by providing a binary representation of speech presence, dependent on the spectral power threshold in each frame. It was noted that there were instances of false positives in speech detection, where silent frames are detected as speech due presence of impulsive ticks in the frame. Since they are silent

frames, no equalisation is required. This problem is addressed by clustering speech frames, the silent frames are expected to cluster together, hence avoiding the possibility of misclassifying silent frames and equalising them.

On pre-processing, AEM and corresponding IEM speech are required to be enframed into separate matrices specified frame length . The enframing function (`v_enframe`), which is adapted from the VOICEBOX [39] implementation, divides a time-domain signal into matrix of overlapping time frames. The dimension of this matrix is of number of frames by the frame size. This process is iterated over the entire training dataset. Each iteration creates two matrices of enframed AEM and corresponding IEM speech. At each iteration, the AEM and IEM enframed speech is concatenated to a training AEM and IEM matrix respectively. Therefore, on preparation of training data, speech utterances are framed and concatenated into a matrix of overlapping frames.

LPCC and K-means Implementation

Once the training AEM and IEM matrices are prepared, LPC coefficients of IEM speech frames are computed. The computed LPC coefficients of IEM speech correspond to each frame of the training IEM matrix. LPC coefficients are computed using the MATLAB implementation, which uses autocorrelation method of AR models. The lpc function requires a time-series input, which is the IEM speech frame, and the filter model order, M. The function output is a vector of size $M+1$ containing the LPC coefficients for that speech frame. Cepstral coefficients are computed using the expression in eq.4.6 [39].

The K-means algorithm as discussed in section 2.4, is implemented using MATLAB implementation of the k-means++ algorithm. The function inputs matrix, with columns indicating points (IEM-LPCC) and rows representing the variables (number of LPCC coefficients).

Average Cluster Filter ($|\mathbf{H}_{\mathrm{EQ}}|$) Computation

Next step in the implementation is computation a codebook of per-cluster equalising functions ($|\mathbf{H}_{\mathrm{EQ}}|$). These cluster are computed using the code shown in Fig.7.3.The function takes inputs from the result of the k-means clustering algorithm. These are an index vector of the size on number IEM speech frames, and matrix containing the LPCC cluster centres. The index vector contains an index, which corresponds to a cluster, indicating its membership. In regards to the output, the function returns the average spectral gain, or equalisation function ($|\mathbf{H}_{\mathrm{EQ}}|$) for each cluster.

The function first iterates over the number of clusters established from the trained k-means clustering. For each cluster, first all the cluster members are grouped in a vector using the find function. These are stored in the vector cluster_idx(refer Fig. 7.3). On establishing the cluster members, a nested for loop, which iterates over the intra-cluster members is constructed. It is inside this loop, the intra-cluster equalisation function is computed. Firstly, the corresponding AEM-IEM pair in the cluster is loaded. Here the AEM-IEM speech pair is a single framed utterance, from the training set. Spectral gain function is computed as shown in eq.4.2. On computing $|\mathbf{H}_{\mathrm{EQ}}|$ for each cluster member, there are several estimates for the same phoneme mapping. Hence, an average is derived to obtain a single $|\mathbf{H}_{\mathrm{EQ}}|$ estimate for the cluster. This process is iterated over all clusters. The result is a matrix storing codebook of phoneme transfer function of size : number of clusters by frequency bins ($\frac{N-FFT+1}{2}$).

Equalisation Filter Selection

With establishing a codebook of filters, which represent a phoneme-based mapping of IEM to AEM speech, a method for picking an optimal equalisation filter on a frame basis is required. This is performed using the code shown in Fig.7.4.

Test files are evaluated as individual utterances. Similar to producing the training dataset. Firstly, the test utterance is pre-processed, which is then followed by the enframing procedure. After enframing the test utterance is in form of a matrix which spans the dimensions of the number of frames in the utterance by the frame size. For each frame its respective IEM LPCC coefficients are computed. The LPCC coefficients are then classified using the pre-trained k-means IEM LPCC model. Classification is done using the Euclidean dissimilarity measure; where the LPCC coefficients of each test frame are assigned to the cluster with the minimum distance from the trained LPCC coefficient cluster centres. On determining the cluster membership, the corresponding phoneme-dependent equalisation function can be applied to the frame. It should be noted the equalisation function is applied in the spectral domain, hence STFT coefficients of IEM speech were also derived using the spectrogram function from MATLAB.

4.4 Chapter Summary

In this chapter, a spectral-domain technique is proposed from a baseline spectral equalisation method in literature [22]. Baseline Spectral gain method applies an average gain per frequency bin, computed through FFT magnitude ratio of the AEM and corresponding IEM speech utterance. The method is however, contains no information on the variability of the bone-conduction in modulating sounds in different manner. Therefore, phoneme-based mapping is suggested. The phoneme-based approach proposes to model phoneme variation using LPCC features. Therefore LPCC clusters are clustered using k-means clustering. Subsequently, cluster centres are computed. Subsequently, cluster specific equalisation transfer functions are computed. During testing new utterances are classified using Euclidean dissimilarity, and corresponding equalisation gain is applied, to convert IEM speech into AEM speech. In the next chapter, testing is performed on the two approaches in Chapter 3 and Chapter 4.

Chapter 5

Testing and Results

In this chapter, the testing and results are presented for the two approaches - Artificial Bandwidth Extension and Spectral-domain methods. Before delving into testing and results, objective error metrics to evaluate the performance of the two approaches is discussed as well. Testing is conducted to determine the functionality of the proposed approaches and also determine the optimal value for parameters, which led to optimal performance of methods in terms of the error metric. Along with the objective evaluation of the error metric, the results are also evaluated through visual inspection of spectrograms in identifying the key aspects of the proposed methods.

5.1 Evaluation Metric

Evaluation metrics were surveyed in the literature survey in section 2.5. These include both subjective and objective measures. The evaluation metric used for this project is a distortion measure called Log-spectral Distortion (LSD). LSD is a spectral distance metric which is denoted in eq.5.1.

$$LSD(dB) = \sqrt{\frac{1}{B}\sum_{\omega=0}^{B} 10[\log(S(\omega)) - \log(\hat{S}(\omega))]^2} \tag{5.1}$$

Here, $S(\omega)$ and $\hat{S}(\omega)$ denote the PSD of reference (clean) and enhanced speech respectively, and B representing the bandwidth of the signal. It should be noted that due to the log-transformation of the PSD, LSD is calculated in dB. Intuitively LSD can be interpreted as the mean-squared distance between the reference and enhanced signal. According to [41], two signal are indistinguishable in terms of speech quality when LSD $\leq$ 1dB. For the purposes of evaluation LSD is calculated for 25ms window with 15 ms overlap. Furthermore, LSD is calculated in frequency ranges, to infer of the improvement in various frequency bands. These frequency ranges are as follow: low-frequency (0-2kHz), mid-frequency (2-4kHz) and high-frequency ranging between 4-8kHz.

5.2 IEM-AEM speech Data

The training and test data was derived from the database established in section 2.2.3. Out of the 120 utterances, which spanned approximately 8 mins in total. The dataset is recorded from a set pf phonetically balanced sentences from the Harvard Speech database, which are spoken by a single male speaker. The utterances are recorded under quiet conditions in an anechoic chamber, at a sampling rate 44.kHz, but later downsampled to 16kHZ.

5.3 Artificial Bandwidth Extension

The essential parameters for the A-BWE method and their respective default values are as follows. Firstly, speech is segmented into uniform frame length of 400 samples (25ms). The LPC whitening filter is of 18-th order FIR filter. In regards to the filtering stage, Butterworth 3rd-order low and high-pass filters are used. Lastly, a bandpass filter is used to eliminate any out of passband information. Initially, both the low pass and high pass filter have a cut-of frequency of 1800Hz. The bandpass filter's passband is in the range of voice frequencies of 300-3800Hz.

5.3.1 Testing and Results

Firstly the major parameters influence is determined on the performance of bandwidth extension approach. These are as follows: **(1)** length of speech frame, **(2)** LPC whitening filter order.

Frame length of Speech

The frame length is parameter which determines the context which is provided to the LPC whitening filter, to conduct inverse prediction in estimating the narrowband excitation function. Initially the frame length is set to 400 samples. Fig.5.2 shows the effect of varying frame length. The plot also displays the corresponding LSD of the utterance used. The reference LSD between the AEM and raw IEM speech is 3.847. It is evident that increasing the frame size also increases the LSD of BWE-IEM speech. Although on decreasing the frame length, the LSD decreases; there was a significant fall in the perceptual quality of the speech. This can be attributed to the increased number of artefacts (impulsive vertical manifestations), which cause popping sounds. Therefore, there is a trade-off between reducing LSD and perceptual quality with choosing frame. Hence, frame length value 2000 was chosen as most artefacts are disappeared then.

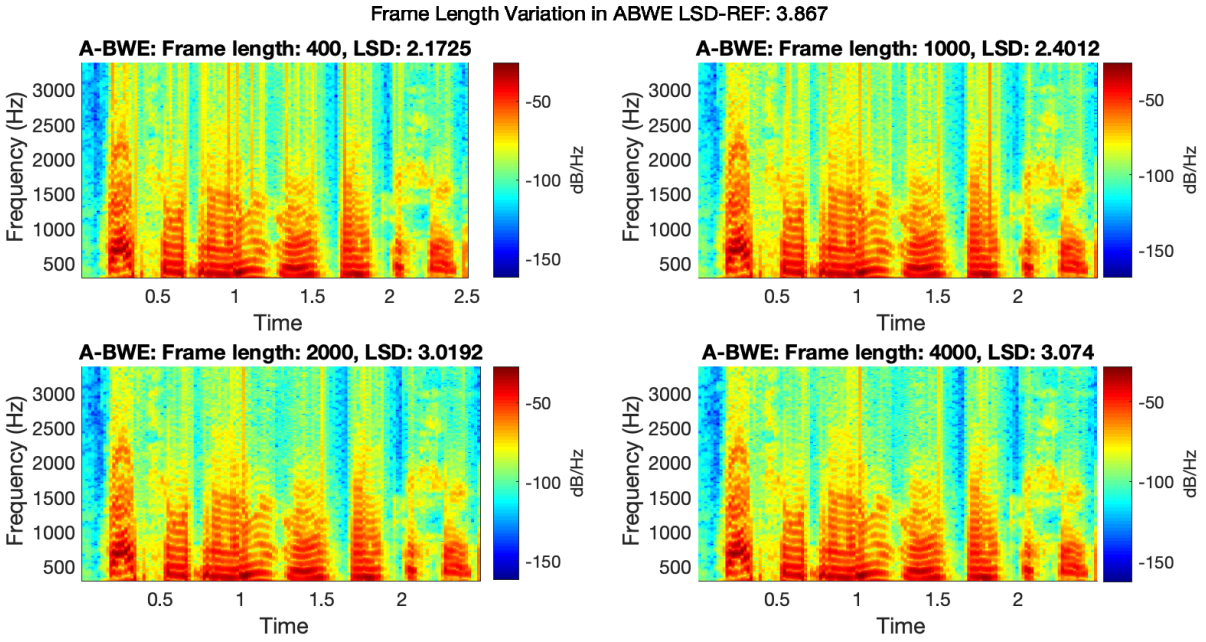


Figure 5.1: Results of applying artificial bandwidth extension on IEM speech for a test utterance

LPC whitening filter order

As discussed the LPC whitening filter extracts the narrow band excitation signal from the speech frame, which is then passed through the non-linearity function to the attain wideband excitation signal. Therefore, filter order is an essential parameters to be optimised. In this investigation, the frame length was set to 2000, as derived from the previous experiment. Subsequently, LPC filter order was varied from the range 10-30, which was chosen from typical values chosen in the literature, and LSD values for a particular utterance were calculated. These are shown in Table.5.1. It should be noted that no discernible changes were witnessed in the spectrograms, with varying the filter order. Table shows 5.1 minor variations in LSD for the given range of LPC filter order. Therefore, 18 was chosen as optimal filter order to avoid under or over modelling of the LPC power spectra.

REF LSD : 3.867	
LPC Filter Order	LSD (dB)
10	3.0275
18	3.0192
25	3.0306
30	3.0382

Table 5.1: Relationship between LSD and LPC filter order

Next using the derived values for the frame length and the LPC whitening filter order, testing is performed to assess the performance of the ABWE approach. Since bandwidth extension is a blind method, i.e, it does not require any training, hence there is no testset for the approach as well. The test utterance and the ABWE extended IEM speech, along with the reference speech (AEM) are shown in Fig.5.2.

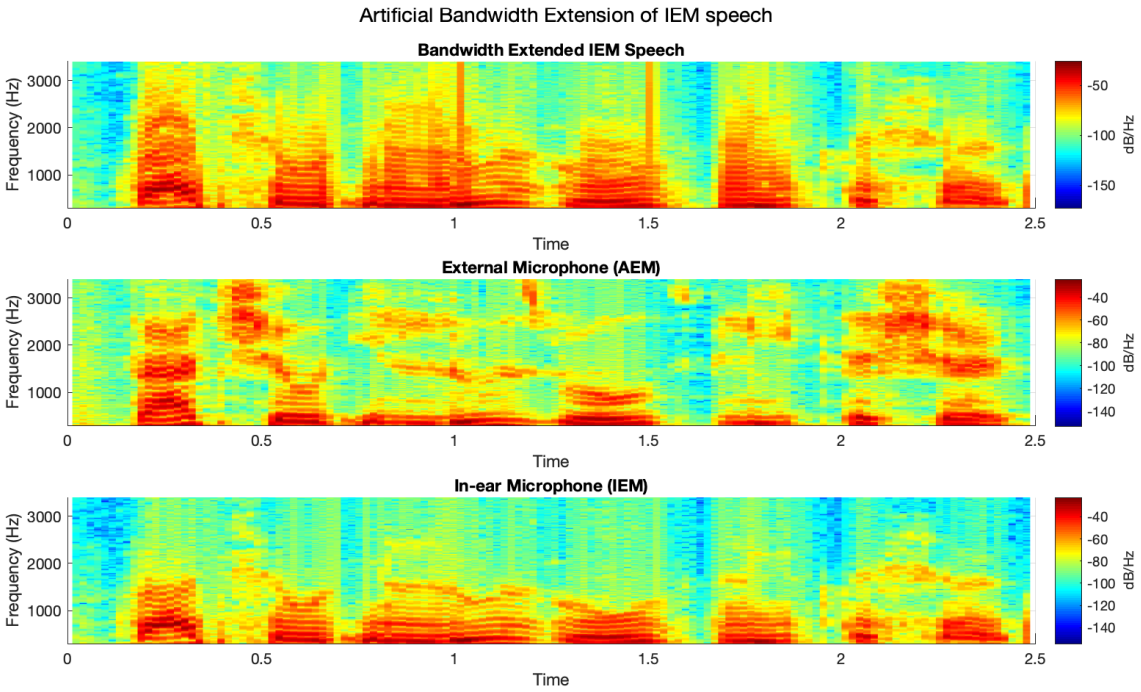


Figure 5.2: Results of applying artificial bandwidth extension on IEM speech for a test utterance

It is evident from the spectrogram in Fig.5.2 can only map limited aspects of AEM speech given the IEM speech. The first chunk on the spectrogram is mapped well using the using the BWE approach, as the odd harmonics generated share common information between the low and cut-off band. However, with features with dissimilar frequency content between bands, the BWE approach extends the content from the low-frequency bands, which consequently degrades the speech quality even further. On the notion of degradation, artefacts can also be noticed on the Bandwidth extended speech. These are mainly observed in form of impulsive popping sounds, as evident on the BWE extended speech in Fig.5.2(top). It was also observed with the non-linear characteristics function, there is some unpredictable behaviour in regards, to which parts of speech are band extended, while which ones are not. Furthermore, referring to the other test utterances, placed in appendix section in Fig.7.5- Fig.7.7, these common attributes are observable. The sub-optimal performance of the bandwidth extension method can be attributed to there not being sufficient mutual information between the lower frequency and attenuated or missing higher frequency content, which was an assumption placed on using the artificial bandwidth extension method. The rationale behind the assumption is due to the source-filter model. Since both narrowband and wideband speech are produced by the same process, they were assumed to have common information. However, using the non-linear characteristics this assumption does not hold. Hence justifying the sub-optimal performance of the bandwidth extension. From a quantitative perspective, Average LSD (ALSD) is calculated over the 10 test utterances in the testset. ALSD for raw speech (AEM vs IEM) speech was 4.74dB while the ALSD between AEM vs BWE enhanced speech is 4.04dB. In addition, per utterance results can be seen in the Appendix: Table.7.1. It should be noted the change in LSD is not perceptible in the sound quality of speech, as various artefacts are also introduced.

In summary, the testing and results are shown for the artificial bandwidth extension approach for blind estimation of AEM speech using the IEM speech. Firstly, relationship between the parameters frame length, LPC whitening filter order with LSD is explored. It is established that with increasing the frame length causes decrease in LSD, however; with decrease in frame length the causes artefacts in BWE enhanced IEM speech. In regards to the LPC filter order, no significant relationship between the filter order and LSD is found. Next results are computed using these parameters. It is seen non-linear bandwidth extension is unable to model the AEM speech especially harmonics and fricatives outside the cut-off frequency. In quantitative terms, although LSD lowers on performing BWE, but the improvement is not translated to noticeable change in audio quality. In the next section, the spectral-domain equalisation approach testing and results are discussed.

5.4 Spectral-Domain Methods

In this section, results and testing on spectral-domain approaches discussed in chapter 4 is performed. Firstly the essential parameters and their respective default values are outlined.

5.4.1 Experimental Setup

- **Training and Testing set**: The training and test data was derived from the database established in section 2.2.3. Out of the 120 utterances, which spanned approximately 8 mins in total, 110 utterances were used for the training and the remaining 10 utterances for testing. Due to the speaker-dependent nature of the transfer function, the database contained data from a single speaker. Furthermore, testing is performed in an unseen manner, such that there are no common utterances between the testing and training data.

- **FFT Parameters**: The FFT parameters include all the core parameters required in extracting the spectra of speech and calculating the spectral gain, or transfer function. The FFT spectra was computed using a periodic Hamming window of size 25ms (400 samples), and 75% overlap, which equates to the hop size being 160 samples. Lastly, the number of Fourier components (N-FFT) was set to 2048.

- **LPCC parameters**: Cepstral coefficients which are used for clustering speech utterances are calculated from the LPC coefficients of filter order 30. The reduce the dimensionality of the problem, only significant LPCC are chosen, i.e, non-zero LPCC coefficients. Therefore, the LPCC coefficients were simplified to only 18 components.

- **Smoothing Window**: As discussed in section 4.1.2, a moving average window is used to eliminate transient characteristics in the computed transfer function. The length of the smoothing window is set to 5 samples. This implies the average is computed for 5 samples ahead and behind the current frequency bin.

- **Cluster Size**: Cluster size is also an optimisable parameter, which determines the partitioning of phonemes for the K-means clustering in the phoneme-dependent transfer function approach. The cluster size is set to 10 cluster initially.

5.4.2 Testing

In this section, the proposed Phoneme-based mapping transfer function performance is compared with the linear spectral gain equaliser. In regards to the performance, the testing is done through analysing the differences in transfer functions of the two spectral-domain methods. In addition, the spectrograms and the resulting power spectra mismatch between the approaches is discussed. On the notion of spectral mismatch, the LSD, discussed in section 5.1, is also used as a performance indicator for the spectral-domain methods.

Testing is begun by analysing the primary objective of the project - the AEM-IEM speech transfer function. The transfer function is plotted for both Linear Spectral Gain and the Phoneme-mapping based approach in Fig.5.3. The spectral gain (SG) transfer function (Fig.5.3 - left) is the average transfer function calculated over the frames of the training data. While the cluster-based transfer function, is a codebook of equalising functions, where each equalising transfer function corresponds to a K-means cluster centre. In regards to this, the linear spectral gain can then be interpreted as average of theses codebook transfer functions in the clustering method.

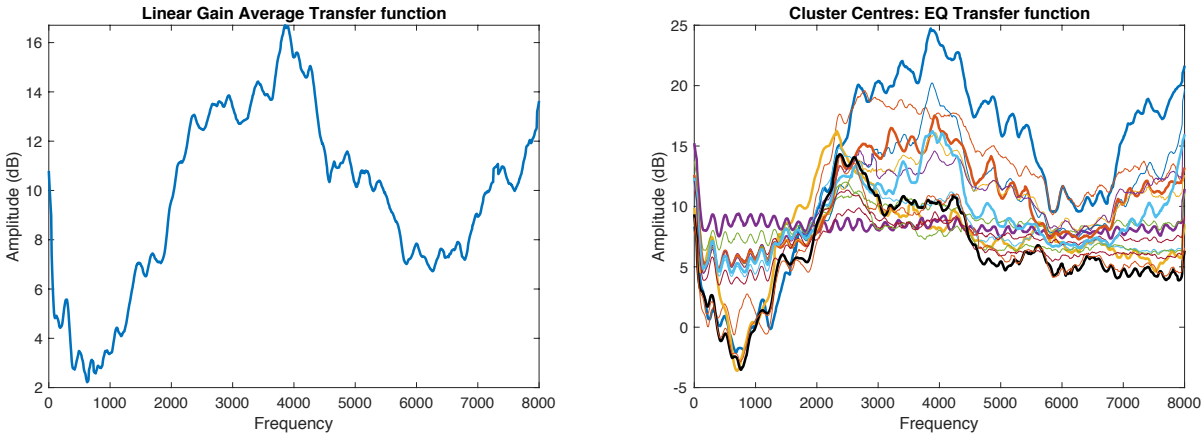


Figure 5.3: AEM-IEM transfer function for Linear Spectral Gain (left) and transfer function codebook (right) for Phoneme-based mapping approach. Wider line width indicate significant transfer functions

The general trend of the linear spectral gain transfer function follows such that the largest gain is applied in the 2-4kHz region, ranging from 13-16dB, while the gain is reduced in the high-frequency region between 4-8kHz. For the low-frequency region in 0-2kHz, the gain applied is minimal, suggesting large similarity between the low frequency bands of AEM and IEM speech. Since the gain of the transfer function varies, it suggests that the bone-conduction channel is a frequency-selective filter, in contrast to the previous claims of being a simple low-pass filter.

In regards to the phoneme-based transfer functions, it is evident the phoneme-mapping have similar characteristics of the averaged transfer function. Although the codebook transfer functions have similar shapes, the codebook transfer functions vary in gain magnitude. For instance, some transfer functions (Fig.5.3(right) - bold black and amber lines) attenuate the IEM speech characteristics in the frequency regions 500-1300Hz. In addition, the solid blue line in Fig.5.3 resembles very closely the linear spectral gain transfer function. This suggests that this transfer function characteristic has the greatest influence on the average transfer function of IEM-AEM speech. Another interesting frequency response is of the violet curve, which provides a constant gain at approximately 10dB. Such a frequency response can be attributed to the silent frames in the frequency spectrograms. During prepossessing, a spectral-energy based VAD was used, however; silent frames were still observed post pre-processed speech. Since, the in-ear recorded speech is occluded external noise is suppressed, therefore; in silent regions in IEM speech form a cluster with constant gain transfer characteristics.

Furthermore, the evidence of clusters having similar phoneme characteristics can be witnessed through the varying frequency response of codebook transfer functions. Along with the transfer function resembling the average spectral transfer function, few transfer functions have low to moderate amplification with flat gain characteristics past 4kHz. Such frequency can be attributed to vowel sounds like /a/. In contrast, transfer functions with high amplification in high-frequency bands, such as the orange plot, can be mapped to sounds such as /s/, /z/, called unvoiced fricatives. Unvoiced fricatives are seen on spectrograms as bursts of energy with static noise like characteristics, which mainly traverse the high-frequency regions. Since they will be band-limited due to the bone-conduction channel, therefore; high frequency amplification is required to map unvoiced fricatives. Next resultants spectrograms along with their PSD plots are shown from these approaches.

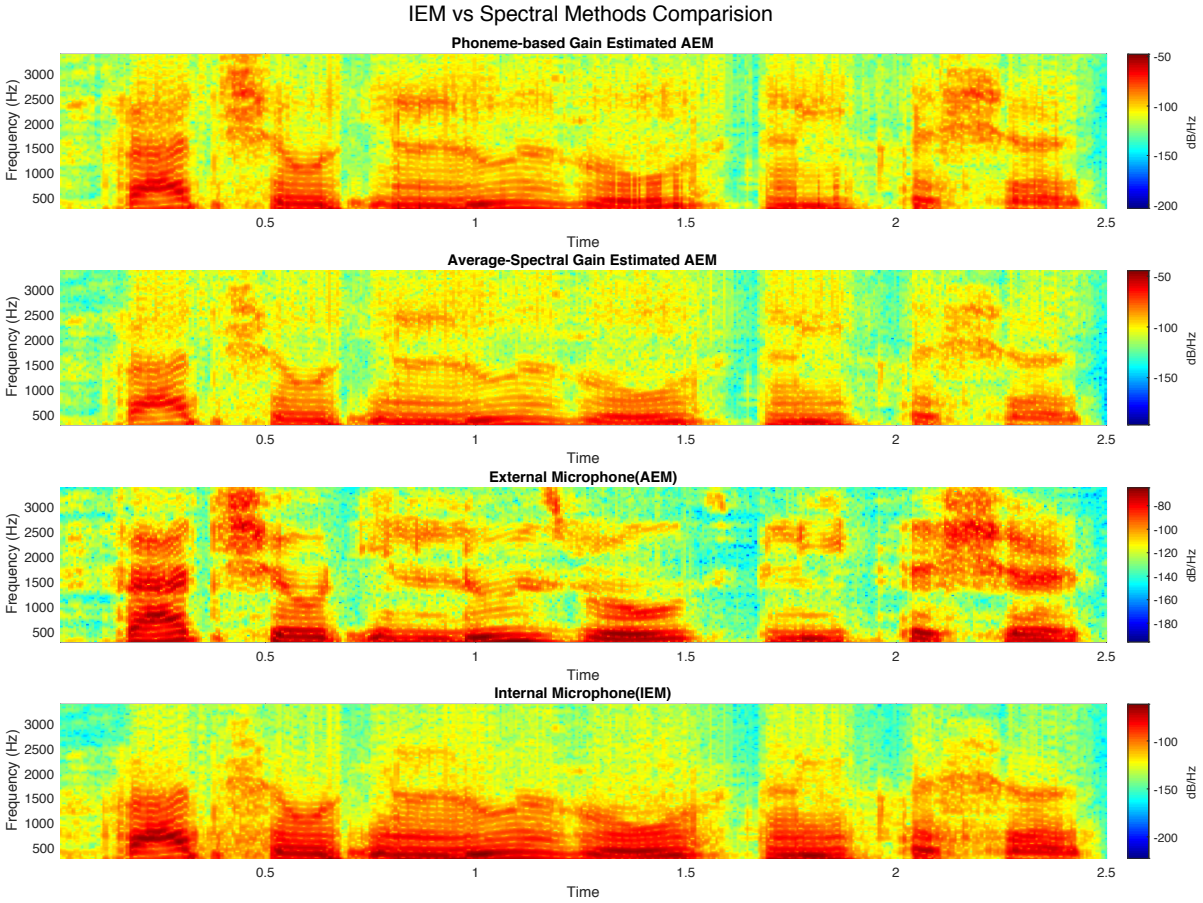


Figure 5.4: Spectrograms comparing the spectral-domain approaches. From top to bottom: Estimated AEM- Phoneme mapping, Estimated AEM- Linear Transfer Gain, Reference Speech, IEM speech

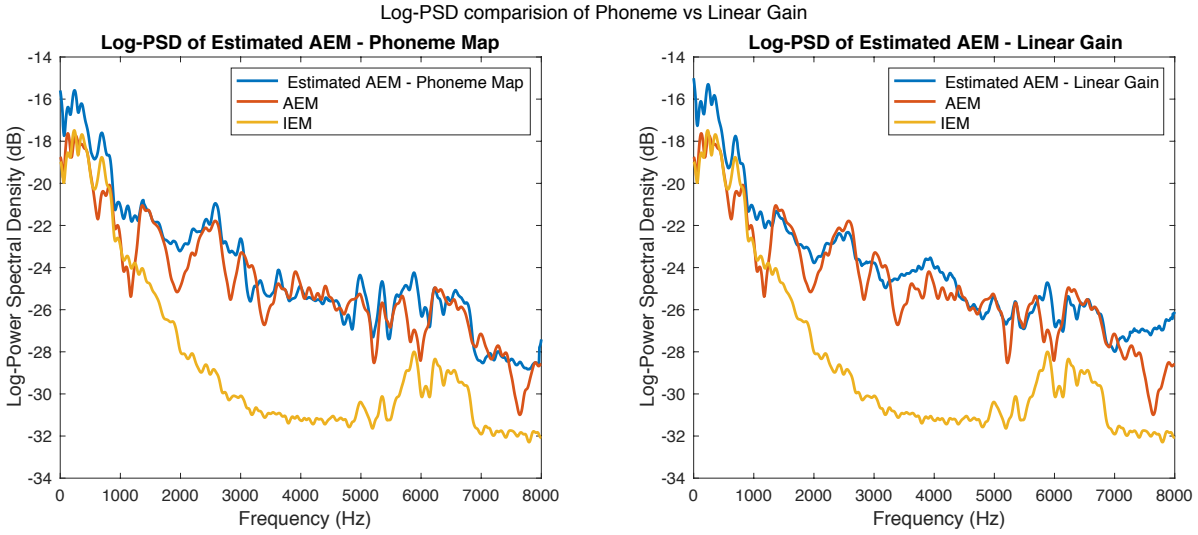


Figure 5.5: PSD plots comparing the spectral-domain approaches. Left: Estimated AEM- Phoneme mapping. Right: Estimated AEM- Linear Transfer Gain, Reference Speech, IEM speech

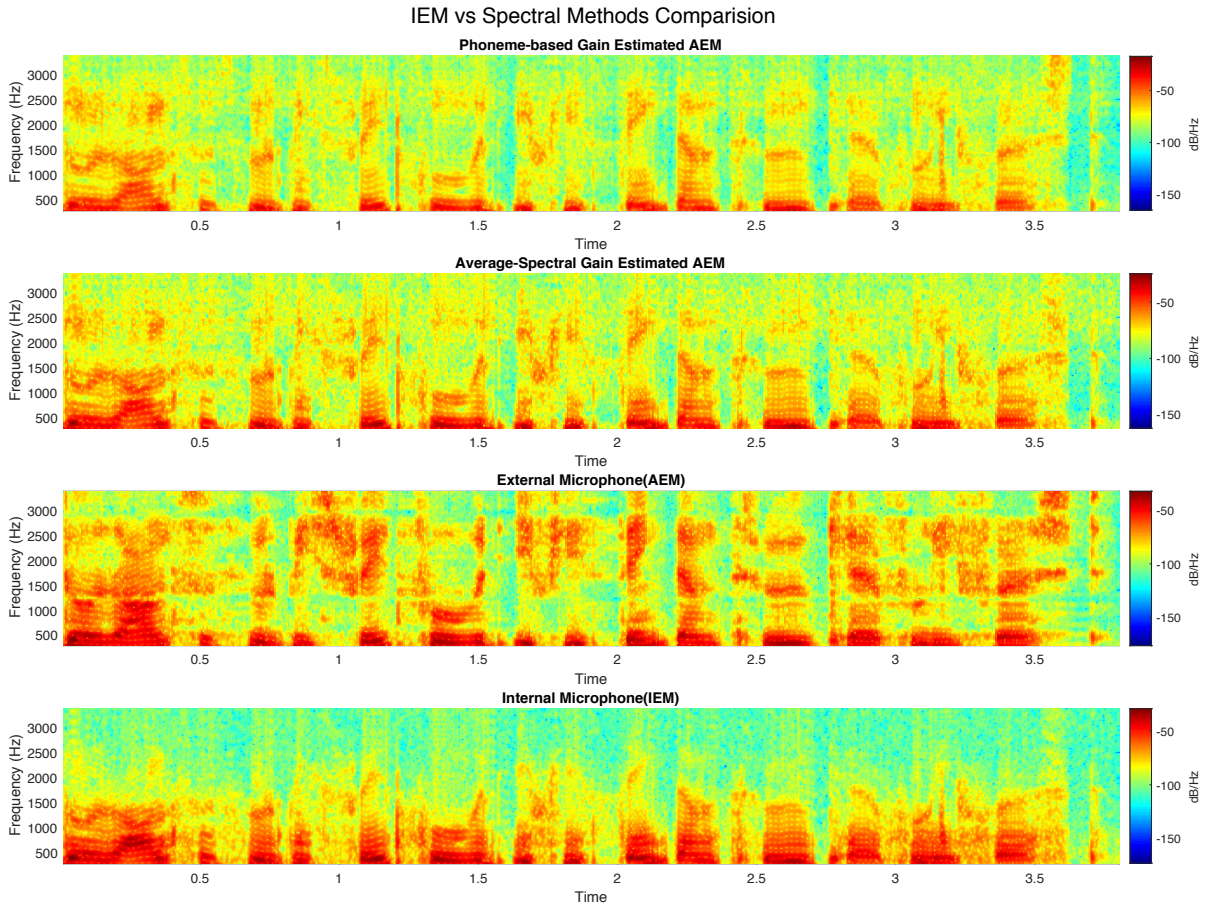


Figure 5.6: Spectrograms comparing the spectral-domain approaches. From top to bottom: Estimated AEM- Phoneme mapping, Estimated AEM- Linear Transfer Gain, Reference Speech, IEM speech

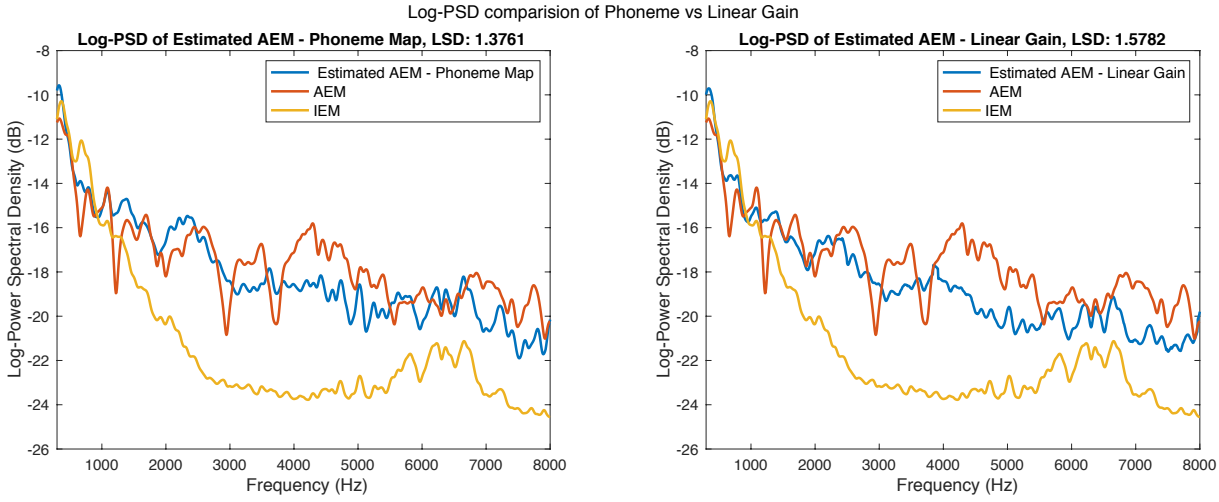


Figure 5.7: PSD plots comparing the spectral-domain approaches. Left: Estimated AEM- Phoneme mapping. Right: Estimated AEM- Linear Transfer Gain, Reference Speech, IEM speech

Fig.5.4 and Fig.5.6 shows the spectrograms comparing the spectral-domain approaches of Linear Gain and Phoneme-dependent Transfer function. Along with these the Reference (AEM) and IEM spectrograms are also displayed for comparison. The spectrogram is constructed using the experimental setup outlined in section 5.4.1. It should be noted that the two examples provided are picked from the testset, and more examples are referred in the Appendix: Fig.7.8-Fig.7.13.

Firstly, it can be seen in Fig.5.4 that the spectral gain methods are able to estimate the damped harmonics between 2-2.5kHz in the first chunk of the spectrograms, before the 0.5s mark. The estimation of the particular chunk is more accurate using the Phoneme-based mapping, than compared to the Linear gain transfer function. However, it should be noted both spectral approaches struggle to model any formants beyond the cut-off frequency of 1.5kHz. In addition, the phoneme-based approach also performs well in modelling fricatives. Fricative begin from approximately 1.5kHz on the spectrogram. On Fig.5.4 these fricatives are seen at time instances of 0.7s and 2.4s. These fricatives were found to be /s/ sounds. Therefore, it is as predicted from the transfer function response, that the phoneme-method is better able to model these fricatives. This is because the codebook approach can assign the specific transfer function to the /s/ sounds. While the SG transfer function dampens the high frequency content in fricatives, due to averaging effect from other phoneme sounds with moderate amplification in the high-frequency region. Also evident in Fig.7.14 which shows the fricative comparison approaches.

Fig.5.6 shows another example spectrogram obtained using the proposed spectral domain methods. There are large similarities seen between the SG transfer function and phoneme-based methods in the voiced frequency regions, ranging 300-3800Hz, which as seen in Fig.5.4 as well. The transformed speech closely resembles the external microphone spectrogram in the low-frequency band region of 0-1500 Hz, i.e, before cut-off. Both methods are able to filter out the dense structure of the band limited IEM speech. However, the due to the lack of information in frequency regions past cut-off frequency region, equalisation methods struggle to reproduce harmonic information, with formants outside the cut-off frequency.

Furthermore, inspecting the spectrograms in Fig.7.8 -7.12, it is evident the phoneme-based clustering approach is successfully clustering silent frames in speech utterances, and not applying gain in those sections. On the other hand, SG transfer function applies a constant gain, irrespective of speech presence, in the 2-4kHz region. This further improves the LSD score for the phoneme-based transfer function.

A part of degradation in the IEM captured speech quality is through artefacts caused due to the IEM capturing bodily functions. For instance, the breathing artefacts are visible in the IEM spectrograms. Breathing patterns recorded using and IEM are shown in Fig.2.9 in section 2.3.3. Breathing artefacts are manifested on the spectrograms in form of low-frequency fricative patterns, and can be seen on IEM spectrogram on Fig.5.6 in the low frequency regions at time instances prior to speech content near 3.5s. Observing the same section in the SF-enhanced and the phoneme-based spectrograms in Fig.5.6, it is evident that these artefacts have been attenuated. Phoneme-based transfer function provides better attenuation performance, as the frequency structure of breathing signal is clustered, rather than being applied a linear gain.

Next the discussion is done on the PSD performance of the two spectral-domain approaches. Fig.5.5 and Fig.5.7 show the PSD plots for both spectral-domain approaches. More specifically, the log-PSD is plotted to convert to dB scale. From the 2 sets of PSD plot it is evident that the nature of IEM bandwidth limitation is similar across utterances, with degradation being the

worst in the 2-4kHz region. In addition, there a significant similarities in the SG and Phoneme-based estimated AEM PSD. Both approaches perform very similarly in the 0-2kHz, frequency region. Differences are observed in the 2-4kHz section. This is seen as some of the harmonic structure in speech between 2-4kHz is lost due to the averaging effect, whereas this is reduced using the phoneme-based transfer function approach. This is because individual characteristics of phonemes are used in calculating the per-frame equalisation gain.

LSD Error Evaluation

In this section, the LSD error metric described in section 5.1 is used to quantify the performance of the transfer function between the SG and phoneme-based methods. LSD can be seen as the Euclidean distance between two PSD spectra. Fig.5.8 shows the spectral mismatch between AEM vs IEM, and AEM vs Proposed Method (bottom). The LSD will be the mean of the spectral mismatch between for this particular test utterance.

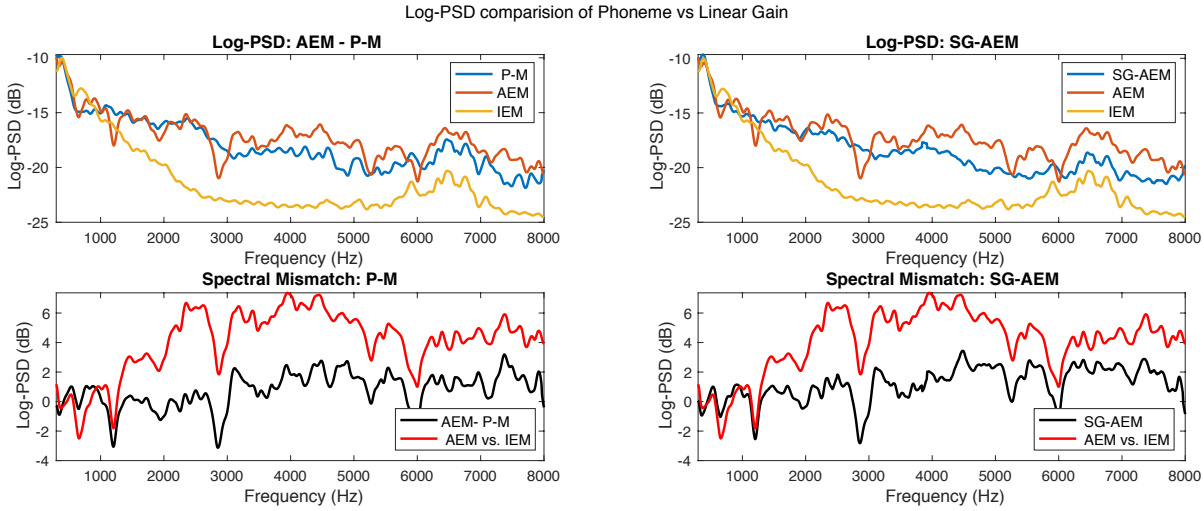


Figure 5.8: Spectra of estimated AEM speech using Phoneme-based and SG transfer function (top) of test utterance. Spectral mismatch between AEM vs IEM, and AEM vs Proposed Method (bottom)

In this case, the LSD is measured between the three scenarios: AEM vs IEM, AEM vs SG-AEM (spectral gain estimated AEM) and the AEM vs P-M (phoneme mapping method), for all test utterances. In addition, instead of computing the average LSD (ALSD), distortion is measured in frequency segments, and then ALSD is computed for each frequency segment. These frequency segments are as follows: low-frequency (0-2kHz), mid-frequency (2-4kHz) and high-frequency ranging between 4-8kHz. LSD results of single test utterances using Spectral Gain and Phoneme-based method are shown in Table.7.2 and Table.7.3 respectively. Using the LSD for each test utterance, ALSD is calculated for frequency segmented speech. These are shown in Table.5.2.

Firstly, it should be noted at the 4-8kHz range is omitted from the calculation of ALSD in Table.5.2. This decision was made as there is not much significant speech segments available in the range. Speech in that frequency range are just fricatives, which were discussed in the testing section. Since the frequency region is empty, its LSD will be similar to that of the AEM speech, which can bias the ALSD metric. The results can however, be referred to in Table.7.2 and Table.7.3 in the Appendix section. Now we can advance to the results shown in Table.5.2. In the 0-2kHz section it is seen that the raw speech (IEM) performs better than the spectral-domain methods. In addition, the spectral domain methods in that frequency regions are close to the

Method	ALSD (dB)		
	0-2kHz	2-4kHz	0-4kHz
AEM vs IEM	1.863	6.040	4.020
AEM vs AEM-SG	1.957	1.219	1.588
AEM vs AEM-P-M	1.879	1.118	1.499

Table 5.2: Average Log-spectral Distortion comparison of the proposed spectral-domain methods. Where SG represents the Linear spectral gain method, and P-M being the phoneme mapping approach.

raw ALSD value as well, where the AEM P-M approach outperforms the SG method. The larger spectral distortion in the Spectral Gain approach can be attributed to the averaging nature of the approach. In the lower bands, the AEM-SG spectra has varying values which are inherently dependent on the nature of phonemes. Since such phoneme specific information is unavailable in the SG approach, it cannot model such variations. In case of the AEM-P-M, the results are still inferior to the raw IEM speech, however; it outperforms the spectral method, as some degree of phoneme identification is applied through LPCC clustering approach. The largest gain in ALSD are seen in the 2-4kHz frequency segment. However comparing this enhancement to spectrogram, it could also be due to the non-harmonic artefacts produced when the equalising gain is applied.

In summary, the testing section featured the following significant findings. Firstly the transfer functions between the SG and phoneme-based approach are compared. It is seen that the SG transfer function can be interpreted as an average a codebook of transfer functions which are derived using k-means clustering in the phoneme-mapping transfer function. These clustered transfer functions were assigned to specific phoneme depending on their frequency amplification characteristics. During testing, it is found that the phoneme-mapping approach can alter the harmonic structure of the IEM speech which is congested in the low frequency bands (0-2kHz) and map it to AEM speech. Performance in this energy band is slightly improved to the Spectral Gain approach, but inferior to the raw speech input. In the higher frequency regions the phoneme-based approach is also able to model the high-frequency fricatives due to sound specific mapping, which the spectral gain function is unable to model as well. Overall the performance is of the phoneme-based approach is better than the baseline spectral gain method, however, the difference improvement is not excessively large. The reasons for this will be discussed in section 6. In the next section, experiments are conducted on the phoneme-based methods to derive optimal parameters.

5.4.3 Results and Experiments

After discussion of the testing of spectral domain system, several experiments are conducted in identifying the optimal parameter options for the proposed phoneme-based transfer function. In this section, the following relationship between the parameters and the proposed system will be investigated:

- Investigation of Optimal number of cluster, K, for LPCC coefficients.
- Effect of smoothing window and FFT window size.
- Number of LPC Cepstral coefficients to be used.

Number of Clusters

The number of clusters, K are a prior parameter required cluster phoneme features, which are in form of the LPCC coefficients. Initially the number of clusters are chosen to 16, as referred from the literature, [21]. However, since K-means algorithm is not a globally optimal clustering method, and is also sensitive initial conditions; deriving the optimal number of clusters is of importance. Since the k-mean approach finds clusters by reducing the total variance, the optimal cluster number can be found by choosing the one with the minimum distortion measure. The distortion measure is the mean of the within-cluster variance. Where the within-cluster variance is the summation point-wise squared distance from the cluster centroid. The distortion measure (D) is written as follows:

$$D = \frac{1}{N_c} \sum_{j=S_1}^{S_n} \sum_{i=1}^{N_p} (x(j,i) - \bar{x}(j))^2 \tag{5.2}$$

In eq.5.2 S_j denotes the notation for a cluster, while $\bar{x}(j)$ is the respective centroid of cluster S_j. N_p represents the number of intra-cluster points, while N_c denoting the number of clusters.

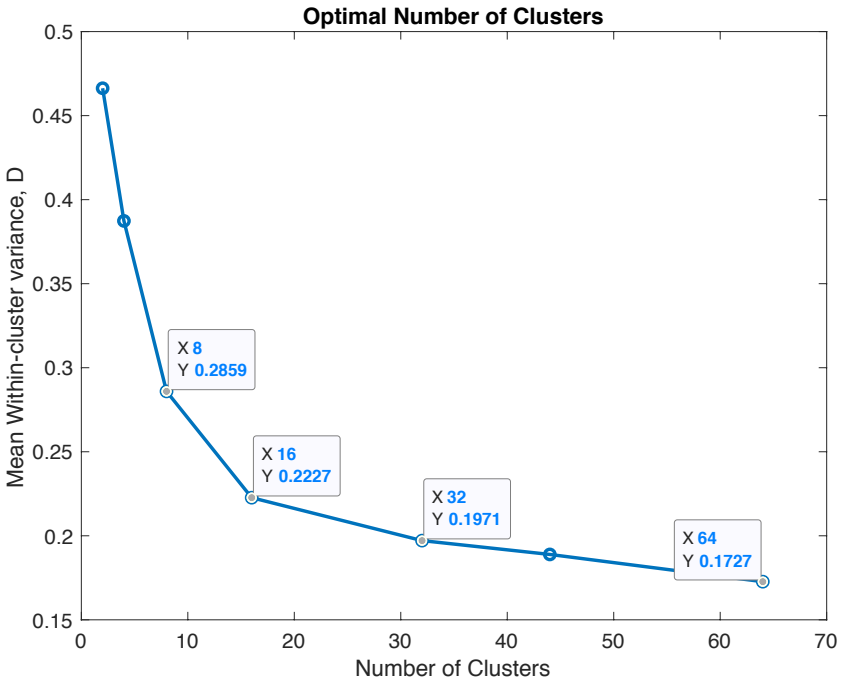


Figure 5.9: Mean Within-cluster variance for varying cluster sizes

The method used to evaluate the number of clusters is called the elbow-method. As seen in Fig.5.11 it can be seen that there is a trade-off in cluster accuracy and computational complexity. The mean within-cluster variance reduces with increasing number of clusters. According to the elbow plot, it is established 16 is indeed an optimal choice given the trade-off. It should also be noted that there is a small decrease (37%) in the cluster variance from 2 to 16 clusters. This could be indicative of the fact that clusters a located also together. In addition, clusters being in close proximity can be potential cause for similar performance of the SG and the phoneme-based method in lower-frequency regions.

Effect of FFT Parameters: Smoothing window and FFT window length

Smoothing window and the FFT window length are shown to be influencing parameters in the performance of equalisation approaches in literature [22]. In this investigation, firstly the smoothing window is kept constant at 10 samples, while the FFT window length is varied. On the flip side, when investigating the effect of smoothing window the FFT window length is set constant to 400 samples (25ms). It should also be noted that the overlap factor is kept constant at 75% overlap. The remaining parameters are as mentioned in section 5.4.1. Furthermore, this experiment is conducted over a single unseen utterance from the testset. The results are as follows:

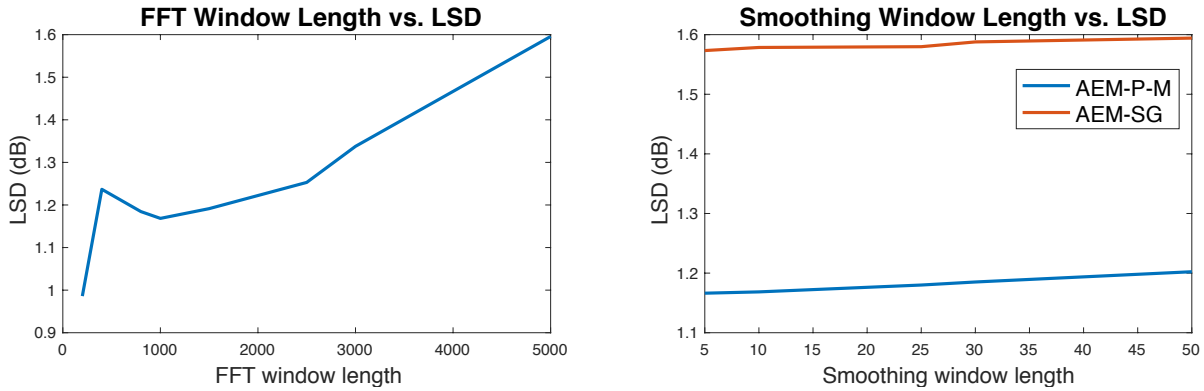


Figure 5.10: Effect of FFT window and smoothing window size on the LSD on test utterance

Firstly, it is evident that there is no significant variation in the LSD with respect to changes in the length of the smoothing window, and this is the case with both the linear spectral gain and the phoneme-mapping methods. This is not per expectation, as an optimal point was expected. This is because of the balance between the transient behaviour of the window and the averaging action of the smoothing window reducing the harmonic information in the estimated AEM speech. In regards to the FFT window length, the optimal point chosen is 1000. Although the minima is at 200, FFT window length chosen such that there is balance between the variance-resolution trade-off in windowing. Adopting a short window size, and lead to observable loss of resolution. While larger window sizes, will not be able to cluster phonemes accurately, as frames will contain more than single sounds. Furthermore, using larger window size increases the variance in the PSD estimate as well, hence affecting the LSD.

Number of LPCC coefficients

Next parameters which can potentially influence model performance is the number of LPCC coefficients. For the phoneme-mapping based transfer function, the LPCC coefficients are the feature vector over which clustering is conducted. Hence, the length of the feature vector determines the dimensionality of the problem, and is directly related to the computational complexity of the proposed approach. In this investigation, LPCC coefficients are varied, and the LSD score, dB, is obtained for all utterances in the unseen testset. The rest of the default parameters are same as mentioned in section 5.4.1. It is seen that 8 LPCCs are the optimal size for the clustering feature vector. This is the case for all but two test utterances. The general trend reflects decline in LSD with increasing number of LPCCs. However, larger LPCC sizes are not suitable due to two reasons: (1) increasing dimensionality of LPCC features directly increases computational complexity, and furthermore (2) featuring LPCC coefficients in higher dimensional space with

increasing the LPCCs coefficients the grouping in the LPCC vectors can become sparse, according to the curse of dimensionality theorem. This will adversely affect the performance of the k-means clustering, and finally affecting the LSD.

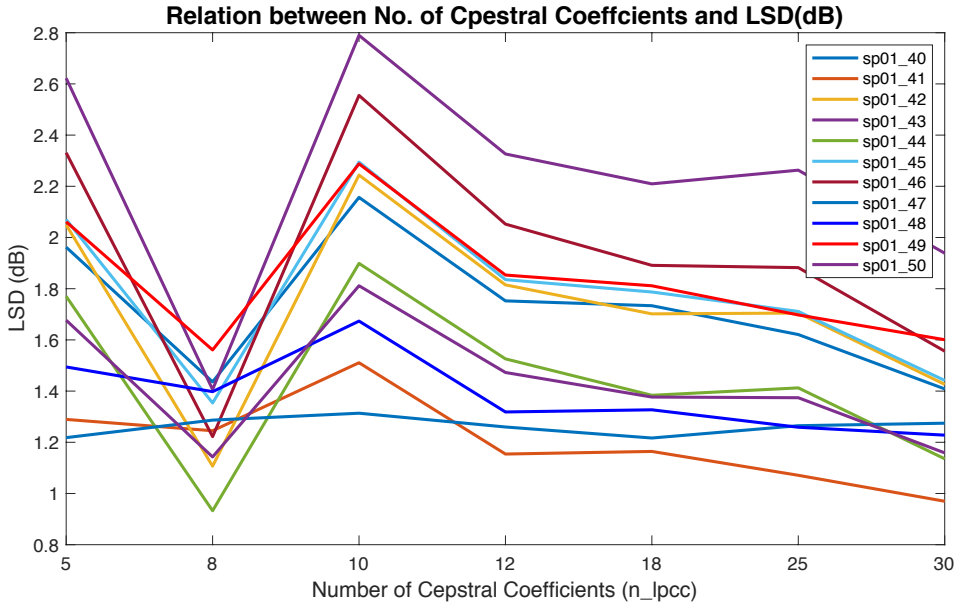


Figure 5.11: Per-utterance LSD for varying the number of LPCC coefficients.

5.5 Chapter Summary

In this chapter, testing and results are discussed on the approaches to establish a transfer function between AEM and IEM speech, which are discussed in the Chapter 3 & 4. Firstly, discussion is conducted on the Artificial Bandwidth extension approach. It is established that although the ABWE method achieve improvement in LSD score(4.74 to 4.04 dB), there is no perceptual increase in the audio quality. This is attributed to the breaking of assumption of there being significant mutual information in the low frequency and the missing high frequency sections.

Next the spectral domain approaches are explored. In the testing section, it is seen that the phoneme-mapping based transfer function can map low frequency content from IEM to AEM speech. This performance is superior to that of the baseline spectral-gain method, however; still outperformed by raw speech. In the higher frequency bands, the phoneme-mapping approach outperforms both the raw and the spectral-gain method. Due to phoneme-mapping relying on identifying the individual sounds, the approach is able to map high frequency components, in contrast to the spectral-gain method, this is shown in Fig.7.14. Lastly, in order to determine the optimal parameter performance for the phoneme-mapping based transfer function, the following parameters investigated. These the number of clusters in K-means algorithm, length of FFT and smoothing window and the size of LPCC coefficients for clustering. Results show that optimal clusters are 16, as seen in Fig.5.2. In regards to the FFT parameters, the LSD score is found to be independent of the smoothing window, however the optimal value for FFT window is 1000. The optimal number of LPCC coefficients are found to be 8, this is done in consideration to the cost increasing model complexity, and clustering performance. The next chapters the testing and result outcomes are critiqued on their advantages and shortcomings.

Chapter 6

Evaluation and Concluding Remarks

In the Evaluation and Concluding Remarks chapter, project objectives are revisited and critiqued on basis of the project work done to meet the requirements established during the project. The approaches utilised to solve the object objectives are briefly outlined. Furthermore, their merits and drawbacks are discussed in context to the project. Lastly in the concluding remarks section, the achievements of the project are placed in line with the project objectives. Furthermore, with objectives aligned with project work, suggestions and recommendations on potential future work to advance the project are discussed.

6.1 Project Evaluation

The project aimed solve the problem of limited bandwidth of the in-ear microphone signal compared to recordings from an external microphone, under quiet conditions. This objective is planned to be achieved firstly establishing a dataset of simultaneously recorded in-ear and external speech. Once the dataset is established, the primary objective was to create an inverse transfer function to map the frequency characteristics of IEM to AEMs. The transfer function is required mainly to recover the missing or the attenuated frequency content due to the bone-conduction channel.

The objective of creating an inverse transfer function is approached from two perspectives, one being a time-domain method of bandwidth extension, as outlined in [30], and next being the from spectral domain. In the time domain, it is assumed that speech in lower frequency region share mutual information with the missing or the attenuated frequency components. While the spectral-domain approach creates a frequency-dependent gain equalisation function between the IEM and AEM speech, and therefore, can be bandwidth extended to recover attenuated frequency content. The equalisation method proposed in [22] is used as the baseline result. [22] proposes computing the frequency specific equalisation gain using the magnitude ratio of the FFT of the AEM-IEM speech pair. It is aimed to extend the approach by establishing a codebook transfer functions, which corresponds to specific phonemes. Where the codebook is compiled using k-means clustering method. These methods were applied to the established dataset, and were evaluated for objective quality using the log-spectral distortion measure, which is measured in dB.

The objective of mapping IEM speech to that from the external microphone was achieved successfully using the phoneme-based transfer function approach. The proposed method outperforms the baseline spectral gain method, in regards to objective measures in form of the LSD distance measure. This is highlighted in chapter 5.4.2. The phoneme-base method outperforms the baseline in all frequency regions - 0-2kHz, 2-4kHz and 0-4kHz. In terms of perceptible evaluation, the spectrograms of the estimated AEM and IEM speech are compared. This is discussed in section 5.4.2. Further results on comparison on provided in Appendix Fig.7.8- 7.13. Although improvement is noted compared to spectral method, in the 0-2kHz frequency range even the phoneme based method underperforms that the raw IEM speech, while outperforming the raw IEM speech in range 2-4kHz, and overall in the 0-4kHz range.

The BWE algorithm was successfully implemented, but the application of this method did not show significant improvements in LSD score. In comparison to spectral domain methods, the time-domain approach BWE is shown to underperform in LSD score sense. This is shown in chapter 5.3.1 and Table.7.1 in appendix. With the ALSD score being 4.74dB for raw comparison, and 4.04dB after BWE enhancement. Furthermore, it was observed the BWE approach further degraded IEM audio quality through adding artefacts in forms of popping sounds, as evident in Fig.5.2 in Chapter 5.3.1. The sub-optimal performance of the BWE method is attributed to the violation of the assumption that there is sufficient mutual information between 0-2kHz and 2-8kHz frequency bands to apply bandwidth extension. However bandwidth extension method does have its advantages, as it is a blind method does not require any training, unlike the proposed spectral-methods. In addition, since BWE operates on estimating the narrow excitation signal and extending it using non-linearity, there is no dependence on speaker identity for BWE, which again is not the case for spectral-domain method.

Next the merits and drawbacks of the proposed phoneme-based method is compared to the linear spectral gain approach. The primary advantage of the proposed phoneme-based approach is improving the on the averaging effect of the baseline method. The baseline approach computes FFT magnitude ratio of speech frames, with having multiple transform estimates, these are averaged to obtain the baseline transfer function. Such averaging causes loss in the harmonic information of different sounds. The primary advantage of the phoneme-based model is its ability model the variation in the transfer function characteristics dependent on phoneme characteristics in speech, which is achieved using K-means clustering. Secondly using LPCC coefficients, in clustering information provides improved clustering performance, compared to STFT coefficients as, LPCC coefficients encode harmonic information in speech utterances.

Potential disadvantages of the phoneme-based method is the increased training time and computational complexity as it requires training k-means clustering, in addition to the transfer function training. In addition, issues with k-means clustering were observed as well. With K-means being a partitioning algorithm, which is not globally optimal. Therefore, the resultant cluster map is not optimal as well. Furthermore, the resultant k-means clusters were of relatively equal sizes. This implies there must be misclassification of phoneme sounds, hence influencing the phoneme-specific transfer functions. In the following section, the project achievements are concluded and future work recommendation are considered.

6.2 Conclusion and Future Work

In conclusion, the project solves the problem of limited bandwidth of the IEM, and aims to map it to external microphone speech. Through experiments it is found that the cut-off frequency in the bone-transmission channel is approximately 2kHz. Furthermore, it was discovered that attenuation in the frequencies past cut-off, varies as well (refer to Fig.5.3-left in section 5.4.2). Approaches explored are through BWE, in time domain, and spectral domain approaches of equalisation gain, and proposed extension of the equalisation gain method.

The phoneme-mapping is successfully implemented to establish an inverse transfer function, which achieves good performance in objective sense using the ALSD score measure. The approach also outperforms the baseline spectral-gain method, as shown in Chapter 5.4.2. Specifically, in the 0-4kHz frequency range, the raw speech (AEM vs. IEM) LSD decreases from 4.020 to

1.499dB, which reflects a 38% reduction in the ALSD. This is in contrast to the baseline method, which provides ALSD of 1.588. Furthermore, the phoneme-based method is able to regenerate most speech patterns in the 2-4kHz frequency, which were previously attenuated or missing due to transmission through the bone-conduction channel. It is also noted that the phoneme-based method generated fricatives as well, which were previously attenuated as shown in Fig.7.14. BWE approach was implemented, however; it did not show significant improvements, as sufficient not being present information between the low-frequency and attenuated bands.

The most significant design choices made were in the implementation of the phoneme-based approach. The method extends the average gain equalisation method outlined in [22]. It was shown that the characteristics of the bone-conduction channel are dependent on variation in sounds, as phoneme are produced due to varied movement of mouth, jaws and nasal response. In order to classify this variation, it was chosen to utilise k-means clustering, due to its fast convergence and computational load capabilities. In addition, it was decided to use LPCC coefficients as features for classifying phonemes. This decision is made to assign priority to significant LPC coefficients, while assigning them to cluster using Euclidean dissimilarity, without introducing a weighted Euclidean distance measure, as it will further contribute the complexity of the proposed method, refer to section 4.2.2. These design choices revealed the transfer characteristics of various phonemes through their respective transfer functions. These are shown in Fig.5.3 in Chapter 5.4.2. However, as explained in section 6.1, k-means clustering is not optimal; a potential extension will be to utilise globally optimal clustering methods, as distribution or density based clustering, for instance, using Gaussian Mixture Models (GMM). Since distribution based clustering method perform soft-assignment to assign members to cluster, they do not necessary bound to have equal cluster sizes. Furthermore, using a GMM any non-spherical geometries in clustering feature space can also be modelled accurately. It should be note that this come at cost of increasing model parameters and complexity.

Experiments are conducted on the implemented phoneme-based mapping approach, to determine optimal parameter values. The experiment include finding the optimal K-means cluster, determine optimal FFT window and smoothing window length, and lastly optimal number of LPCC coefficients. In regards to the cluster, $K = 16$ is the optimal number of cluster, as shown in Fig.5.11 in section 5.4.3. It was determined that the ALSD is independent of the smoothing window length, while there is a linear relationship between FFT window length and LSD measure. Optimal FFT window is chosen to be 1000, considering the variance-resolution trade-off in windowing. Lastly, the LPCC were chosen to be 8 coefficients, with respect to computational complexity and avoiding sparse data representation associated with higher feature dimensions.

In regards to the future work, the spectral-domain equalisation approach can be adopted into a time-domain filtering method to reduce the computation complexity of applying the phoneme-specific transfer function. Furthermore, using time-domain filtering also eliminate reconstruction issues. In addition, as discussed distribution or density based clustering can improve clustering performance by more accurately identifying phoneme clusters, and mapping noin-spherical disributions. This can be done using GMM method. Furthermore, an interesting application of the proposed approach can be the field of speech denoising. Since the IEM is immune degradation from environmental noise, the proposed method can be used convert IEM into estimate of AEM signal. Using the estimated AEM signal from the inner-ear, the noisy signal can be denoised by extracting the mutual information between the AEM signals.

Chapter 7

Appendix

7.1 Artificial Bandwidth Extension Algorithm - Software Implementation

```
function [x_bwe,fs] = BWE(x,fs, frm_length, lpc_order, cut_off)
% Bandwidth extension function
% Input: x_iem: In-ear Speech, fs: Sampling Frequency
%      : lpc_order: LPC filter Order, cut_off: filter cut-off frequency
% Output: x_bwe: Bandwidth Extended Signal, fs: Sampling frequency

frm_idx = 1:frm_length:length(x);
x_bwe = [];

for i=1:length(frm_idx)-1
    % Upsampled x2 speech
    up_x_iem = upsample(x(frm_idx(i):frm_idx(i)+frm_length-1), 2);
    % whitening LPC filter
    lpc_coef = lpc(up_x_iem,lpc_order);
    % apply whitening filter to speech frame
    est_x_aem = filter([0 -lpc_coef(2:end)],1,up_x_iem);

    % x^3 excitation signal
    ex = est_x_aem.^3;
    % band-extended excitation signal
    sum = ex + est_x_aem;

    % filtering stage
    hpf = highpass(sum,cut_off,fs*2);
    lpf = lowpass(up_x_iem,cut_off,fs*2);
    sum2 = hpf+lpf;

    % Band-pass filtering [160, 8000]Hz
    bp = bandpass(sum2,[160,8000],fs*2);
    % downsample signal to original fs
    bwe = downsample(bp,2);

    % band-extended speech
    x_bwe = cat(1,x_bwe,bwe);
end
end
```

Figure 7.1: Software Implementation of the Artificial Bandwidth Extension Algorithm

7.2 Pre-processing module - Software Implementation

```
function [x_aem,x_iem, Fs_new] = preproc(x_aem,x_iem, fs, downsample_fs)
% Preprocessing module function
% Input: x_aem : REF speech, x_iem : In-ear Speech
%      : fs : original Samp. Frequency, downsample_fs: required downsample fs.
% Output : x_aem, x_iem : downsampled, preprocessed speech
%        : Fs_new: downsample fs

Fs_new = downsample_fs;

% resampling speech to downsample_fs
[Numer, Denom] = rat(Fs_new/fs);
x_aem = resample(x_aem, Numer, Denom);
x_iem = resample(x_iem, Numer, Denom);

% STSA-MMSE algorithm
x_aem = v_ssubmmse(x_aem, Fs_new);
% x_iem = v_ssubmmse(x_iem, Fs_new);

% Voice Activity Dectection Silence Removal
[x_aem,x_iem] = RemoveSilence(x_aem,x_iem,Fs_new);

% Peak Normalisation
x_aem = x_aem - mean(x_aem);
x_iem = x_iem - mean(x_iem);

x_aem = x_aem./max(x_aem);
x_iem = x_iem./max(x_iem);

end
```

Figure 7.2: Software Implementation of the Pre-processing Module

7.3 Computation of Cluster Centres- Software Implementation

```
%% Compute average Cluster centre Transfer Functions

% Matrix containing average filter weights
% Dimensions : Number of Clusters x (NFFT/2 +1)
avg_cluster_weights = zeros(n_clusters, n_fft/2 +1);

% Iterate of each cluster
for i=1:n_clusters

    %find indicies of data frame in particular cluster
    cluster_idx = find(idx==i);

    iem_cluster = zeros(length(cluster_idx), win_length);
    aem_cluster = zeros(length(cluster_idx), win_length);
    cluster_w_mat = zeros(length(cluster_idx), n_fft/2 + 1);

    for j=1:length(cluster_idx)

        % get aem-iem pair from a particular cluster
        iem_cluster(j,:) = train_iem(cluster_idx(j),:);
        aem_cluster(j,:) = train_aem(cluster_idx(j),:);

        % Extract spectral features
        [s_aem, ~, ~, ~] = spectrogram(aem_cluster(j,:), win_length, hop_length, n_fft,
F, 'yaxis');
        [s_iem, ~, T, ~] = spectrogram(iem_cluster(j,:), win_length, hop_length, n_fft,
F, 'yaxis');

        %per-cluster per aem-iem pair tansfer function
        h = abs(s_aem)./abs(s_iem);
        h = smoothdata(h, 'movmean', smoothing_win_length);
        %matrix containing transfer function for all aem-iem instances in a
        %cluster
        cluster_h_mat(j,:) = h;

    end
```

Figure 7.3: Software Implementation of computation of cluster centres

7.4 Classification of Test Speech Utterances - Software Implementation

```
% Test utterance classification loop
% Iterate over test speech IEM frames
for i=1:nfrm_test
    %get LPCC coefficients of test data frames
    lpc_iem_test  = lpc(seg_test_iem(i,:),lpc_order);
    lpcc_iem_test = v_lpcar2cc(lpc_iem_test, n_lpcc);

    %fit new data frames into the existing clusters
    %Euclidean dissimilarity to find closest cluster centre
    [~,idx_test] = pdist2(C,lpcc_iem_test,'euclidean','Smallest',1);

    %get corresponding average cluster weights for data frames
    h = avg_cluster_weights(idx_test,:);
    % apply smoothing window of length : smoothing_win_length
    h = smoothdata(h, 'movmean', smoothing_win_length);

    % get estimated of AEM speech frame using selected mapping h
    [s_aem_frm, f, T, ~] = spectrogram(seg_test_iem(i,:), win_length, hop_length,
n_fft, F, 'yaxis');

    % apply the classified transfer function to the test speech frame
    s_aem_est(i,:) = h .* abs(s_aem_frm).';
    si(i,:) = abs(s_aem_frm).';
end
```

Figure 7.4: Software Implementation of classification of new test utterances, using the k-means cluster centres.

7.5 Artificial BWE enhanced IEM spectrogram

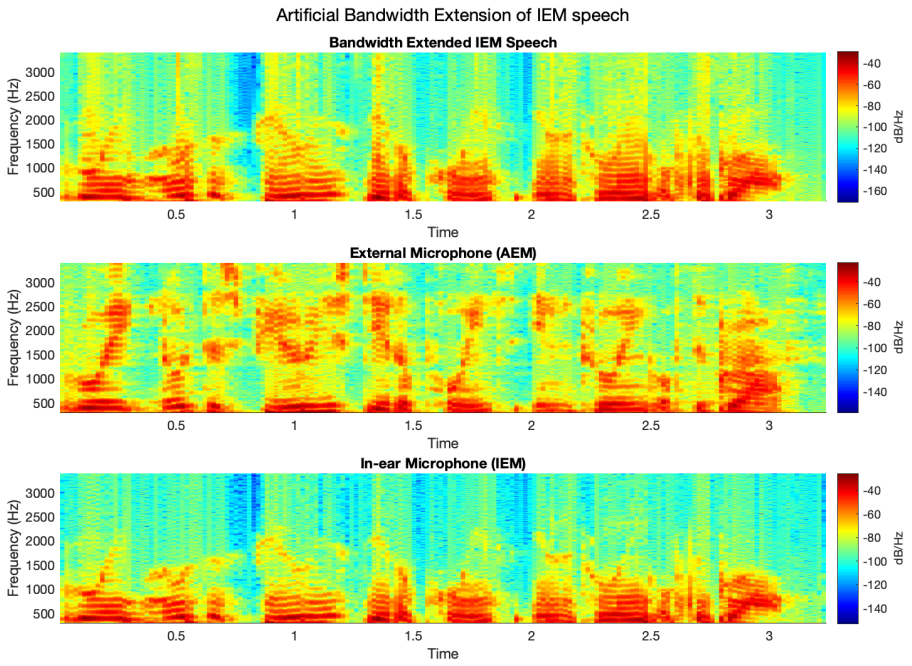


Figure 7.5: Spectrograms comparing the time-domain bandwidth extension approach. From top to bottom: Estimated AEM- BWE speech, Reference Speech, IEM speech

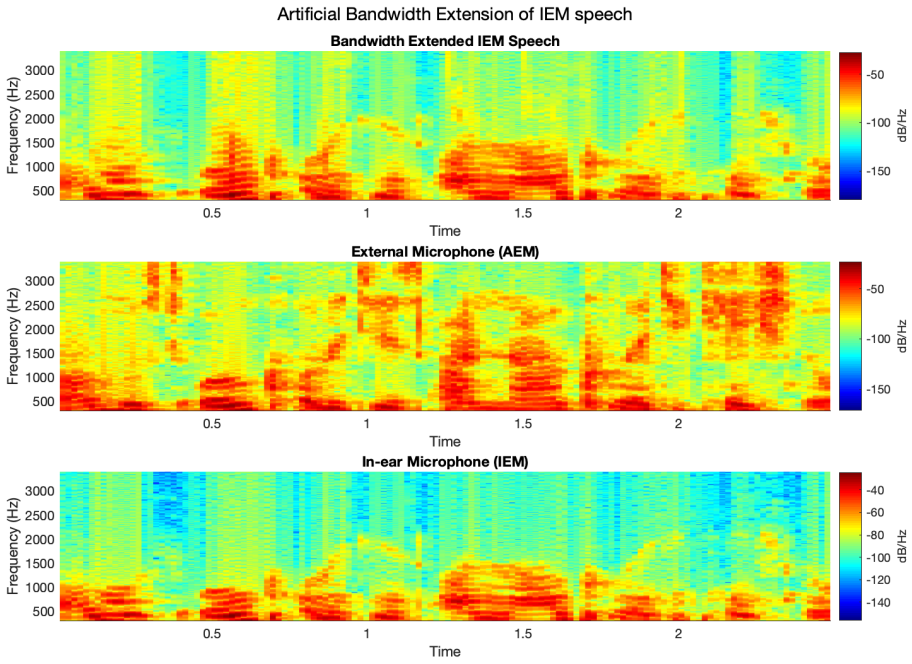


Figure 7.6: Spectrograms comparing the time-domain bandwidth extension approach. From top to bottom: Estimated AEM- BWE speech, Reference Speech, IEM speech

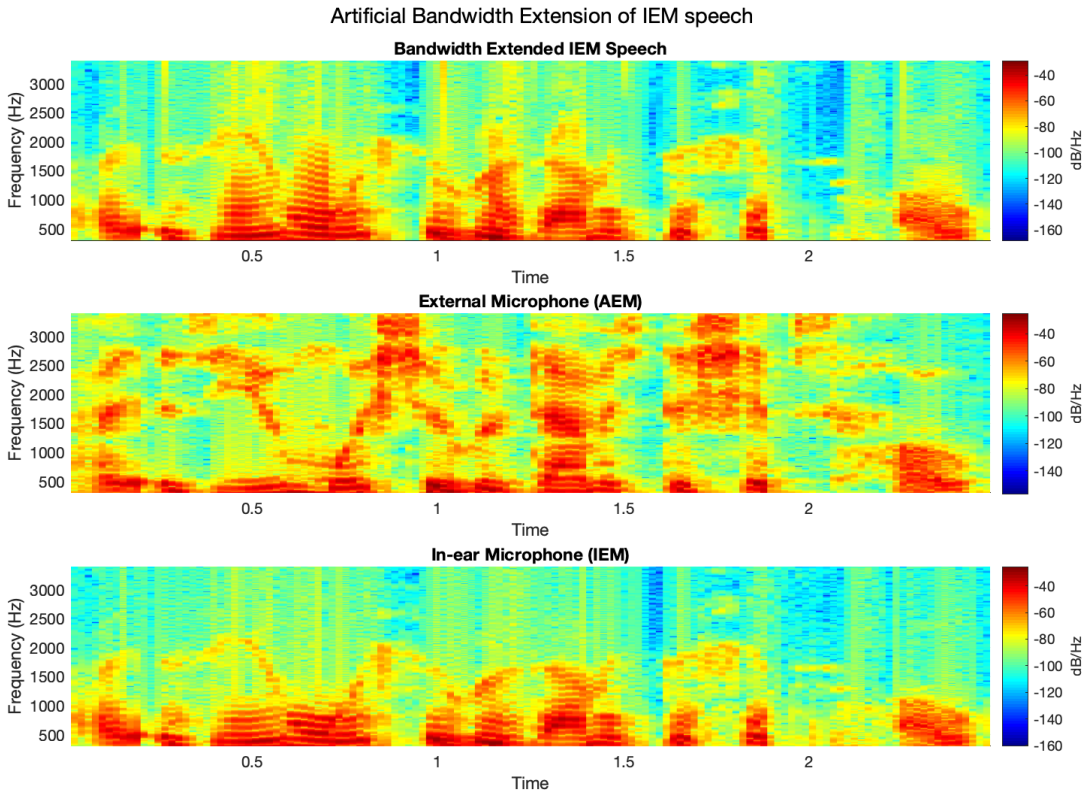


Figure 7.7: Spectrograms comparing the time-domain bandwidth extension approach. From top to bottom: Estimated AEM- BWE speech, Reference Speech, IEM speech

7.6 Log-Spectral Distortion Results: Artificial BWE

Utterance	Method	LSD (dB)
sp1_41	AEM vs IEM	5.982
	AEM vs BWE	5.197
sp1_42	AEM vs IEM	3.867
	AEM vs BWE	3.019
sp1_43	AEM vs IEM	4.791
	AEM vs BWE	3.620
sp1_44	AEM vs IEM	5.395
	AEM vs BWE	4.307
sp1_45	AEM vs IEM	4.939
	AEM vs BWE	4.433
sp1_46	AEM vs IEM	5.728
	AEM vs BWE	5.480
sp1_47	AEM vs IEM	3.357
	AEM vs BWE	3.010
sp1_48	AEM vs IEM	3.926
	AEM vs BWE	3.112
sp1_49	AEM vs IEM	4.498
	AEM vs BWE	3.502
sp1_50	AEM vs IEM	4.712
	AEM vs BWE	3.856
ALSD (dB)	**AEM vs IEM**	**4.742**
	AEM vs BWE	**4.044**

Table 7.1: Table shows the LSD in dB for Utterances in the Test Set. LSD calculated for two scenarios: AEM vs IEM speech, and AEM vs BWE method improved speech

7.7 Comparison of Spectral Domain methods: spectrograms

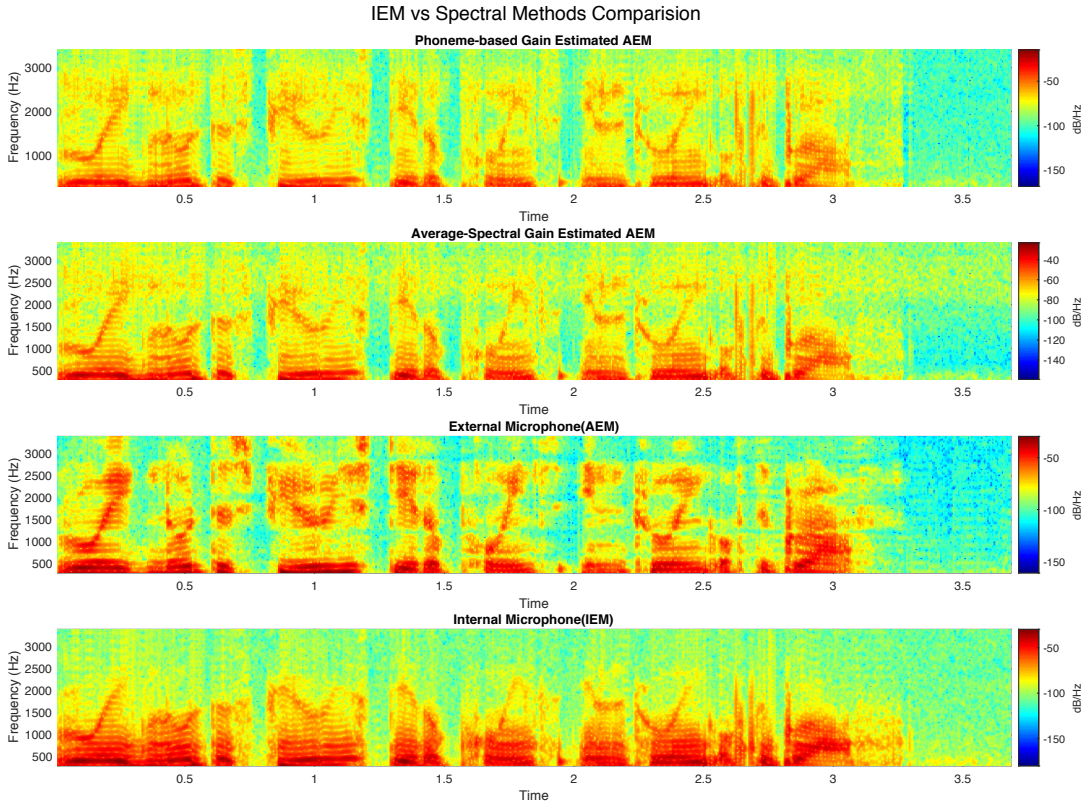


Figure 7.8: Spectrograms comparing the spectral-domain approaches. From top to bottom: Estimated AEM- Phoneme mapping, Estimated AEM- Linear Transfer Gain, Reference Speech, IEM speech

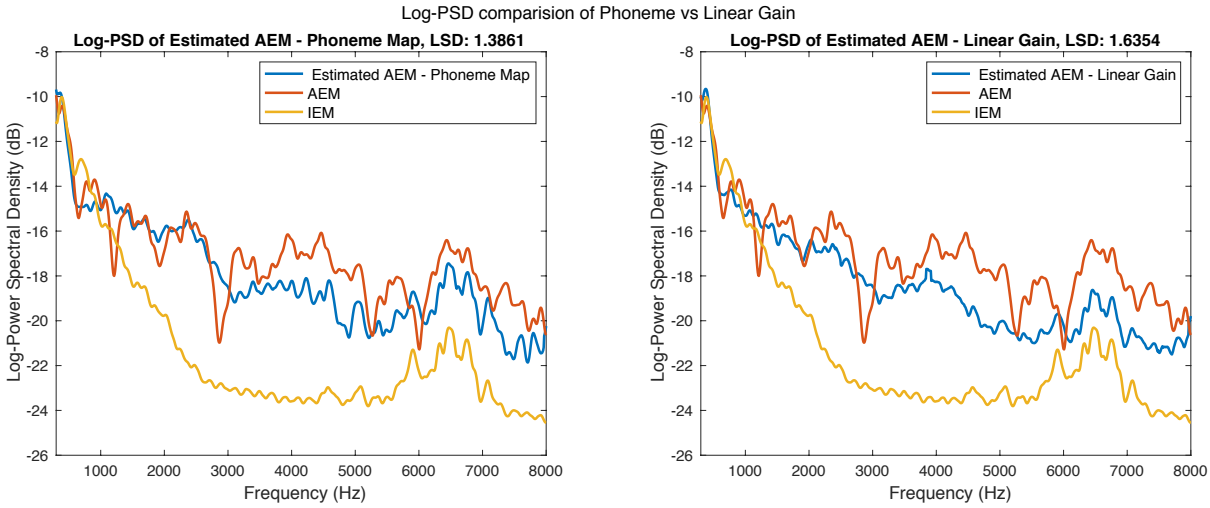


Figure 7.9: PSD plots comparing the spectral-domain approaches. Left: Estimated AEM- Phoneme mapping. Right: Estimated AEM- Linear Transfer Gain, Reference Speech, IEM speech

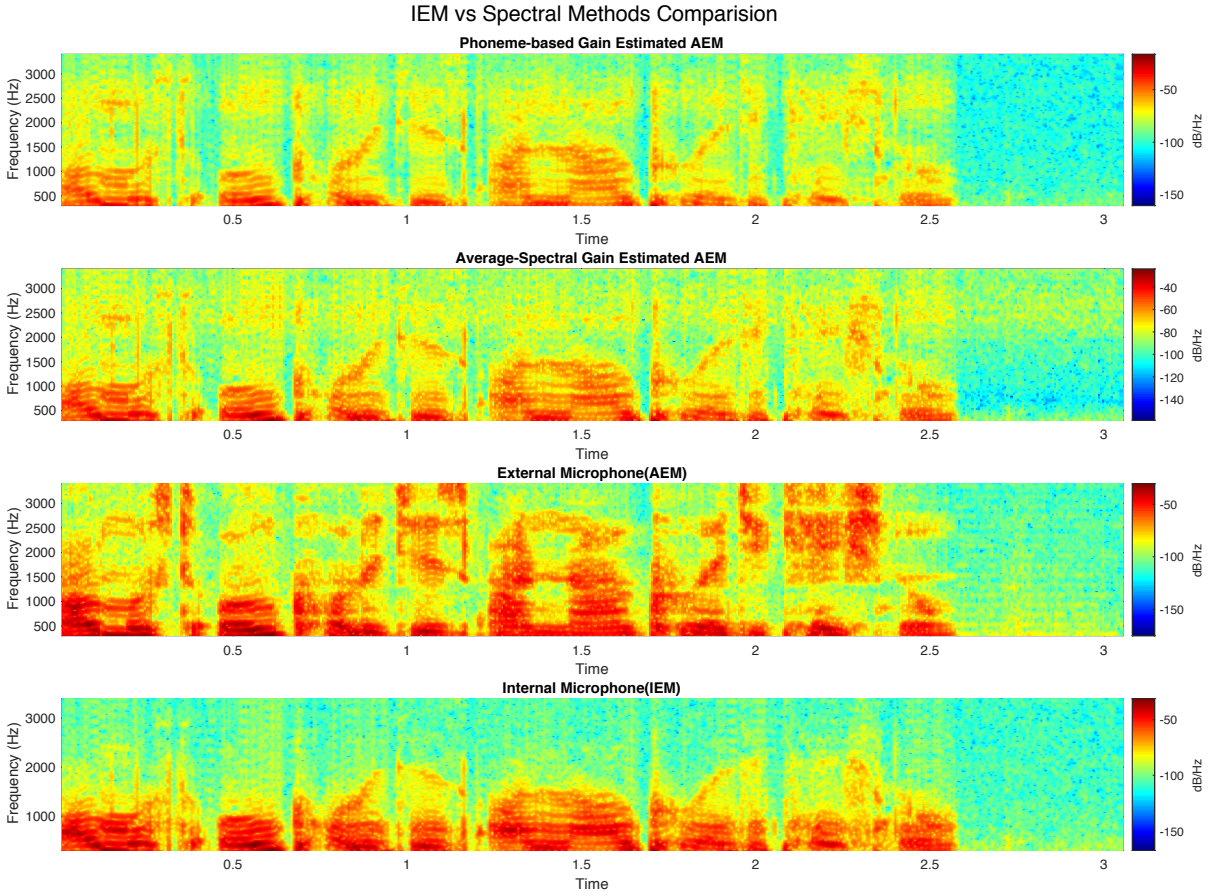


Figure 7.10: Spectrograms comparing the spectral-domain approaches. From top to bottom: Estimated AEM- Phoneme mapping, Estimated AEM- Linear Transfer Gain, Reference Speech, IEM speech

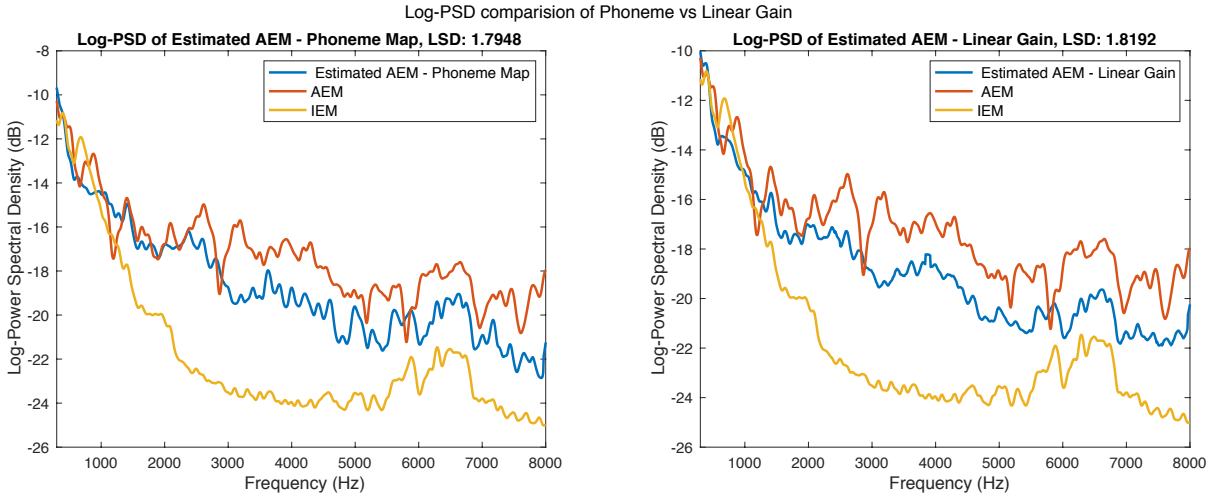


Figure 7.11: PSD plots comparing the spectral-domain approaches. Left: Estimated AEM- Phoneme mapping. Right: Estimated AEM- Linear Transfer Gain, Reference Speech, IEM speech

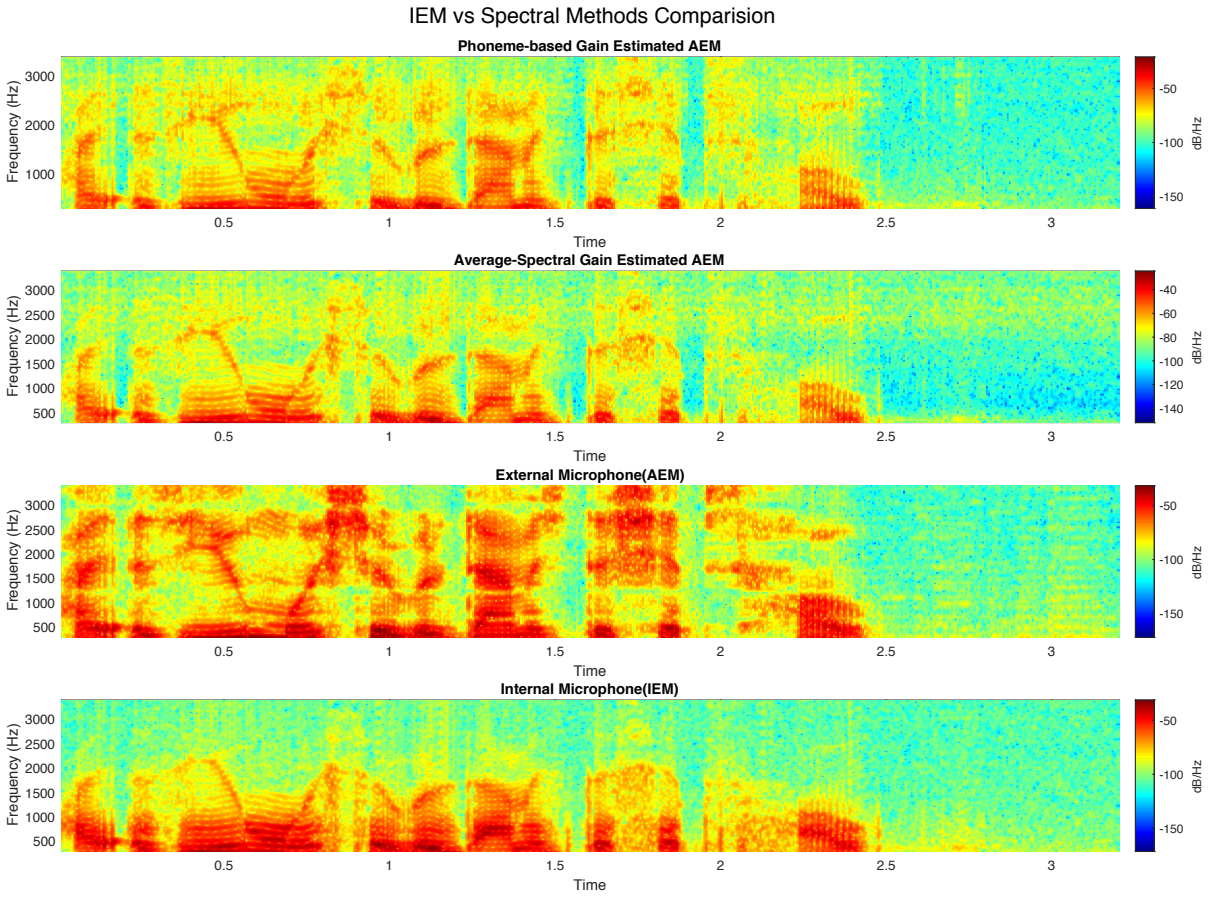


Figure 7.12: Spectrograms comparing the spectral-domain approaches. From top to bottom: Estimated AEM- Phoneme mapping, Estimated AEM- Linear Transfer Gain, Reference Speech, IEM speech

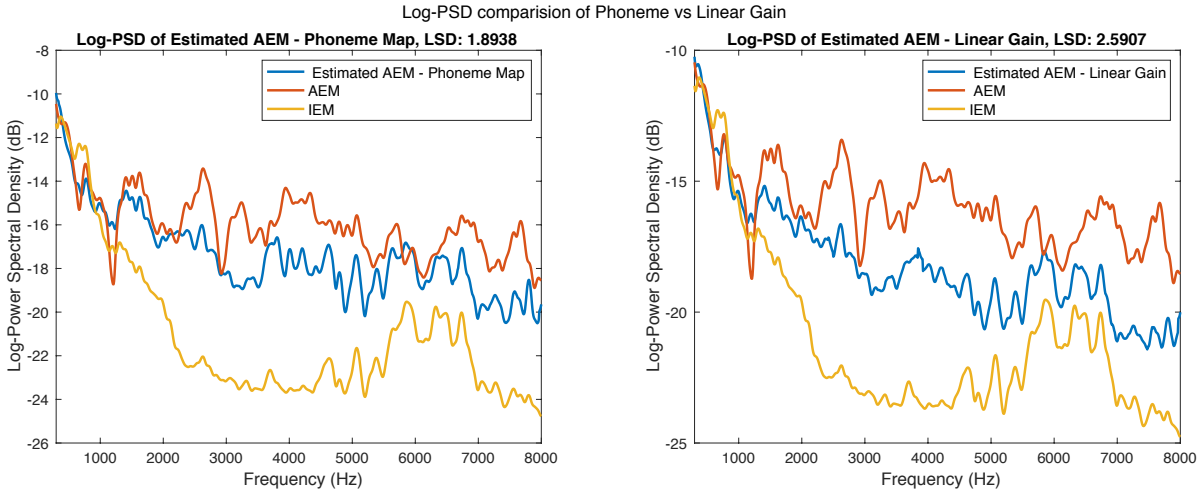


Figure 7.13: PSD plots comparing the spectral-domain approaches. Left: Estimated AEM- Phoneme mapping. Right: Estimated AEM- Linear Transfer Gain, Reference Speech, IEM speech

7.8 Example of Fricative modelling using spectral-domain methods

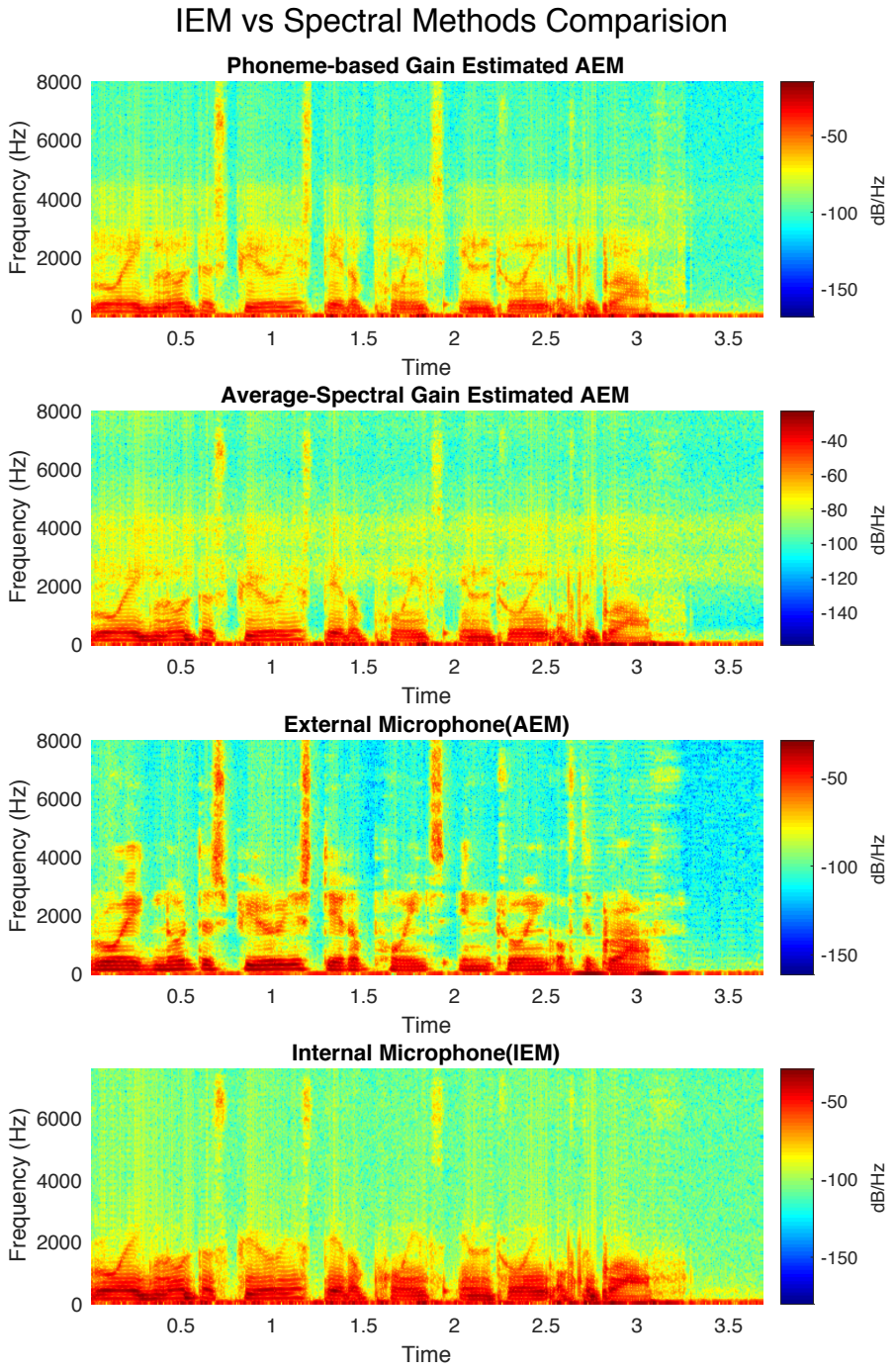


Figure 7.14: Comparison of fricative modelling using the spectral-domain methods - spectral gain, and phoneme-mapping transfer function

7.9 Log-Spectral Distortion Results: Spectral Gain Method

Utterance	Method	LSD (dB)		
		0-2 kHz	2-4kHz	4-8kHz
sp1_41	AEM vs. IEM	1.88	6.10	4.76
	AEM vs. SG	1.81	1.03	1.02
sp1_42	AEM vs. IEM	1.90	5.93	4.18
	AEM vs. SG	1.75	0.96	1.29
sp1_43	AEM vs. IEM	1.94	6.02	5.67
	AEM vs. SG	2.31	0.96	1.30
sp1_44	AEM vs. IEM	1.86	6.32	4.77
	AEM vs. SG	2.31	0.96	1.30
sp1_45	AEM vs. IEM	1.64	5.25	4.67
	AEM vs. SG	1.95	1.66	0.86
sp1_46	AEM vs. IEM	1.75	6.19	4.70
	AEM vs. SG	1.79	0.86	0.65
sp1_47	AEM vs. IEM	2.01	6.89	5.69
	AEM vs. SG	1.98	1.12	1.30
sp1_48	AEM vs. IEM	1.59	5.84	4.03
	AEM vs. SG	1.69	1.64	1.30
sp1_49	AEM vs. IEM	1.81	5.93	4.21
	AEM vs. SG	1.87	1.53	1.47
sp1_50	AEM vs. IEM	2.14	5.76	5.41
	AEM vs. SG	1.86	1.32	1.04

Table 7.2: Table shows the LSD in dB for Utterances in the Test Set. LSD calculated for two scenarios: AEM vs IEM speech, and AEM vs Spectral Gain (SG) method improved speech. LSD is decomposed into frequency segments: 0-2kHz, 2-4kHz, and 4-8kHz.

7.10 Log-Spectral Distortion Results: Phoneme-based mapping

Utterance	Method	LSD (dB)		
		0-2 kHz	2-4kHz	4-8kHz
sp1_41	AEM vs. IEM	1.88	6.10	4.76
	AEM vs. P-M	1.86	0.97	0.89
sp1_42	AEM vs. IEM	1.90	5.93	4.18
	AEM vs. P-M	1.82	1.00	0.95
sp1_43	AEM vs. IEM	1.94	6.02	5.67
	AEM vs. P-M	1.61	1.17	1.25
sp1_44	AEM vs. IEM	1.86	6.32	4.77
	AEM vs. P-M	1.97	1.28	0.88
sp1_45	AEM vs. IEM	1.64	5.25	4.67
	AEM vs. P-M	1.86	1.08	0.86
sp1_46	AEM vs. IEM	1.75	6.19	4.70
	AEM vs. P-M	2.01	1.19	1.00
sp1_47	AEM vs. IEM	2.01	6.89	5.69
	AEM vs. P-M	2.16	1.12	1.29
sp1_48	AEM vs. IEM	1.59	5.84	4.03
	AEM vs. P-M	1.73	1.18	1.16
sp1_49	AEM vs. IEM	1.81	5.93	4.21
	AEM vs. P-M	1.67	1.15	1.28
sp1_50	AEM vs. IEM	2.14	5.76	5.41
	AEM vs. P-M	1.80	1.21	1.05

Table 7.3: Table shows the LSD in dB for Utterances in the Test Set. LSD calculated for two scenarios: AEM vs IEM speech, and AEM vs Phoneme-based Mapping (P-M) method improved speech. LSD is decomposed into frequency segments: 0-2kHz, 2-4kHz, and 4-8kHz.

Cover Page: Imperial College Crest Source

Imperial College Coat of Arms on the Title page is sourced from [42]

Bibliography

[1] S. Mitra W.S. Gan and S.M. Kuo. Adaptive feedback active noise control headset: Implementation, evaluation and its extensions. *IEEE Trans. Consumer Electron*, 51:975–982, 2005. pages 4

[2] T.H. Falk R.E. Boucheral and J. Voix. Integration of a distance sensitive wireless communication protocol to hearing protectors equipped with in-ear microphones. *Proceedings of Meetings on Acoustics*, 19, 2013. pages 4

[3] Yanli Zheng, Zicheng Liu, Zhengyou Zhang, Michael Sinclair, Jasha Droppo, li Deng, Alex Acero, and Xuedong Huang. Air- and bone-conductive integrated microphones for robust speech detection and enhancement. pages 249 – 254, 01 2003. pages 4

[4] J.G. Casali and E.H. Berger. Technology advancements in hearing protection circa 1995: active noise reduction, frequency/amplitude-sensitivity, and uniform attenuation. *Am Ind Hyg Assoc J.*, page 175-185, 1996. pages 4

[5] M. A. Tuğtekin Turan and E. Erzin. Enhancement of throat microphone recordings by learning phone-dependent mappings of speech spectra. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7049–7053, 2013. pages 4, 8

[6] Valentin Goverdovsky, Wilhelm von Rosenberg, Takashi Nakamura, David Looney, David Sharp, C. Papavassiliou, Mary Morrell, and Danilo Mandic. Hearables: Multimodal physiological in-ear sensing. *Scientific Reports*, 7, 09 2016. pages 4, 8, 16, 18

[7] Emily Mullin. Voice analysis tech could diagnose disease" mit technology review, 2017. pages 4

[8] OSHA. Occupational noise exposure: Hearing conservation amendment, final rule. *Occupational Safety and Health Administration*, page 9738–9797, 1983. pages 4

[9] D.C. Byrne J.R. Franks W.J. Murphy, R.R. Davis. Advanced hearing protector study: Conducted at general motors metal fabricating division. *Flint Metal Center*, 2005. pages 4

[10] GMåns Eeg-Olofsson. Transmission of bone-conducted sound in the human skull based on vibration and perceptual measures. *Institute of Clinical Sciences at Sahlgrenska Academy University of Gothenburg*, 2012. pages 7

[11] Yvan Petit Martin K. Brummund, Franck Sgard and Frédéric Laville. Three-dimensional finite element modeling of the human external ear: Simulation study of the bone conduction occlusion effect. *The Journal of the Acoustical Society of America 135*, 135:1433, 2014. pages 9

[12] GMåns Eeg-Olofsson. Transmission of bone-conducted sound in the human skull based on vibration and perceptual measures. *Institute of Clinical Sciences at Sahlgrenska Academy University of Gothenburg*, 2012. pages 9

[13] J. Tonndorf. Bone conduction. studies in experimental animals. *Acta otolaryngologica Supplementum*, 1966. pages 9

[14] EH. Huizing. Bone conduction-the influence of the middle ear. *Acta otolaryngologica Supplementum*, 155:1–99, 1960. pages 9

[15] T. H. Falk R. E. Bouserhal and J Voix. On the potential for artificial bandwidth extension of bone and tissue conducted speech: A mutual information study. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, Australia*, page 5108–5112, 2015. pages 9

[16] William M. Fisher Jonathan G. Fiscus David S. Pallett Nancy L. Dahlgren Victor Zue John S. Garofolo, Lori F. Lamel. Timit acoustic-phonetic continuous speech corpus, 2014. data retrieved from TIMIT, `https://www.ldc.upenn.edu/language-resources/data`. pages 10

[17] Beerends J. Hollier M. Hekstra A. Rix, A. Ieee recommended practice for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics*, 17 no.3:225–246, September 1969. pages 10, 16

[18] Kenneth W. Bozeman Svante Granqvist Nathalie Henrich Christian T. Herbst David M. Howard Eric J. Hunter Dean Kaelin Raymond D. Kent Jody Kreiman Malte Kob Anders Löfqvist Scott McCoy Donald G. Miller Hubert Noé Ronald C. Scherer John R. Smith Brad H. Story Jan G. Švec Sten Ternström Joe Wolfe Ingo R. Titze, Ronald J. Baken. Toward a consensus on symbolic notation of harmonics, resonances, and formants in vocalization. *The Journal of the Acoustical Society of America*, 137:5108–5112, April 2015. pages 13

[19] Cluster analysis methods for speech recognition. Speech enhancement using a minimum mean square error short-time spectral amplitude estimator. *Department of Speech, Music and Hearing Royal Institute of Technology*, 32 no.6:1109–1121, February 2005. pages 15

[20] Vassilvitskii . Arthur, D. k-means++: the advantages of careful seeding. *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics Philadelphia, PA, USA.*, 32 no.6:1027–1035, 2007. pages 15

[21] Tomas Dekens and Werner Verhelst. Body conducted speech enhancement by equalization and signal fusion. *IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING,*, 21:12, DECEMBER 2013. pages 16, 43

[22] Tomoe & Nakagawa Kiyoshi Kondo, Kazuhiro & Fujita. Equalization of bone conducted speech for improved speech quality. *Symposium on Signal Processing and Information Technology*, page 5108–5112, April 2006. pages 16, 17, 31, 44, 46, 48

[23] Yong-Shik SHIN Hochong PARK and Seong-Hyeon SHIN. Speech qualityenhancement for in-ear microphone based on neural networks. *The Institution ofElectronic, Information and Communication Engineers*, E104-D:8, April 2019. pages 16, 17

[24] Antoine & Voix Jérémie Bouserhal, Rachel & Bernier. An in-ear speech database in varying conditions of the audio-phonation loop. *The Journal of the Acoustical Society of America*, 145:1069–1077, April 2019. pages 16

[25] P. C. Loizou. Speech quality assessment. *Multimedia Analysis, Processing and Communications (Springer, Berlin)*, page 623–654, 2011. pages 16

[26] Beerends J. Hollier M. Hekstra A. Rix, A. Perceptual evaluation of speech quality(pesq) - a new method for speech quality assessment of telephone networks and codecs. *Proc. IEEE Int. Conf. Acoust, Speech, Signal Processing*, 2:749–752, 2001. pages 16

[27] R. Kubichek. Mel-cepstral distance measure for objective speech quality assessment. *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing, Victoria, BC, Canada*, 1:125–128, 1993. pages 16

[28] Tomikura T. Shimamura, T. Quality improvement of bone-conducted speech. *In: Proceedings of the ECCTD*, page 1–4, 2005. pages 17, 18

[29] A. Bernier and J. Voix. Signal characterisation of occluded in-ear versus free-air voice pickup on human subjects. *Canadian Acoustics*, 38:78–79, 2010. pages 17

[30] T.H. Falk R.E. Bouserhal and J. Voix. In-ear microphone speech quality enhancement via adaptive filtering and artificial bandwidth extension. *The Journal of the Acoustical Society of America*, 141:1321–1331, 2017. pages 17, 19, 46

[31] Cohen I. Li, M. and S. Mousazadeh. Multisensory speech enhancement in noisy environments using bone-conducted and air conducted microphones. *IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP), Chengdu, China*, page 1–5, 2014. pages 17

[32] Acero A. Seltzer, M. L. and J. Droppo. Robust bandwidth extension of noise-corrupted narrowband speech. *INTERSPEECH, Lisbon, Portugal*, page 1509–1512, 2005. pages 17

[33] T. Shimamura D. Watanabe, Y. Sugiura and H. Makinae. Speech enhancement for bone-conducted speech based on low-order cepstrum restoration. *2017 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), Xiamen*, pages 212–216, 2017. pages 17

[34] T. Shimamura S. Handa and H. Makinae. Importance of spectral envelope in bone-conducted speech. *IEICE Technical Report*, 115-208:75–80, 2015. pages 17

[35] You-Jin & Tsao Yu & Huang Jen-Wei & Wang Hsin-min Chang-Le & Fu, Szu-Wei & Lee. Multichannel speech enhancement by raw waveform-mapping using fully convolutional networks. 2019. pages 18

[36] Tulay Adali Simon Haykin. *Adaptive and Learning Systems for Signal Processing, Communications, and Control, Adaptive Signal Processing, Chapter 7*. Hoboken, NJ, USA: John Wiley Sons, Inc., Reading, Massachusetts, 2010. pages 20, 21

[37] Esfandiar Zavarehei. Log spectral distance - silence removal, 2020. retrieved from MATLAB Central File Exchange, `https://www.mathworks.com/matlabcentral/fileexchange/9998-log-spectral-distance`. pages 24

[38] Y. Ephraim ; D. Malah. Speech enhancement using a minimum mean square error short-time spectral amplitude estimator. *IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING*, 32 no.6:1109–1121, DECEMBER 1984. pages 24

[39] Mike Brookes. VOICEBOX: Speech Processing Toolbox for MATLAB, 2011. retrieved on 8 June 2020 from: `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`. pages 25, 30

[40] S. Furui. *Digital Speech Processing, Synthesis, and Recognition*. Marcel Deckker, Inc., New York, 1989. pages 28

[41] K. Paliwal and W. Kleijn. Quantization of lpc parameters. *Speech Coding and Synthesis (Elsevier Science B.V., Amsterdam)*, 32 no.6:433–466, 1995. pages 32

[42] Shadowssettle / cc by (https://creativecommons.org/licenses/by/3.0) - File:Shield of Imperial College London.svg. Accessed: 16/06/2020. pages 61