Explicit memory inclusion for efficient artificial bandwidth extension

Pramod Bachhav

EURECOM
Sophia Antipolis

SORBONNE UNIVERSITÉ

Explicit memory inclusion for efficient artificial bandwidth extension

Dissertation

submitted to

Sorbonne Université

in partial fulfilment of the requirements for the

degree of Doctor of Philosophy

by

Pramod BACHHAV

14th November, 2019

Thesis Advisor	**Prof. Nicholas EVANS**, EURECOM, France
Reviewers	**Prof. Tim FINGSCHEIDT**, Technische Universität Braunschweig, Germany
	Prof. Tom BÄCKSTRÖM, Aalto University, Finland
Examiners	**Dr. Christophe BEAUGEANT**, Renault, France
	Mme. Aurélie DONJON, NxP Software, France
	Prof. Marc DACIER, EURECOM, France

Abstract

Legacy telephony terminals and infrastructure typically operate with bandwidths of 0.3-3.4kHz. At such narrow bandwidths, speech quality and intelligibility can be poor, especially for consonant sounds. Today's terminals and infrastructure, in contrast, operate at wider bandwidths for which speech quality and intelligibility is greatly improved. Naturally, though, the complete transition from narrowband to wideband (0.05-7kHz) and super-wideband (0.05-14kHz) communications will require considerable time. As a result, wide and super-wideband technology must interoperate with narrowband technology. In this case, users will experience substantial variations in speech quality and intelligibility. Artificial bandwidth extension (ABE) algorithms have been developed to improve speech quality and intelligibility in situations where wideband (or super-wideband) capable technology is used alongside narrowband (or wideband) terminals or infrastructure. ABE involves the automatic estimation of missing higher frequency components from available lower frequency components.

Most ABE algorithms exploit contextual information or memory captured via the use of static or dynamic features extracted from neighbouring speech frames. The use of memory leads to higher dimensional features and increased computational complexity. When information from look-ahead frames is also utilised, then latency also increases. Past work points toward the benefit to ABE of exploiting memory in the form of dynamic features with a standard regression model. Even so, the literature is missing a quantitative analysis of the relative benefit of explicit memory inclusion. The research presented in this thesis assesses the degree to which explicit memory is of benefit and furthermore reports a number of different techniques that allow for its inclusion without significant increases to latency and computational complexity. Benefits are shown through both a quantitative analysis with an information-theoretic measure and subjective listening tests. Key contributions relate to the preservation of computational efficiency through the

use of dimensionality reduction in the form of principal component analysis, semi-supervised stacked autoencoders and conditional variational auto-encoders. The two latter techniques optimise dimensionality reduction to deliver superior ABE performance.

The potential gain in speech quality when extending from wide to superwide band speech is much less than when extending from narrow to wideband speech. In this case, increases to computational complexity can be difficult to justify. The final key contributions reported in this thesis involve the development of an especially efficient approach to super-wideband ABE based on linear prediction analysis-synthesis which avoids the statistical estimation of missing higher frequency components. In addition to computational efficiency, the solution delivers speech of superior quality to wideband speech signals processed with an adaptive-multirate wideband codec.

Contents

List of Abbreviations

1G	1st generation
2G	2nd generation
3G	3rd generation
3GPP	3rd Generation Partnership Project
3GPP2	3rd Generation Partnership Project 2
4G	4th generation
ABE	artificial bandwidth extension
ACELP	algebraic code-excited linear prediction
ACR	absolute category rating
ADPCM	adaptive differential pulse code modulation
AE	auto-encoder
AM	amplitude modulation
AMPS	Advanced Mobile Telephone System
AMR	adaptive multi-rate
AMR-WB	adaptive multi-rate wideband
ASR	automatic speech recognition
CCITT	International Telegraph and Telephone Consultative Committee
CCR	comparison category rating

CDMA	code division multiple access
CELP	code-excited linear prediction
CGM	conditional generative model
CI	confidence interval
CMOS	comparison mean opinion score
CS-ACELP	conjugate-structure algebraic code-excited linear prediction
CVAE	conditional variational auto-encoder
DCR	degradation category rating
DFT	discrete Fourier transform
DNN	deep neural network
EFR	enhanced full rate
ELBO	evidence lower bound
ETSI	European Telecommunications Standards Institute
EVS	Enhanced voice services
FDMA	Frequency Division Multiple Access
FFT	fast Fourier transform
FM	frequency modulation
FT	Fourier transform
GMM	Gaussian mixture model
GMMR	Gaussian mixture model regression
GPU	graphics processing unit
GSM	Global System for Mobile
HB	highband
HD	high definition

HF	high frequency
HMM	hidden Markov model
HPF	highpass filter
IMT-2000	international mobile telecommunications-2000
IRS	intermediate reference system
ISDN	integrated services digital network
ITU	International Telecommunication Union
ITU-T	Telecommunication Standardization Sector of the International Telecommunication Union
kbps	kilobits per second
KLT	Karhunen–Loève transform
LDA	linear discriminant analysis
LF	low frequency
logMFE	log-Mel filter energy
LP	linear prediction
LPC	linear predictive coding
LPF	lowpass filter
LPS	log power spectrum
LSD	logarithmic spectral distortion
LSF	line spectral frequency
MAP	maximum a posteriori
MDCT	modified discrete cosine transform
MFB	Mel filterbank
MFCC	mel-frequency cepstral coefficients

MI	mutual information
ML	maximum likelihood
MMSE	minimum mean square error
MOS	mean opinion score
MSIN	mobile station input
MVN	mean and variance normalisation
NB	narrowband
NMT	Nordic Mobile Telephone (NMT)
OLA	overlap and add
PCA	principal component analysis
PCM	pulse code modulation
PDF	probability density function
PESQ	perceptual evaluation of speech quality
PGM	probabilistic graphical model
PS	power spectrum
PSTN	public switched telephone network
RE	reconstruction error
RMS	root-mean-square
RPCA	robust principal component analysis
RPE-LTP	regular pulse excitation with long-term prediction (codec)
SAE	stacked auto-encoder
SGD	stochastic gradient descent
SGVB	stochastic gradient variational Bayes (SGVB)
SLP	selective linear prediction

SNR	signal-to-noise ratio
SSAE	semi-supervised stacked auto-encoder
SWB	super-wideband
SWBE	superwide bandwidth extension
UMTS	Universal Mobile Telecommunication System
VAE	variational auto-encoder
VoIP	voice over Internet Protocol
VQ	vector quantisation
WB	wideband

List of Figures

List of Tables

Chapter 1

Introduction

The quality of speech offered by communication systems is highly dependent on the bandwidth of speech signals. Due to the bandwidth limitations imposed by communication systems, speech signals often lack higher frequency content and thus suffer from limited quality and intelligibility. Artificial bandwidth extension (ABE) algorithms have thus been developed to improve the quality of speech signals by artificially estimating the missing frequency components. Higher bandwidths lead to better, more comfortable conversations due to increased speech quality and intelligibility.

This thesis concerns the research topic of ABE for speech signals and its applications. This chapter provides an introduction to the topic. Section 1.1 provides a brief history on the evolution of telephony systems from analog to digital and from wired to wireless modes of communication. The speech production mechanism and physiology is discussed in Section 1.2 and describes different speech sounds and their spectral characteristics. The effect of bandwidth on speech quality and intelligibility is also explained. In Section 1.3 different types of speech coding methods (or codecs) based on their operational bandwidths are explained. Section 1.4 then introduces narrowband-to-widebandwith extension and its applications. Wideband-to-super-wide bandwidth extension is explained in Section 1.5. Sections 1.6 and 1.7 present contributions and an outline of this thesis respectively.

1.1 Evolution of communication systems

This section presents a brief overview of the evolution of communication systems from analog to digital and from wired to wireless modes of communication.

1.1.1 Analog and digital telephony

The transmission of the very first sentence "Mr. Watson, come here, I want to see you," uttered by Graham Bell over an electric *telephone* in 1876 laid the foundation for enormous progress in communication systems [1]. This led to the installation of over 3000 telephones and the first public telephone exchange in the US by 1878 [2]. Different operators started providing telephone services, however, subscribers to different services could not communicate with each other. American Telephone and Telegraph (AT&T) started providing an universal service and an unified telephone network to subscribers, allowing them to make long-distance telephone calls by the 1950s [3, Section 1.2.4]. By then, the telephone networks, referred to as public switched telephone networks (PSTNs)[1], were still analog utilizing frequency division multiplexing (FDM). The speech signals, limited to a frequency range of 0.3-3.4Hz referred to as *narrowband* (NB), were transmitted over different frequency channels with a frequency separation of 4kHz. The narrowband limitation of transmission comes from the characteristics of the transducers and hardware (such as copper lines) employed in PSTNs[2].

A demonstration of wireless the transmission of Morse code signals by Guglielmo Marconi in 1895 started parallel developments of radio communications. The first wireless voice transmission in 1915 signaled a start to the convergence of radio and telephony [3, Section 1.2.3]. In 1937, Alex Reeves conceived the idea of pulse code modulation (PCM) based on the time-division multiplexing principle [5]. The development of PCM marked the first step towards digitization for voice communications. Due to the invention of transistors, the commercial use of PCM was possible only in the late 1950s [6] when the era of digital transmission of speech over telephone networks started. In accordance with the then existing

[1] Also known as plain old telephone services (POTSs).

[2] The lower limit, i.e., 300Hz was chosen to decrease susceptibility to interference caused by AC electric power lines. The analysis presented in [4] showed that a bandwidth of 3 or 3.1kHz provided good quality both in terms of articulation and naturalness. The suggested bandwidth was a economical choice to achieve desired transmission quality and thus the upper limit of the telephone band was set 3.4kHz.

PSTNs, PCM adopted the typical bandwidth of 0.3-3.4kHz[3] for communication and therefore, for many years, subscribers were offered only NB communication services.

1.1.2 Wireless cellular networks

After Marconi's successful attempt at wireless transmission, engineers and scientists started research on developing efficient means of communications using radio frequency (RF)/radio waves. The idea of cellular telephone systems started to be explored in the 1970s. The concept involved the division of a geographical area into adjacent, non-overlapping, hexagonal-shaped cells [3, Section 1.2.6]. In this scheme, all mobile units in a given cell could communicate via a transmitter and a receiver dedicated to each cell (referred to as the base station); communication (or handoff) between the units crossing cell boundaries was coordinated via a mobile switching station. The first generation (1G) wireless mobile phone system was developed by Martin Cooper at Motorola in 1973 but not commercialised until 1984. Wireless communications have progressed remarkably in last few decades. Mobile handsets have also advanced alongside the generations (from 1G to 4G) with added functionalities. An illustration of the typical mobile devices introduced in different generations is shown in Fig 1.1. As of today, current wireless mobile telephone systems can be divided into four generations.

The first generation cellular systems, introduced in the 1980s, used analogue cellular and cordless telephone technology. The cordless telephone was connected to PSTNs over radio. The Advanced Mobile Telephone System (AMPS) and Nordic Mobile Telephone (NMT) are notable examples of 1G analogue standards. A frequency division multiple access (FDMA) technique was utilised enabling multiple users to share the same frequency spectrum. Transmissions over radio were susceptible to eavesdropping and could be easily intercepted by a standard radio receiver [5, Section 1.4.1].

Second generation (2G) systems, introduced in the late 1980s, used digital

[3]In 1960s, the bandwidth of the PSTN was standardized by Consultative Committee for International Telephony and Telegraphy (CCITT) to 0.3-3.4Hz. CCITT was later renamed to International Telecommunication Union for Telecommunication standardization sector (ITU-T). According to ITU-T Rec. G.120 [7] the attenuation in the NB should not exceed 9dB compared to the value for 1020Hz whereas the attenuation distortion at 0.3kHz and 3.4kHz should never exceed 3dB.

Figure 1.1: An evolution of mobile handsets with advancements in the cellular communication systems reproduced from [9].

speech transmission. These systems supported additional services such as voice mail, text transmission, speed dialing, roaming, etc. 2G systems used advanced coding and compression techniques to utilise the allocated spectrum more efficiently. Network-control techniques were improved to conserve bandwidth and privacy was improved to prevent eavesdropping [5, Section 1.4.2]. Widespread 2G standards are the Global System for Mobile communications (GSM), IS-136 and IS-95[4]. 2G services were designed specifically for voice transmission and were not efficient for data transmission [8, page 321].

Third-generation systems, introduced in the early 2000s, provide advanced voice and high-speed data services that could not be delivered via 2G technology. While data transmission is done via packet switching, voice is transmitted using circuit-switching technology. The most common 3G technologies are Universal Mobile Telecommunication System (UMTS) or wideband CDMA (WCDMA), time division-synchronous CDMA (TD-SCDMA) and CDMA2000. These are collectively referred to as International Mobile Telecommunications-2000 (IMT-2000) [8, page 324]. Most GSM operators upgraded (from 2G) to UMTS/WCDMA 3G services. Upgrades usually required changes to existing infrastructure in the form of more base stations and the replacement of time-division access by code-division access.

Fourth generation (4G) systems employ an all-internet packet infrastructure

[4]The standards IS-136 and IS-95 employ time-division multiple access (TDMA) and code division multiple access (CDMA) techniques respectively

that supports data rates of 100Mbps. Packet switching technology is used for both voice as well as data. Worldwide Interoperability for Microwave Access 2 (WiMAX 2) and Long-Term Evolution Advanced (LTE-Advanced) are the two popular 4G protocols, also referred to together as IMT advanced [8, page 332].

1.2 Speech production

A study of the underlying anatomy and physiology of the human speech production system provides useful insights in order to analyse different acoustic as well as articulatory properties of speech signals. These properties help to understand spectral and temporal characteristics of various speech sounds. A model of the human speech production system is illustrated in Fig. 1.2. A *speech signal* is produced by a speaker at his/her mouth or lips in the form of pressure waves. The organs that are involved in the speech production mechanism are: the lungs, larynx and vocal tract. The *lungs* produce an airflow which is modulated by *vocal chords* or *vocal folds* of the larynx. The airflow passing through the *glottis* – a slit-like orifice between the two vocal folds – is converted to either a quasi-periodic or noisy airflow by vibration of the vocal folds. The resultant airflow source excites the vocal tract that comprises oral, nasal and pharynx cavities. The *vocal tract* performs spectral shaping or colouring of the excitation source. The subsequent variation of air pressure at the lips is radiated in the form of travelling waves called speech [10, Section 3.1]. Speech signals can be seen as the output of a filtering operation in which the vocal tract system (or filter) is excited by the modulation of an excitation source or airflow. The mechanism is typically known as the *source-filter model* of speech production which allows the modelling of speech signals as a convolution of the impulse response of the vocal tract filter[5] and the excitation source (also referred to as *glottal flow*).

Speech is a *non-stationary* signal consisting of different speech sounds, each of which is characterised by a distinct position of the vocal tract articulators (vocal folds, tongue, lips, teeth, velum, jaw) [11, Section 3.1]. Speech sounds are, therefore classified according to: the nature of the excitation source (which is mainly categorised as periodic, noisy, impulsive or combinations of the three) and the shape

[5]The vocal tract can be modelled as a linear time-invariant filter which exhibits resonance frequencies, typically known as *formant frequencies* or *formants*. Generally vocal tract takes form an all-pole filter – the approach referred to as *linear prediction analysis* – where the conjugate poles represent the formants.

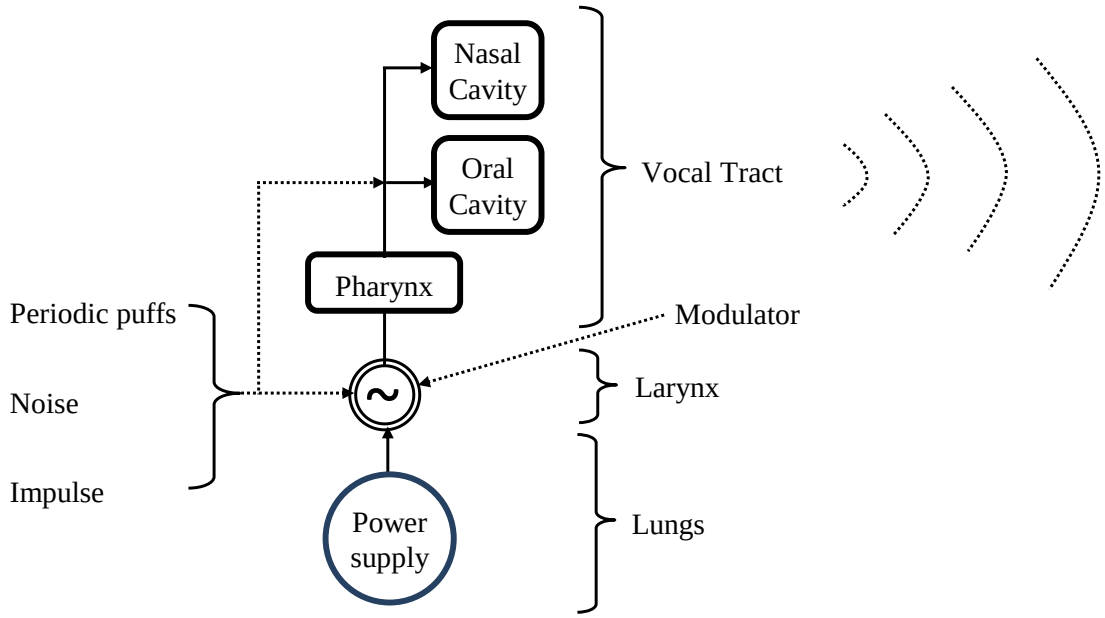


Figure 1.2: Model of human speech production mechanism (adapted from [10]).

of the vocal tract (which can be described by the place and manner of articulation (or constriction) in the vocal tract) [10, Section 3.4]. This section provides a brief overview of different speech sounds, their spectral characteristics and how the bandwidth limitations imposed by telephone filter results in intelligibility and quality.

1.2.1 Speech sounds

Speech sounds are broadly divided into two categories: vowels and consonants.

Vowels form the largest group of *phonemes*. The characteristics of vowels differ based on the position of the tongue – that mainly determine the vocal tract shape – towards the front, centre or back of the oral cavity. Each vowel phoneme thus corresponds to an unique, stable vocal tract configuration during most of the sound generation. The excitation source has quasi-periodic nature which is generated by vibration of the vocal folds at a certain fundamental frequency (also knowns as *pitch frequency* or *pitch*).

Consonants form the second largest group of phonemes which are sub-categorised into nasals, plosives, fricatives, whispers and affricates.

- *Nasals* or *murmurs* are closest to vowels and are produced at the nostrils by the quasi-periodic airflow only through the nasal cavity; the oral cavity remains constricted. Depending upon the place of constriction that is formed by the tongue across the oral cavity, nasals are distinguished, e.g. /m/ as in "mo" and /n/ as in "no".

- *Fricatives* are of two types. Unvoiced fricatives (e.g., /sh/ in "should") are characterised by a noise source generated due to turbulent airflow near the vocal tract constriction. The noise source is spectrally shaped depending upon the location and the degree of constriction formed by the tongue at the teeth or lips or along the oral cavity. In contrast, voiced fricatives (e.g. /z/ as in "zebra") are generated by the simultaneous generation of noise at the constriction and vibration of the vocal folds. These sounds thus are formed by a combination of a noisy and periodic airflow.

- During production of *plosive* or *stop* sounds, the air pressure is first built up due to closure of the oral cavity. The pressure is then released over a very short duration. This results in a burst or impulsive source that excites the vocal tract at the constriction. In unvoiced plosives (e.g., /k/ in "baker"), the burst is followed by aspiration[6] caused by turbulence at the open vocal folds. In voiced plosives (e.g., /g/ in "go") there is little or no aspiration as the vocal folds are also vibrating.

- *Whisper* sounds are similar to unvoiced fricatives, however, the turbulence occurs at the glottis rather than at a vocal tract construction, e.g, /h/ as in "he". The size of the glottis influences the spectral characteristics of whisper sounds.

There is another category of phones which represents *transitional* speech sounds. These are associated with changes or transitions during movement of articulators from one position to another. Such sounds are "non stationary" and are associated with the rapid spectral changes during transition between two articulatory states corresponding to two different sounds. This phenomenon is known as *coarticulation*.

- *Diphthongs* are produced by vibrating vocal folds similar to vowels, however, the vocal tract does not remain steady (as in vowels) but varies smoothly

[6] *Aspiration* is caused due to turbulence of the glottal airflow at the glottis when the vocal folds are open. The vocal folds do not vibrate, or remain fixed partly or completely leading to whispered and breathy voice respectively [10, page 64].

between two vowel configurations. Diphthongs are thus characterised by formant transitions as the vocal tract articulation changes gradually between two vowel positions. The examples of diphthongs are /Y/ as in "hide", /W/ as in "out", /O/ as in "boy" and /JU/ as in "new".

- *Semi-vowels* or *sonorants* are categorised into glides and liquids. *Glides* (e.g., /w/ as in "we" and /y/ in "you") are dynamic and transitional sounds that often occur before a vowel or between vowels. In the later case, they are similar to diphthongs, however the constriction of the oral tract is narrower and the transition between two vowels in quicker than in diphthongs. Thus glides are characterised with faster formant transitions and weaker articulation. *Liquids* (e.g. /r/ as in "read" and /l/ as in "let") exhibit different types of constriction formed by the tongue than in glides.

- *Affricates* are the sounds representing transitions from plosives to fricatives, e.g., affricate /tS/ as in "chew" representing a transition from the plosive /t/ to the fricative /S/.

1.2.2 Spectral characteristics of speech sounds

Different speech sounds are characterised by different temporal and spectral characteristics. They can be distinguished from each other based upon their acoustic properties such as rapid transitions in spectral content, abrupt changes in amplitude, presence or lack or combination of voicing and aspiration and, the spectral shape attributed to the vocal tract configuration [12]. The spectral content of speech sounds is mainly dependant upon the nature of the glottal airflow or excitation source and the shape of the vocal tract[7]. Based on their spectral properties, speech sounds can be classified into voiced and unvoiced sounds.

- The voiced sounds are the result of excitation of the vocal tract by a quasi-periodic glottal airflow. They are characterised by a quasi-periodic time-domain waveform with large variations in amplitude. The periodicity is measured in terms of the *pitch period*. The magnitude spectra of voiced sounds thus exhibit harmonic structure with peaks at integer multiples of the fundamental frequency or *pitch*, especially in the lower frequency region. The

[7]The shape of the vocal tract is defined by the place (or manner) of constriction (also referred to as articulation.

spectral envelope (smoothed version of the speech spectrum) is characterised by the presence of peaks which correspond to formants of the vocal tract filter.

Vowels and diphthongs (except when whispered) are mainly voiced sounds which ranging from 50 to 400 ms duration in normal speech. While vowels exhibit line spectra with energy concentrated at multiples of F0, the spectra primarily are characterised by the first three formants: F1, F2 and F3. They occur on average every 1kHz for adult males [11, Section 3.4.2].

Voiced sounds generally have low-pass characteristics and most of the energy is concentrated in the lower part (below 1kHz) of the audio spectrum. Thus, a significant portion of the voiced spectrum is covered adequately by NB or telephone speech despite the bandwidth limitations. While the fundamental frequency might be missing, the human ear is nonetheless capable of *hearing* pitch properly [13, Section 2.2.1].

- Unvoiced sounds are characterised by time domain waveforms with relatively lower amplitude than voiced sounds, however, with rapid variations due to noise-like nature of the excitation source. Thus, the spectrum of unvoiced speech extends over the entire audio spectrum. Unvoiced sounds contain significant energy above 3.4kHz and thus a larger portion of information is missing in NB speech signals.

  Fricatives fall into the category of unvoiced sounds. Fricatives are characterised by a lack of energy at lower frequencies. They have a highpass spectrum and therefore, most of the energy is concentrated above 2.5-3.2kHz frequency region.

1.2.3 Effect of bandwidth on speech quality and intelligibility

The choice of cut-off frequencies (0.3kHz and 3.4kHz) and characteristics of the telephone filter specified in ITU-T Rec. G.120 [7] was based on the bandwidth limitations imposed by analog transmission systems. The choices were motivated by results obtained from subjective listening tests [14, Section 10.1]. In order to retain compatibility with existing analogue PSTNs, initial progress in digital telephony have occurred with bandwidth constraints of 0.3-3.4kHz, e.g, the bandwidth of PCM. After digitisation of voice transmission, speech coding techniques were

developed in order to compress speech signals to lower bit rates than that in PCM, without degrading speech quality. Therefore, the bandwidth available for voice transmission was dictated by both the development of cellular systems and speech coding techniques. However, the frequency content of speech signals range from 60Hz to 20kHz. The limitations upon the telephone bandwidth thus result in a lack of distinctive spectral properties of speech sounds.

Some fricatives such as /f/ and /s/ differ in the location of the lowest spectral peak, which occur typically around 2.5kHz and 4kHz for a male adult speaker respectively. Such distinctive properties are lost in typical telephone speech, which often causes troubles to the listener in distinguishing between different fricative sounds [15, 1.1.3.1]. Some plosives (/t/ and /d/) are characterised by higher energy bursts occurring around 3.9kHz. Others (/b/, /p/) also exhibit similar energy bursts albeit with less intensity. These distinct properties are lost in NB speech, once again leading to reduced intelligibility and naturalness for plosives. Nasals are too affected. They are dominated by the first formant F1 which typically occurs around 250Hz. This is also lost due to the lower cut-off of the telephone band.

Due to the loss of such important information, the intelligibility of syllables can be degraded. During a telephone call, sometimes high frequency sounds such as /f/ and /s/ (or /p/ and /t/) are thus difficult to differentiate and, in the absence of informative contextual information, must be spelled out for effective communications. Improvements in intelligibility are thus necessary to reduce the listening effort, in order to provide comfortable communications [16].

The perceived quality as well as the intelligibility of speech signals increases with increases in acoustic bandwidth, particularly for unvoiced sounds. This is because they contain a substantial portion of their spectral content above 3.4kHz. A transition from NB to WB communication then naturally leads to an increase in syllable and sentence intelligibility from 90% and 99.3% to 98% and 99.9% respectively [16, 17]. The speech quality is also improved by 1.42 MOS points [17]. The quality of communication is further improved at super-wide bandwidths.

Even though intelligibility relates to the recognisability of speech sounds, speech quality measures are mainly used to evaluate the performance of speech codecs in transmission systems. Similarly, ABE algorithms are also evaluated in terms of speech quality by using either subjective or objective measures.

1.3 Speech coding

Speech coding or *speech compression* algorithms have the goal of estimating the compact digital representations of analogue voice signals for their efficient storage and transmission [18]. In the 1990s, the increased number of mobile phones and the increase in demands to mobile communications brought new challenges for digital speech transmission systems, especially regarding the limitations of bandwidth and its implications upon PCM speech quality [13, Chapter 1].

The digital transmission of speech involves the use of codecs in order to convert analogue signals into digital format. A codec consists of an *encoder* at the transmitter end that converts an analogue signal into compressed, digital bits. These bits can be transmitted over digital landlines or wireless networks. At the receiver, the digital bitstream is converted back to an analogue signal using a *decoder*. Codecs are used in telephones, cellular networks, televisions, set-top boxes and TV transmitters and receivers [8, page 18]. The purpose of *speech coding* is to compress data so that speech transmission can be performed at lower bit rates while maintaining high speech quality [19, Section 3.5]. Speech codecs should thus operate at low bit rates with low complexity and limited delay; these requirements are especially important in mobile communications considering the limited number of radio network resources and requirement for low power consumption in mobile, battery powered devices [14, Section 8.1], [19, Section 3.5].

Various speech coding standards and speech codecs have proposed in the past decades. They can be categorised in terms of their operational bandwidths.

1.3.1 Narrowband coding

NB coding techniques compress speech signals to between 0.3-3.4kHz.

Pulse-code modulation (PCM)

PCM [20] is a simple *waveform*[8] based coding method that performs discrete-time, discrete-amplitude approximation of analog signals in the time domain. Signals are sampled at a sampling rate of 8kHz and quantised using non-uniform quantisation levels at 64kbps. PCM supports narrowband (0.3-3.4kHz) communication that is compatible with older analog telephone systems. The transmission characteristics of PCM for voice signals are standardised in ITU-Recs. G.711 [21] and G.712 [22]. PCM coding is widely used in PSTNs and mobile networks.

For economical and complexity reasons, higher bit rates (e.g., 64kbps employed with PCM) are reduced to between 4 and 32kbps in radio cellular systems such as cordless phones and cellular radio networks while maintaining similar narrowband speech quality of PCM. To achieve this, most coding algorithms are based on the source-filter model of speech production and also exploit properties of the human auditory system [14, Chapter 8].

An extension of PCM, known as adaptive differential PCM (ADPCM), standardised in ITU-T Rec. G.726 [23], supports multiple bit rates of 16, 24, 32 and 40kbps. In ADPCM quantisation of residual error signal instead of the speech waveform itself is performed [24, Chapter 10]. ADPCM is typically used in cordless phones.

GSM full rate (FR) and enhanced full rate (EFR) codecs

The regular pulse excitation with long-term prediction (RPE-LTP) codec, also referred to as the GSM full rate (GSM-FR) codec, is based on linear predictive coding (LPC) which uses short-term LP analysis for spectral envelope modelling and long-term LP analysis to obtain residual error signal. The error signal is then quantised using ADPCM [24, Section 10.2]. The scheme operates at 13kbps and was adopted by ETSI in the GSM 6.10 standard (ETSI Rec. GSM 06.10 [25]) in

[8] *Waveform coding* methods aim to achieve lower bit rates through quantisation of either speech signal itself or residual error obtained via linear prediction (LP) analysis (e.g, PCM, ADPCM). *Parametric coders* (also called *vocoders*), e.g., linear predictive coding (LPC) vocoder, encode a set of model parameters instead of the time domain waveforms. These parameters represent the vocal tract system configuration, e.g., LP coefficients. While sufficient intelligibility is achieved at lower bit rates, produced speech still sounds synthetic. *Hybrid methods* exploit advantages of both the schemes; coefficients of the synthesis filter are transmitted as side information whereas quantisation of LP residual error signal is performed, e.g., code-excited linear prediction (CELP) coding [14, Section 8.1]

1992 for digital mobile radio communications.

GSM-FR was further improved using an efficient vector quantisation technique for residual error signal using algebraic code-excited LP (ACELP) algorithm [26,27]. It was standardised as the GSM enhanced full rate (GSM-EFR) codec in ETSI Rec. GSM 06.60 [28] in 1996 and operates at 12.2kbps. The GSM-EFR codec achieved speech quality equivalent to that of ADPCM at 32kbps [14, Section 8.5.3.3].

G.729

The codec standardised in ITU-T Rec. G.729 [29] in 1995 operates at 8kbps and is based on so-called conjugate-structure ACELP (CS-ACELP). It is widely used in VOIP infrastructures.

Adaptive mulit-rate (AMR)

An extension of the EFR codec with eight possible bit rates ranging from 4.75 to 12.2 kbps, known as the AMR codec, was standardised for GSM (2G) and UMTS networks (3G) by ETSI (ETSI Rec. GSM 06.90 [30]). The quality at the highest bit rate is equivalent to that of the EFR codec. 3GPP adopted AMR as the default speech codec for 3G wideband systems such as UMTS and CDMA2000 (3GPP TS 26.090 [31]). AMR coding involves the transcoding of AMR-coded speech signals to/from PCM format [32]. Signals are thus encoded and decoded twice in succession which leads to added complexity and reduced speech quality.

1.3.2 Wideband coding

In order to support wideband transmission in telephone networks (at least) new terminals need to include front-ends with better electro-acoustic, improved analogue to digital (A/D) converters and new speech codecs. Cellular radio networks also need expensive base station modifications [14, Section 10.1]. As WB transmission improves the quality of voice transmission, there is increasing demand for WB communication services in fixed and mobile networks at lower bits rates. WB coding aims at voice communication at broader bandwidths of 0.05-7kHz.

The first WB speech codec for ISDN and teleconferencing was standardised

in 1985 by CCITT [14, Section 10.1]. The ITU-T Rec. G.722 [33] specifies the characteristics of an audio WB coding system in which a frequency band is split into two, lower and higher, sub-bands such that both are encoded using ADPCM (the technique referred to as sub-band ADPCM (SB-ADPCM)). The codec supports bit rates of 4, 56 and 64 kbps and a fall-back capability to NB standards. The G.722 standard is usually used as a reference for the evaluation of other codecs [24, Appendix C]. In 1999, a low-complexity WB codec was introduced in ITU-T Rec G.722.1 [34] for hands-free applications. It achieved comparable speech quality at reduced bit rates of 24 and 32 kbps.

Adaptive multi-rate wideband (AMR-WB)

The breakthrough in WB speech coding and quality was brought by a wideband extension to the AMR codec, referred to as AMR-WB, which encodes speech within bandwidth of 0.05-7kHz. The AMR-WB codec was first standardised in 2001 by 3GPP (3GPP TS 26.190 [35]) for 3G cellular networks. It is based on the ACELP technique and employs artificial bandwidth extension for signal resynthesis beyond 6.4kHz. It is also adopted by ITU-T for WB speech coding and is specified in ITU-T Rec. G.722.2 [36]. It supports nine bit rates ranging from 6.6 to 23.85 kbps. The AMR-WB codec operating at 8.85 kbps achieves higher speech quality than AMR at 12.2kbps [37].

Voice transmission by AMR-WB provides significantly better quality than NB telephony due to the increased bandwidth. Therefore, conversations are more natural, thereby improving the user experience. AMR-WB technology is generally referred to as *high-definition (HD) voice*. HD voice also improves hearing in noisy environment. By May, 2016, 164 mobile operators (17 on GSM (2G), 130 on UMTS (3G) and 63 on LTE (4G) networks)[9] have launched commercial HD voice services in 88 countries [32]. HD voice services work best when two HD mobile phones are in communication over a HD-voice compatible network. Improved acoustic properties and noise reduction capabilities of the most recent HD smart phones also improve call quality.

The transmission of WB voice over GSM or UMTS networks needs tandem-free (TFO) or transcoder-free operation (TrFO). In TFO, the coded WB parameters are transmitted within the PCM bitstream to achieve WB speech quality. However,

[9]Some operators offer HD voice service on more than one network.

this requires a bit rate of 64kbps. In TrFO, double encoding can be avoided when the end-to-end transmission links both employ the same type of codec [13, Section 3.4]. Therefore, the combination of TFO and TrFO operations makes WB calls possible between all types of network.

The transmission of voice packets over LTE networks is referred to as voice over LTE (VoLTE). LTE systems, being all-IP and optimised specifically for data transfer, do not support circuit switching which is needed for voice and SMS services. Voice calls are thus handled using circuit switched fall back (CSFB) when the data connection "falls back" to 2G or 3G network connection before call initiation [32].

G.729.1

ITU-T Rec. G.729.1 [38] defines an extension to the G.729 codec providing for the scalable narrowband and wideband coding of speech and audio signals from 8-32kbps. G729.1 is the first layered[10] codec that is designed with an embedded scalable structure in order to extend the functionalities of the existing G.729 standard [39, Section 4.2.1.1].

1.3.3 Super-wideband or full band coding

SWB or FB voice communications, also referred to as *full* HD voice services, transmit almost the entire human voice spectrum which makes conversations much more natural and understandable than in the case of NB or WB communications. Full HD voice thus improves the call experience beyond that obtained by HD voice services, bringing quality closer to what is achieved in face-to-face conversations.

[10]Before G.729.1, the WB codecs did not extend the operation of existing NB codecs. Such extension requires, (i) detection of bandwidth of an incoming signal before encoding and (ii) the interoperability among various mobile networks and devices (which support different coding standards). *Layered* extension of existing codecs provides these features. In layered coding, the core layer of a codec is overlaid with multiple enhancement layers, e.g, the core layer is designed for NB coding whereas the enhancement layers provide more improvements at the cost of additional bit rate. Such scheme provides the features of bandwidth scalability [39, Section 4.1.2].

G.729.1 Annex E

G.729.1 Annex E [40] extends the 32kbps mode of the G.729 codec to super-wideband mode providing bit rates in the range of 36-64kbps. The codec uses MDCT coding. A SWB extension to the scalable WB codec G.729.1 proposed in [41] achieves improved audio quality (especially for music signals) in comparison to the existing SWB extension G.722.1 Annex E with $\approx 18\%$ reduction in bitrate.

Extended AMR-WB (AMR-WB+)

The AMR-WB+ standard (3GPP TS 26.290 [42]) is an super-wide extension to the AMR-WB codec that operates up to an increased frequency range of 16kHz and bit rates up to 32kbps [14, Section 10.1]. It is a hybrid codec that combines linear predictive and transform coding techniques depending on the signal type, e.g., speech or audio. AMR-WB+ provides high quality for audio or music signals while meeting the strict requirements for multimedia codec bit rates and complexity [43].

G.719

A low-complexity coding algorithm for full-band speech and audio signals is described in ITU-T Rec. G.719 [44]. The coding technique offers bitrates from 32 up to 128 kbps.

HE-AAC

The high-efficiency advanced audio codec (HE-AAC) uses a so-called spectral band replication (SBR) [45, 46] approach for efficient coding of audio signals. HE-AAC was developed by the International Organization for Standardization/International Electrotechnical Commission (ISO/IEC) Moving Picture Experts Group (MPEG) by subsequent extension of the established Advanced Audio Coding (AAC) architecture [47]. The SBR technique exploits properties of the human auditory system – that the higher frequencies in audio spectrum contribute marginally to perception – and it is mainly focused on generic audio (speech and music) signals. The audio content above 8kHz can thus be encoded efficiently to achieve higher compression rates. HE-AAC is employed in mobiles, internet streaming, TV digital radio and TV broadcasting mainly for music and audio content [47].

Over the top (OTT) conversational codecs

Over the top (OTT) service providers (e.g., Skype) provide point-to-point services for VoIP. The use of proprietary codecs such as SiLK allows for conventional NB voice services to be shifted towards WB and SWB communications via the use of broadband IP services [48].

Opus[11] is another high quality codec which combines technologies from Skype's SilK and Xiph.Org's CELT[12] audio codecs, providing bit rates of 6 kbps (for NB mono speech) to 510 kbps (for high quality stereo music). It was standardised by the Internet Engineering task force (IETF). It is used for interactive speech and music transmission (supporting audio bandwidth from NB to FB) over the Internet. Opus performs hybrid coding which involves the coding of frequencies up to 8kHz using the SiLK codec. Frequencies above 8kHz are coded using CELT [39, 4.2.2.1]. Opus provides SWB or FB transmission at, and above 24 kbps. Unfortunately, however, the bit rate is too high for efficient use of radio resources for speech and audio in mobile systems [48].

Enhanced voice services

3GPP carried out a preliminary study [49] in 2010 in order to investigate use-cases of enhanced voice services (EVS) over the packet system of LTE networks. The study further led to the standardisation of the EVS codec in 2014. EVS is the first conversational (low delay) codec that can encode speech as well as other audio signals with a SWB (0.05-14kHz) at bit rates as low as 9.6kbps [39, 4.2.3]. It operates at four different bandwidths, namely, NB (0.02-4kHz), WB (0.05-7kHz), SWB (0.05-16kHz) and FB (0.05-20kHz)[13]. The key features of EVS include: (i) enhanced quality for mixed signal content such as speech as well as music leading to improved user experience via features such as *in-call music*; (ii) improved coding efficiency for NB and WB communications thereby providing better quality than existing AMR and AMR-WB codecs at similar bit rates; (iii) backward compatibility to the AMR and AMR-WB codecs; (iv) improved robustness to packet loss, frame erasures and jitter [48]. EVS supports twelve bit rates ranging from 5.9 to 128 kbps with SWB and FB services starting at or above 9.6 and 16.4kbps respectively [48].

[11] http://www.opus-codec.org/

[12] Constrained Energy Lapped Transform

[13] The quality offered by the use of these bandwidths is equivalent to that of NB telephones, AM, FM and compact discs (CD) respectively.

EVS combines LPC for speech and modified discrete cosine transform (MDCT) based coding for audio signals. It switches automatically between these two coding modes in real time depending upon the type of signal. Subjective listening tests have shown that EVS outperforms all existing conversational voice and audio codecs across all bit rates and bandwidths [50]. Detailed technical details of EVS can be found in [51, 52, 53, 54].

Due to the key features that the EVS codec provides, mobile operators have started enabling their networks for EVS support. As of September 2018, 17 mobile operators have already introduced EVS services in their networks; 153 EVS-enabled mobile devices from 12 different vendors are available in the market [55]. EVS services are also marketed as "*HD voice plus*" or "*ultra HD voice*". While deployment of EVS codec enabled devices and networks is speeding up, there is a long way to go before it becomes a ubiquitous technology [55].

1.4 Artificial bandwidth extension

Artificial bandwidth extension algorithms have been developed to improve NB speech quality by estimating missing highband (HB) components at 3.4-8kHz from available NB components. ABE is based on the assumption that spectral content in NB and HB are correlated as the entire speech spectrum is generated by the same physical and acoustical configuration of the human speech production system. ABE can be performed with or without the use of an additional side information for reconstruction of HB components and thus categorised into *blind* and *non-blind* methods.

1.4.1 Non-blind methods

HB frequency components are highly correlated with NB components but contain relatively little information [39, Section 4.1.4.5]. The spectral content at higher frequencies thus can be represented with fewer bits in comparison to NB frequency components. Non-blind ABE methods thus recover missing high frequency components at the receiving-end (or near-end) from auxiliary side information related to higher frequencies which is encoded into a data stream together with NB components. However, the inclusion of such side information typically incurs an additional burden of 1-5 kbps [56]. Non-blind approaches are codec specific and

require a matching decoder in order to recover missing frequency components.

Most speech coding techniques, therefore, usually perform non-blind ABE in order to achieve lower bit rates while maintaining speech quality. Notable examples are Qualcomm's enhanced AMR (eAMR) codec [57], the HE-AAC codec [45, 46], and the AMR-WB codec (in 23.85 kbps mode) [31].

1.4.2 Blind methods

In contrast, blind ABE methods estimate missing HB components using only the available NB components. Such ABE[14] solutions thus exploit the correlation between NB and HB components of speech and estimate missing HB components using a regression model learned from WB speech training data. ABE algorithms thus mainly focus on the better modelling of correlation via improved regression models. In contrast to non-blind alternatives, blind methods do not incur any additional bit-rate burden and are codec-neutral.

While non-blind methods (typically employed in speech coders) provide better WB speech quality (which is quite obvious because the HB information is reconstructed via some side-information), blind methods provide an alternative to WB speech coding where WB speech at the receiver is reconstructed using the input NB speech only. WB services can thus be provided independently of networks and codecs used in mobile devices.

1.4.3 Motivation and applications

This section describes the applications of ABE in different scenarios.

When network and/or mobile terminals do not support WB communication:

As discussed in the previous section, in order to improve speech quality offered by traditional telephony infrastructure, speech signals should be transmitted at higher bandwidths. This requirement has led to the development of coding techniques

[14]The term ABE refers to blind ABE unless mentioned explicitly throughout the remainder of this thesis.

to compress information at higher bandwidths (Section 1.3). Calls at higher-bandwidths are possible only if the entire communication path supports operations at the same bandwidths, e.g. the WB call at the receiving terminal (or near-end) is possible only if the mobile device at the transmitting terminal (or far-end) and the network both support WB communication. Lack of either leads to a reduction in bandwidth and thereby a reduction in speech quality.

In today's scenario, telecommunications involve a combination or interconnection of different networks and mobile devices supporting NB, WB and SWB communications[15]. This is because the entire processing chain that exists between speech codecs and the network terminals requires a complete redesign to support higher bandwidth communications [58]. While deployment of WB codecs and networks is in progress, it is slow as it incurs costs to the network operators as well as end users. Additionally, a phone call may involve a landline device which restricts in the bandwidth to NB by default. Therefore, even today, a significant portion of calls operate in NB mode whereas the migration to WB will take considerable time [59]. NB and WB networks and devices (or terminals) will thus coexist for some years to come, leading to mobile phone calls of different bandwidths at the receiving terminal. An illustration of hybrid NB and WB phone calls is shown in Fig 1.3. The scenarios illustrated in Fig. 1(b), Fig. 1(c) and Fig. 1(d) exploit the potential of ABE to improve speech quality from NB to WB, provided the receiving terminal supports ABE functionality [60].

When a WB-to-NB handover occurs during a phone call:

Due to the presence of heterogeneous communication networks [61], the process of bandwidth switching from WB to NB may occur during an ongoing phone call, especially when the user is moving (e.g., in train, bus or car). This may happen either due to handovers between two different networks (e.g. when the user enters a network cell supporting NB communication from a WB network) or due to decreases in network resources that causes dynamic fall back from WB to NB mode [62,63]. This can lead to abrupt changes in quality and thus an annoying user experience. According to [60], a WB-to-NB handover leads to perceived speech quality even below NB.

A possible solution to avoid this problem is to switch to WB communication

[15]Typically the NB, WB and SWB calls involve the AMR, AMR-WB and EVS codecs.

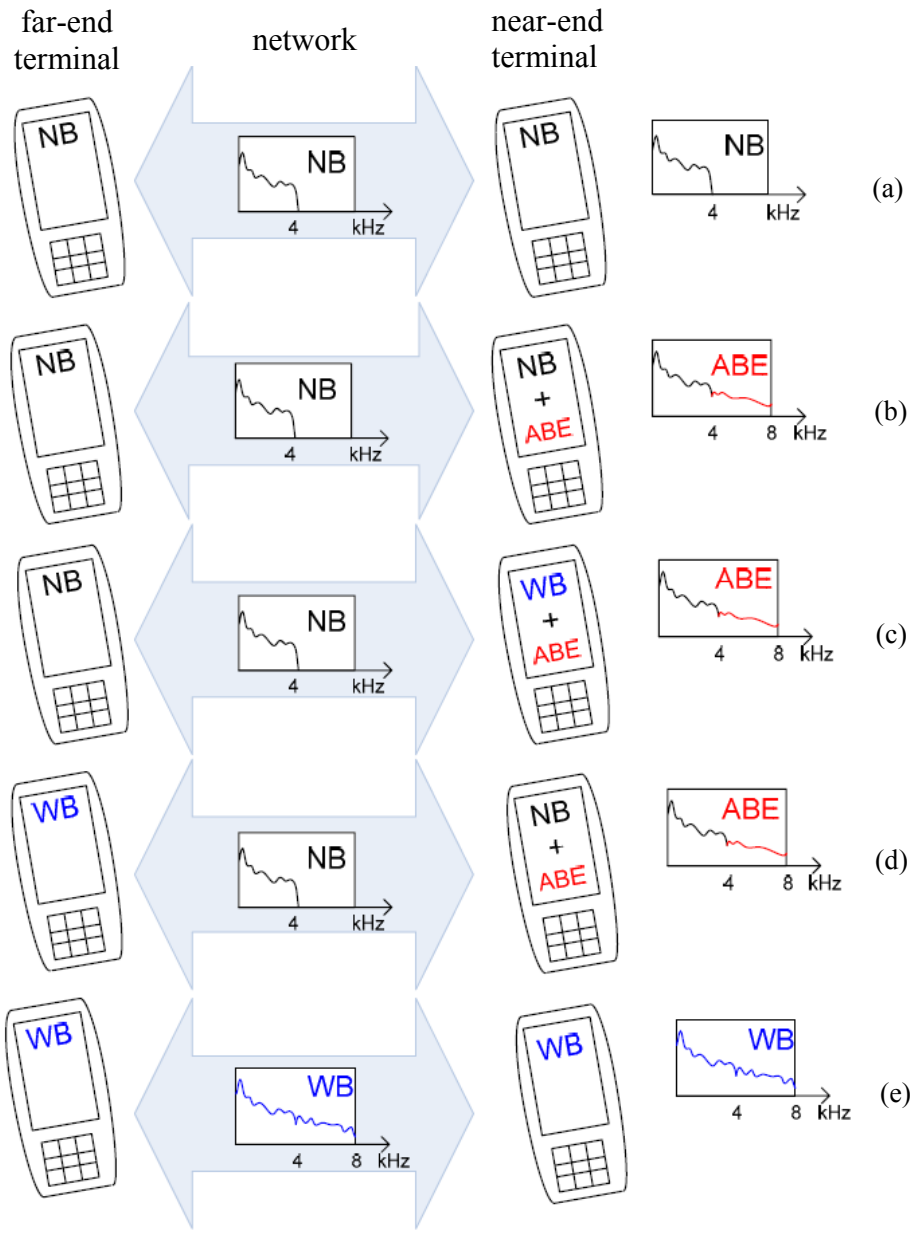


Figure 1.3: An illustration of phone calls at different bandwidths at receiving mobile terminal (adapted from [60]). A NB far-end terminal transmits a NB signal through a NB network and the near-end-user receives (a) NB speech through a NB terminal, (b) artificially bandwidth-extended speech through a NB terminal (with ABE), (c) artificially bandwidth-extended speech through a WB terminal (with ABE). A WB far-end terminal transmits speech in NB if either the network is NB or the receiver is a NB terminal; the user then receives (d) artificially bandwidth-extended speech if the terminal includes ABE. WB transmission is achieved only when (e) both the terminals and the network support WB.

as soon as the network resources are increased. However, subjective tests have indicated that the switching from NB to WB speech is perceived as an impairment unless the transition happens early enough or at the beginning of the call [63]. According to [62], WB transmission should continue at least for 30 seconds after switching in order to benefit from improved speech quality. Therefore, instead of switching from NB to WB (which also includes waiting time until the network resources are reallocated to support WB communication), ABE can be used as soon as the call falls back to NB mode thereby mitigating the need for a true WB call. A comparison between the subjective quality of two switching schemes, namely transitions between WB (AMR-WB coded) and NB (AMR coded) speech and transitions between WB and bandwidth-extended NB (AMR coded) speech is reported in 3GPP TR 26.976 [64]. The case with bandwidth extension was mainly preferred in three different noise conditions (clean, street noise, car noise).

When there is a bandwidth-mismatch between training and testing data:

For certain applications such as speaker recognition, large amounts of NB or telephone speech data is available for model training. In order to operate upon NB data, the conventional approach is to perform downsampling from WB to NB. However, this leads to the loss of useful spectral content in WB speech. ABE can thus be used to reduce the bandwidth mismatch between speech data recorded at different sampling rates. NB speech data can be artificially bandwidth-extended and then used with available WB data. This is helpful in two aspects. First, the amount of WB training data can be increased (using bandwidth-extended NB data) for the training of WB models. Second, already trained WB models can still be used (with the application of ABE) in the case that test data is NB; re-training of the models with NB data is no longer needed.

The use of ABE thus allows for only one model to be trained while still supporting different bandwidths modes. ABE has been investigated for applications such as speaker recognition [65], automatic speech recognition (ASR) [66, 67, 68], speaker identification [69] or speaker verification [70] in order to improve the performance of WB models by increasing the amount of WB training data via ABE.

When users have hearing impairments

People with impaired hearing use hearing-aids or cochlear implants (CIs) and often face difficulties during telephone calls due to bandwidth limitations. ABE can be used to improve intelligibility and quality of NB speech for such users [71]. The study reported in [72] showed that users with hearing impairments can tolerate overestimated energies in the bandwidth-extended speech sounds, thereby providing increased intelligibility.

1.5 Super-wide bandwidth extension

With the progress in recent times in super-wideband or full-band speech coding techniques and the introduction of the EVS codec (Section 1.3), many smart devices and networks now support high-quality speech communication services at super-wide bandwidths. However, in today's heterogeneous networks, SWB devices are often used with other devices and networks which support only narrowband (NB) or wideband (WB) communications. While they usually offer backward compatibility, users of SWB devices will then be restricted to NB or WB communications. Super-wide bandwidth extension (SWBE) thus aims to recover high-frequency (HF) components between 8 and 16kHz in order to improve the quality of WB speech signals.

Similar to ABE, which helps to improve speech quality via NB-to-WB extension (illustrated in Fig 1.3), SWBE[16] has the goal of improving the gap in quality between WB and SWB communications.

1.6 Contributions

ABE performance can be improved either by developing a better regression model, by designing more compact, informative front-end features, or through the combination of both approaches. One approach to improve feature extraction is to exploit *dynamic* (temporal as well as spectral) properties of speech signals in the form of features extracted from neighbouring speech frames. This is in addition

[16]While approaches to NB-to-WB and WB-to-SWB extension both are a form of bandwidth extension, the former is referred to as ABE and the later is referred to as SWBE for simplicity throughout this thesis work.

to the use of conventional *static* features. Contextual information captured via front-end features can be incorporated via the concatenation of instantaneous static features with either static or dynamic *delta* features obtained from neighbouring speech frames. This is referred to as front-end *memory inclusion*. Another solution is to model such temporal dependencies via statistical modelling techniques such as hidden Markov models (HMMs), recurrent neural networks (RNNs) or long short-term memory (LSTM) networks. The use or modelling of *contextual information* or *memory* helps to model interframe dependencies or the dynamics of speech signals, a technique common to many speech processing applications, including ABE.

Memory inclusion has two drawbacks. First, the inclusion of memory produces higher-dimensional features. Traditional statistical models such as GMMs and HMMs tend not to handle high-dimensional data efficiently. This is because the number of parameters and the number of training samples needed for reliable density estimation grows exponentially. This problem is referred to as the *curse of dimensionality*. Deep neural network (DNN) based approaches also require deeper networks and longer convergence time if used with higher-dimensional input features. In other words, the inclusion of memory leads to regression models of increased complexity. Second, the use of memory involves the extraction of contextual information from future (or look-ahead) frames which increases the latency or delay of an ABE algorithm.

The ABE literature addresses the importance of front-end memory inclusion in the form of *delta features* under the constraint of fixed dimensionality. The inclusion of memory has been reported previously, even without affecting the complexity of a standard regression model. This is achieved by replacing higher order static NB and/or HB feature coefficients with lower order dynamic delta coefficients. The use of memory has been studied and investigated from an information theoretic perspective. However, delta coefficients are non-invertible and are discarded during HB reconstruction, resulting in the loss of information and practically suboptimal performance.

While past work points towards the importance of memory to ABE, it raises the questions of what degree of contextual information is of benefit to ABE and how it can best be harnessed without increasing the latency and computational complexity of a standard regression model. The work reported in this thesis addresses these questions; a quantitative analysis via the mutual information (MI) measure is presented which compares the benefit of utilising memory in an otherwise fixed ABE

algorithm. Contextual information captured in the form of static features extracted from neighbouring speech frames is referred to as *explicit memory* throughout this thesis work. The benefit is further confirmed via improvements in ABE performance without affecting the complexity of the regression model. Principal component analysis (PCA) is employed as a dimensionality reduction technique. The topic of ABE is further presented as a feature extraction problem where higher-level, compact NB features are learned from higher dimensional log-spectral data resulting from the inclusion of explicit memory. The use of deep learning architectures such as semi-supervised stacked auto-encoders (SSAEs) and conditional variational auto-encoders (CVAEs) for dimensionality reduction (or feature extraction) is proposed. A comparison of different dimensionality reduction techniques such as principal component analysis (PCA), conventional stacked auto-encoders (SAE), SSAEs, variational auto-encoders (VAEs) and CVAEs is reported. The results show that some form of supervision is important to the optimisation of dimensionality reduction techniques to ABE.

The contributions[17] of the research work in this thesis are divided into two parts. The first part, which forms a significant portion of the work, reports contributions in NB-to-WB extension. The second part reports an approach to WB-to-SWB extension. The proposed approach is based on linear prediction-based analysis synthesis and performs SWBE without statistical estimation of missing frequency components. The following presents a summary of the key contributions:

Explicit memory inclusion under the constraint of fixed dimensionality

This contribution relates to a comparative, quantitative analysis of explicit memory obtained via static features extracted from neighbouring speech frames. Mutual information (MI) is used as a standard information theoretic measure to show the benefit of memory inclusion. Three different front-end features are investigated. Considering practical requirements of ABE solutions, inclusion of explicit memory is performed without significant increases to latency or computational complexity. Specifically, log-Mel filter (logMFE) and linear prediction (LP) coefficients are

[17] In order to support reproducible research, the implementations of all the proposed ABE approaches reported in this thesis are publicly available at: `https://github.com/bachhavpramod/bandwidth_extension`. The speech samples produced during this research work are also available at: `http://audio.eurecom.fr/content/media`.

used as front-end narrowband and highband features respectively; principal component analysis (PCA) is used as a dimensionality reduction transform. In order to highlight the improvements obtained specifically from the modelling of explicit memory under the fixed dimensionality constraint, conventional GMM regression mapping is used as a regression model. Finally, the findings are validated through objective and subjective assessments of an ABE system which uses memory with only negligible increases to latency and computational complexity.

Part of this work was published in:

1. **P. Bachhav**, M. Todisco, and N. Evans, "Exploiting explicit memory inclusion for artificial bandwidth extension," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5459–63, 2018, Calgary, Canada.

Memory inclusion with semi-supervised stacked auto-encoders

This contribution relates to the application of semi-supervised stacked auto-encoders (SSAEs) to ABE for non-linear dimensionality reduction. As an unsupervised, linear approach to dimensionality reduction, PCA aims only to produce a low dimensional representation which retains as much as possible the variation in the input representation. The hypothesis is that supervised or semi-supervised and non-linear dimensionality reduction techniques offer potential to learn lower dimensional representations tailored specifically to ABE, thereby giving better performance. Thus, the ability of SSAEs to learn higher-level representations is exploited to learn compact narrowband features. It is shown that the low-dimensional representations learned from log power spectral (LPS) coefficients lead to discernible improvements to speech quality in comparison to those learned from logMFE features. Features extracted directly from log-spectra can be used by a standard regression model without augmenting complexity. The benefit of compact NB features learned via SSAE is confirmed by an information theoretic analysis. The merit of the approach is further demonstrated with different objective metrics and is confirmed by the findings of informal listening tests.

Part of this work was published in:

2. **P. Bachhav**, M. Todisco, and N. Evans,"Artificial bandwidth extension with Memory inclusion using semi-supervised stacked auto-encoders," in *Proc. INTERSPEECH*,pp. 1185–89, 2018 Hyderabad, India.

Latent representation learning using conditional variational auto-encoders

This contribution relates to the first application of conditional variational auto-encoders (CVAEs) for supervised dimensionality reduction specifically tailored to ABE. CVAEs, a form of directed, graphical model, are used to model higher-dimensional log-spectral data to extract latent narrowband representations. The idea in this work is that the conditioning variable of a CVAE can be optimised via an auxiliary neural network in order to learn higher-level NB features, features that are *tailored* to improve the estimation of missing HB components. Reported is an approach to combine CVAEs with a probabilistic encoder in the form of an auxiliary neural network which derives the conditioning variable. A technique for their joint optimisation is presented. Objective and subjective assessments are reported to show that the probabilistic latent representations learned with CVAEs produce bandwidth-extended speech signals of notably better quality when compared to that obtained with alternative dimensionality reduction techniques.

Part of this work was published in:

3. **P. Bachhav**, M. Todisco, and N. Evans, "Latent representation learning for artificial bandwidth extension using a conditional variational auto-encoder," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7010–7014, 2019, Brighton, United Kingdom.

Efficient approach to SWBE using linear prediction analysis-synthesis

This contribution relates to an efficient approach to SWBE which avoids the use of statistical models for estimation of missing higher frequencies, thereby reducing the complexity. The algorithm is based upon a classical source filter model in which spectral envelope and residual error information are extracted from a WB signal using conventional linear prediction analysis. A form of spectral mirroring is

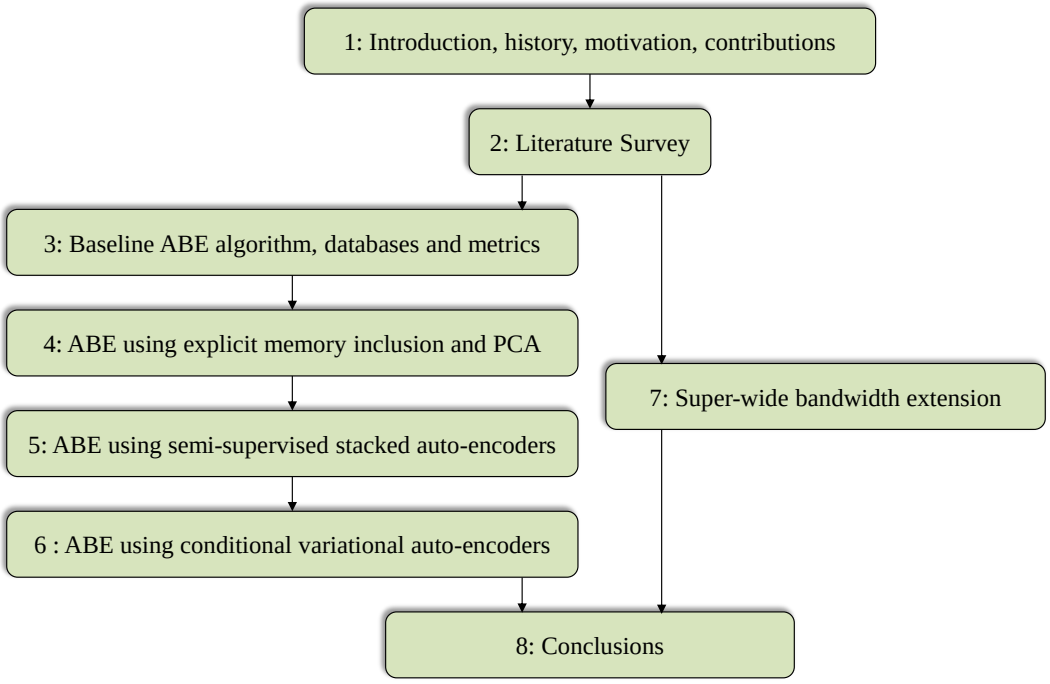


Figure 1.4: Outline of the thesis and connections among various chapters.

then used to extend the residual error component before an extended SWB signal is derived from its combination with the original WB envelope. Improvements to speech quality are confirmed with both objective and subjective assessments. It is demonstrated that the quality of SWB speech, derived from the bandwidth extension of wideband speech, is comparable to that of speech processed with the standard EVS codec with a bitrate of 13.2kbps. In addition, consistent improvements in quality over WB speech processed with the AMR-WB codec with a bitrate of 12.65kbps are reported. Without the need for statistical estimation of missing super-wideband components, the proposed algorithm is highly efficient and introduces only negligible latency.

Part of this work was published in:

4. **P. Bachhav**, M. Todisco, and N. Evans, "Efficient super-wide bandwidth extension using linear prediction based analysis synthesis," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5429–5433, 2018, Calgary, Canada.

Other contributions

Some research work carried out during this PhD is not included in this thesis. This work resulted in the following publication:

5. **P. Bachhav**, M. Todisco, and N. Evans, "Artificial bandwidth extension using the constant Q transform," *in Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5550–5554, New Orleans, USA.

Other publications prior to the undertaking of this PhD include:

6. N. Shah, **P. Bachhav**, and H. Patil, "A novel filtering-based F0 estimation algorithm with an application to voice conversion," in *Proc. IEEE Asia-Pacific Signal and Information Processing Association (APSIPA) Annual Summit and Conference*, 2017, Kuala Lumpur, Malaysia.

7. **P. Bachhav** and H. Patil, "A novel filter bank for epoch estimation," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2017, Kos island, Greece.

8. **P. Bachhav**, H. Patil and T. Patel, "A novel filtering based approach for epoch extraction", in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, Brisbane, Australia.

1.7 Outline of the thesis

An outline of the thesis is shown in Fig. 1.4. An extensive survey of prior works related to artificial bandwidth extension is presented in Chapter 2. Chapter 3 describes a memoryless baseline ABE algorithm which is used to evaluate the benefit to ABE performance of explicit memory inclusion. The databases (used for training, validation and testing of ABE approaches), objective and subjective assessment metrics including the mutual information measure are also discussed. Chapters 4, 5, 6 and 7 report the novel contributions outlined above. Conclusions and, directions for future work are presented in Chapter 8.

Chapter 2

Literature survey

Artificial bandwidth extension algorithms (ABE) aim to estimate missing higher frequency components at 3.4-8kHz. In this chapter, a survey of approaches to ABE is presented. ABE algorithms are mainly divided into four categories: non-model based approaches (Section 2.1), approaches based on the classical source-filter model (Section 2.2), approaches based on direct modelling of spectra (Section 2.3) and end-to-end ABE approaches (Section 2.4). While most approaches use regression models that are trained via standard mean square error (MSE) optimisation criterion, some approaches use perceptually motivated cost functions (Section 2.5). Some approaches focus on modelling of contextual or temporal information to improve ABE performance (Section 2.6). While ABE approaches focus on NB-to-WB extension, SWBE approaches for WB-to-SWB extension are developed to bridge the quality gap between WB and SWB communication (Section 2.8).

2.1 Non-model based ABE approaches

The *non-model based* ABE approaches do not use a priori knowledge about speech production mechanism and employ simple operations to introduce missing frequency components. The operations include spectral translation or shifting (fixed or adaptive), generation of high frequency components via non-linear operations on time-domain signals, bandpass filtering on white noise etc. Some approaches also perform spectral shaping of the generated frequency components using some gain control parameter or an empirically determined filter. The first commercial use of such ABE methods by the British Broadcasting Corporation (BBC) [73] is reported in 1972 where the acoustical bandwidth of telephone speech in broadcast

programmes is improved. The notable examples are [74, 75, 76, 77].

Such methods, however, produce extended WB speech signals with audible processing artefacts and distortions. This is because energy of the generated frequency components is either too weak or too strong compared to the NB component [78, Section 1.3]. Additionally, quality of extended speech signals depends upon the effective bandwidth of the original input signal, e.g., such methods surprisingly work well for SWBE (WB-to-SWB extension) when the input signal is WB but perform poorly when the input signal is NB [79, Section 5.4.1]. Readers are encouraged to refer to [79, Section 5.4] and [78, Section 1.3], [15, Section 2.2] for more details on non-model based algorithms.

2.2 ABE approaches based on source-filter model

The *model-based* ABE algorithms use a priori knowledge about characteristics of speech signals and the human speech production mechanism. Since beginning of the nineties, most ABE algorithms started exploiting the classical source-filter model [80] of speech production where a NB speech signal is represented by an excitation source and a vocal tract filter. The frequency content of these two components can be extended through independent processing before a WB signal is resynthesised. The extension of NB speech is thus divided into two tasks; (1) estimation of HB or WB spectral envelope from input NB features via some form of an estimation technique and (2) generation of HB or WB excitation components via some form of time-domain non-linear processing, spectral translation or spectral shifting methods. The HB component is usually parametrised with some form of linear prediction (LP) coefficients whereas the NB component is parametrised by a variety of static and/or dynamic features.

2.2.1 Extension of spectral envelope

In practice most approaches focus on the extension of the spectral envelope since it has the dominant impact on speech quality. Different techniques such as linear and codebook mapping, Gaussian mixture models (GMMs), hidden-Markov models (HMMs) and deep neural networks (DNNs) are used for estimation.

The use of many different feature representations for NB and HB components

has been reported, e.g. linear prediction coefficients (LPCs) [81, 82], line spectral frequencies (LSFs) [83, 84] and Mel-frequency cepstral coefficients (MFCCs) [85]. A mixed approach reported in [16] uses NB auto-correlation coefficients to estimate HB cepstral coefficients. Some additional features are also added to the NB feature set to improve estimation performance [78, 86, 87].

2.2.1.1 Codebook mapping

The very first approaches to ABE use codebook mapping [88, 89, 90, 91, 92] method for estimation. It involves training of two codebooks. A primary codebook is trained on NB feature vectors $\mathbf{x}$ via vector quantisation (VQ), e.g, using the well-known LBG algorithm [93]. This is equivalent to clustering of the training vectors into N clusters where centroids of the clusters form entries of the primary codebook. For every entry in the primary codebook, average of corresponding WB vectors $\mathbf{y}$ form the corresponding entry in the shadow codebook. During extension, for each NB speech frame, NB feature vector $\mathbf{x}$ is compared with the primary codebook entries. The closest entry is selected and corresponding entry in the shadow codebook gives the estimated WB spectral envelope. Some methods use interpolation methods to improve performance of the codebook based ABE approaches. In this case, instead of choosing one WB envelope from the shadow codebook, the weighted sum of all or most probable codebook entries is used. The use of split codebook is also reported where separate codebooks for voiced and unvoiced frames are used given that the voiced and unvoiced spectral envelopes are characterised by different shapes [94].

The performance of the codebook mapping methods depends on the sizes of the codebooks. Higher the number of entries in the codebooks, better is the estimation performance, however, at the cost of increased memory required to save the codebooks. The performance also strongly depends on the choice of features $\mathbf{x}$ and $\mathbf{y}$. More details on codebook mapping can be found in [95, Section 6.6] and [92, Section 3.1].

2.2.1.2 Linear mapping

In linear mapping based methods [94, 96, 97], the HB or WB feature vectors $\mathbf{y}$ are estimated using a linear transformation defined according to $\hat{\mathbf{y}} = \mathbf{A}^T\mathbf{x}$ where $\mathbf{A}$ is a matrix with dimensions $m \times n$, $\mathbf{x}$ and $\mathbf{y}$ are column vectors with dimensions m and n respectively. The transformation matrix $\mathbf{A}$ is obtained during training.

The relationship between NB and HB frequency components is not simply linear and thus the estimation can be further improved by introduction of a classifier before linear mapping, the method is referred to as piecewise linear mapping [94,98]. This approach helps to improved modelling of non-linear dependency between $\mathbf{x}$ and $\mathbf{y}$. During training, the feature vectors $\mathbf{x}$ are classified into a predefined number of classes (via vector quantisation or hard or soft decision schemes) and then a transformation matrix is learned between only those vectors $\mathbf{x}$ and $\mathbf{y}$ which belong to a particular class. During extension, first classification of input vectors $\mathbf{x}$ is performed and then estimated vectors $\hat{\mathbf{y}}$ are obtained using a transformation matrix corresponding to the estimated class.

2.2.1.3 GMM based mapping

In codebook mapping, the continuous acoustic space is modelled by a discrete set of codebook entries obtained via VQ. This approach thus involves hard classification of input NB feature vectors into N classes defined by the codebooks entries. The resulting extended speech signals thus comprise discontinuities and annoying artefacts. In linear mapping, the acoustic space is assumed to exhibit very simple linear relation between NB and HB components. Linear mapping thus involves an oversimplification of the ABE problem leading to inferior speech quality. Codebook and linear mapping approaches are have deterministic and quantising nature. Statistical modelling techniques, in contrast, provide a probabilistic framework to produce a continuous approximation of the complex, non-linear acoustic space [15, Section 2.3.3.4].

GMM based statistical modelling is widely used in the ABE literature. GMMs are capable of modelling acoustic space of a speaker's voice into set of underlying acoustic classes; the classes are represented by phonetic events such as vowels, fricatives, or nasals. Typically, the spectral shape (represented by some form feature representation) of the i^{th} acoustic class is modelled by the mean μ_i of the i^{th} Gaussian mixture component and the variations of the average spectral shape are modelled by the covariance matrix Σ_i [99]. From another perspective, GMMs can form smooth approximations of arbitrary continuous probability distribution functions via modelling of discrete set of Gaussian functions.

The ability of GMMs (of modelling complex data distributions) is thus exploited in oder to model joint density $p(\mathbf{x}, \mathbf{y})$ of NB and HB feature vectors obtained from

training data. The parameters of the GMM are estimated using the EM algorithm during training and are used to obtain the MMSE estimates $\hat{\mathbf{y}}$ during extension. The joint density modelling via GMMs helps to capture the correlations between NB and HB features spaces.

Drawing upon the successful use for spectral transformation in speaker conversion systems in [100, 101], the first use of GMM based mapping for ABE is reported in [81]. It is shown that the speech signals extended using GMM-based technique are of superior quality that those extended using codebook-based methods. Following their successful use, the numerous approaches [83, 85, 102, 103, 104, 105, 106] employ GMM regression (GMMR) technique[1] for ABE.

2.2.1.4 HMM based mapping

The codebook and GMM based mapping methods do not model the temporal dependencies of the speech signals and thus, the use of the first-order Hidden Markov Models (HMMs) for ABE is proposed in [16, 107, 108]. The states of the HMM are defined via centroids of the codebook which obtained by performing VQ on HB spectral envelopes extracted from training data. During training, the HMM parameters are estimated using Baum-Welch algorithm. The parameters include initial and state transition probabilities, state observation likelihoods which are modelled by continuous probability distribution functions via GMMs. Each state is thus associated with a GMM. During extension, the sequences of input NB feature vectors are decoded using the pre-trained HMM to obtain the posterior probabilities (the probabilities that the input sequence belongs to each state). Based on these posterior probabilities, a pre-trained codebook and different estimation rules, the HB feature vectors are estimated. The estimation rules are based on maximum likelihood (ML), maximum a posteriori (MAP) and MMSE criteria. While ML and MAP based estimation rules take a form of classification and therefore the estimated features are limited to the discrete codebook entries only, continuous estimation is performed via the MMSE rule. Further details can be found in [78, Section 6.4]

In classification-based approaches, the false acceptances and rejections of high frequency sounds (e.g., /s/,/z/) lead to overestimation and underestimation (respectively) of their energies which cause artefacts in reconstructed speech signals. The work in [109] thus aims to modify the HMM based ABE algorithm in [78] in order to

[1]The GMM-based mapping is referred to as GMM regression (GMMR) throughout this thesis.

incorporate *a priori* knowledge of speech acoustic classes via phonetic transcription; the ABE training and estimation both are performed using phonetic information (an offline version). An online version is also presented in [110, 111, 112] where the estimation is done *without* the use of phonetic transcription. The improvements are reported in terms of both speech intelligibility and quality due to reduction in artefacts when ABE application used for automotive hands free systems as well as for hearing-impaired people. The performance of the online version is further improved in [113] by employing a DNN based classifier in order to estimate phonetic classes prior to ABE processing. More details related to these approaches can be found in [87]. There are several other approaches [114,115,116,117,118,119,120,121] which exploit capabilities of HMMs to model inter-frame dependencies.

2.2.1.5 Deep neural networks

Deep neural networks (DNNs) are capable of modelling complex non-linear relationships in the data. The early approaches to ABE [98, 122] which employed DNNs have not shown superior performance in terms of speech quality on speaker-independent ABE task. This is perhaps because of the use of shallow neural networks; it was not possible to train DNNs on large amounts of data due to lack of efficient hardware and training algorithms. However, in last two decades many machine learning techniques and computer hardware have evolved to efficiently train the DNNs that contain many hidden layers with large number of hidden activation units [123].

A modification of HMM-based ABE approach from [109] is presented in [86] in order to perform a direct comparison between the GMM and DNN based acoustic models for HB spectral envelope estimation. The posterior probabilities of the states (as defined by the centroids of a codebook trained on HB feature vectors) are calculated (via a GMM-HMM, hybrid DNN-HMM or DNN-only system) followed by MMSE estimation. While different DNN topologies showed no discernible performance, the improvement in speech quality was shown to attributed towards the improved estimation of the HB energy parameter than that of the spectral envelope. Further investigations of different input NB features (such as autocorrelation coefficients (AC) and log-mel filter bank coefficients), activation functions (such as rectified linear units (ReLUs) and sigmoids) and regularisation techniques (such as dropout) on ABE performance are presented in [124]; a DNN based regression approach is shown to outperform the GMM and DNN-based

classification schemes.

Several other approaches exploit superiority of DNNs to model linear and non-linear relationship between NB and HB components by employing DNNs with unsupervised pre-training [125], Gaussian Bernoulli restricted boltzmann machines (GBRBMs), deep recurrent-neural networks (RNNs) with Long short-term memory (LSTM) cells [126, 127, 128], conditional restricted Boltzmann machines (CRBMs) [129], recurrent temporal restricted Boltzmann machines (RTRBMs) [130].

2.2.2 Extension of excitation

According to the source-filter model of speech production, the excitation signal can be modelled by quasi-periodic impulse train for voiced sounds and by white noise for unvoiced sounds. In both the cases, the spectrum of the modelled excitation signal is flat [78, Section 2.1.1]. ABE algorithms exploit this simple structure of the excitation signal and also the insensitivity of the human auditory system to distortions in the excitation signal above 3.4kHz. The excitation extension is thus usually performed using similar operations that are used for non-model based ABE (section 2.1).

Non-linear processing involves the use of quadratic or cubic functions, half-wave or full-wave rectification on the time-domain excitation signal. Non-linear operations have two key properties: (1) they produce harmonics when applied to periodic signals and (2) depending on the effective bandwidth of the input signal they produce frequency components beyond Nyquist frequency. The NB time-domain excitation signal thus has to be upsampled to adequate sampling frequency before application of non-linear operations.

Modulation in the time domain corresponds to translation in the frequency domain; this property is usually used to generate HB components from the available NB excitation components. The *spectral translation* can be performed by multiplying the time domain NB excitation signal with a real-valued cosine function at a modulation frequency Ω_{M}. The HB components are then extracted via high pass or bandpass filtering. Typically Ω_{M} is chosen to be 3.1kHz which results in translation of NB (0.3-3.4kHz) frequency components from 3.4 to 6.5kHz. Bandwidth of the resulting extended signal (i.e., 6.5kHz) is dictated by the upper limit of NB (i.e.,

3.4kHz). The fixed modulation frequency creates inconsistent pitch harmonics in voiced phonemes of the extended speech signals resulting in metallic sounds. *Pitch-adaptive modulation* technique can then be employed where the modulation frequency Ω_{M} is adjusted in such a way that the shifted frequency components of the NB excitation correspond to harmonics of the pitch in the missing band. In this case, however, a very accurate pitch detection algorithm is required to reduce the artefacts.

When $\Omega_{\mathrm{M}} = \pi$ i.e., the Nyquist frequency, spectral translation corresponds to *spectral folding* or *mirroring*; the spectrum of the input NB excitation is folded or mirrored at the half of the Nyquist frequency, i.e., $\pi/2$. As the HB frequency components are exact folded copy of NB components, no filtering operation is needed. The spectral folding operation is equivalent to zero insertion in the time domain. During extension of telephone speech, a spectral gap is generated between 3.4 and 4.6kHz.

The initial contributions to the approaches to excitation extension which are explained above come from the works reported in [73, 88, 131, 132, 133, 134]. While most ABE approaches, which employ the source-filter model, mainly focus on estimation of spectral envelope, they use one of the excitation extension methods reported in this section. Slightly different approaches to excitation extension are reported in [102, 125, 135]. ABE methods also perform some form of post-processing in order to match the assumption of spectral flatness for the generated HB frequency components; it is important to match energy of the generated HB excitation components to that of the NB components [136, Chapter 4.5]. The additional details on excitation extension methods can be found in [78, Chapter 3], [79, Section 5.5.1], [136, Chapter 4], [15, Section 2.3.2].

The spectral gaps created in extended WB signals due to spectral folding do not adversely affect the speech quality if the envelope extension works reasonably well [16]. The listening tests reported in [137] showed that the adaptive pitch alignment which helps to reduce harmonic distortions in the HB spectrum do not yield significant improvements in speech quality relative to the required additional cost of complexity and memory.

2.3 ABE approaches based on direct modelling of spectra

Some ABE approaches (which are not based on source-filter model) operate directly on the higher dimensional complex speech spectra.

Some handful of ABE approaches [138, 139, 140] directly operate on DFT coefficients in which first missing frequency components are generated by a simple non-linear operation, e.g., spectral mirroring. The generated HB frequency components are then spectrally shaped by a set of parameters. An adaptive spline neural network (ASNN) is employed in [141] to directly map the NB DFT coefficients to the missing HB coefficients. The work in [138] combines the use of spline interpolation and a neural network to estimate a set sub band powers; the parameters which are then used to adaptively tune the spectral shape of the missing HB. A spectral magnitude shaping curve (defined by five control points) is constructed or learned using cubic spline interpolation [142] and neuro-evolutive neural network [139]; the curve is then used to shape the magnitude spectral coefficients of HB components. Further improvements to the ABE approach in [142] are obtained by a more accurate control of the HB spectral shape which are reported in [140, 143].

Other approaches directly estimate the spectral coefficients of the missing HB via statistical models. Magnitude of log-power spectrum (LPS) coefficients for missing HB are estimated using a DNN in [144]. ABE estimation performance using a DNN improved via the use of rich acoustic features for input NB representation; the oversmoothing problem of DNNs is reduced using global variance equalisation as a post-processing technique in [145]. The work in [146] uses deep bidirectional Long Short Term Memory (BLSTM)-based RNN. The target HB spectral features are generated through a weighted linear combination of real target exemplars; the method is used as a post-processing step to reduce the estimation errors[2]. Inspired by their use in automatic speech recognition (ASR), the ABE approach reported in [147] exploits the BN features [148] that are extracted using a DNN-based classifier to capture the linguistic information from the input NB speech. A deep LSTM-based RNN is then employed for estimation of HB components from these BN features which are supposed to capture phone-dependent characteristics and energy distributions of HB spectra.

[2]According to [145, 146], the techniques of global variance equalisation and exemplar based sparse representation are usually used in voice conversion and speech enhancement.

An ABE approach based on sum-product networks (SPNs), a form of probabilistic graphical models (PGMs), for the estimation of missing HB log-spectrograms is reported in [149]. A comparison of SPNs to different generative models such as Gaussian-Bernoulli RBMs (GBRBMs), conditional RBMs, higher order contractive auto-encoders (HCAEs) and generative stochastic networks (GSNs) is presented in [150]. Sparsity of the spectral features is exploited in [151] to learn joint dictionaries for the NB and WB spectral components in a coupled manner; the trained dictionary is then used during estimation of the HB components.

While the magnitude spectrum of HB components is estimated via statistical modelling, the HB phase spectrum is generated via imaged copy of NB phase[3] [144]. Alternately, the extended WB speech signals can also be reconstructed using magnitude spectrum only via Griffin and Lim algorithm (proposed in [152]) as in [150].

2.4 End-to-end approaches to ABE

Convolutional neural networks (CNNs) are capable of extracting useful features by directly operating on raw speech waveforms. Inspired by the success of WaveNet [153] and dilated convolutional architectures [154, 155], an approach to ABE is proposed in [156] using *stacked dilated* CNNs. The method avoids spectral analysis and phase modelling issues via direct modelling and generation of time-domain speech waveforms. The NB speech signals (at a sampling rate of 16kHz) are first fed to dilated CNNs to generate either HB or WB speech signals. The generated waveforms are then added to the available NB signals after appropriate highpass filtering. Modelling of HB speech waveforms at output of the CNNs was found to be more effective than that of WB waveforms. Inspired by SampleRNN architecture [157], the use of Hierarchical recurrent neural networks (HRNNs) composed of LSTM cells and feedforward layers is investigated in [158] for ABE. A comparison of several waveform modelling techniques is presented. HRNNs were shown to achieve better speech quality and run-time efficiency than the dilated CNNs. The major drawback of the waveform-based ABE methods is low run-time efficiency; they are time-consuming during generation of speech samples.

[3]Typically, phase spectrum of input NB speech frame is flipped to the upper half of the spectrum and then minus sign is added to the phase.

A light-weight approach to audio super-resolution is proposed in [159] where a deep CNN with residual (or skip) connections is trained increase sampling rate of input NB time series. The approach reported in [160] exploits information from speech signals in both the time and frequency domains using a time-frequency network (TFNet) that can be trained end-to-end. TFNet comprise two branches which are used to predict the high resolution (HR) audio samples and their spectral magnitude at output layers. Both outputs are then combined via a spectral fusion layer to synthesise the final HR output. An efficient alternative to the WaveNet, a deep learning architecture, referred to as FFTNet [161], is employed in [162] for ABE.

2.5 ABE with modified loss functions

Most ABE approaches usually employ standard mean-square error (MSE) criterion for optimisation which leads to over-smoothing problem. This is because the MSE loss function is minimised by averaging all plausible outputs. A regression model trained with a MSE loss function thus performs reasonably well in the average sense, however it fails to model the energy dynamics of different voiced and unvoiced sounds. Generative adversarial networks (GANs) [163] provide an alternative to the MSE loss function via *adversarial learning* wherein the HB features produced by a *generator* network are compared against the true HB features and classified as real or fake by a *discriminator* network. The goal of the adversarial learning is thus to make generated HB features indistinguishable from the true HB features and thereby producing perceptually better samples. The first application of GANs to ABE is reported in [163]. The work in [164] showed further improvements in ABE performance via the use of conditional GANs [165]. The approach in [166] employs GANs with stabilised training procedure – by adding penalty on the weighted gradient-norms of the discriminator network (proposed in [167]) – for ABE[4]. Another variant of GANs, also referred to as cycleGANs [168], trained with *cycle loss* is explored in [67] for the application of ASR.

In addition to estimation of spectral envelope parameters, ABE algorithms also estimate HB energy parameter; better estimation of HB energy reduces processing artefacts. ABE approaches usually suffer from the problem of HB energy over-

[4]In [166], ABE is alternately referred to as speech super-resolution (SSR). A GAN based approach is investigated for 2x (NB-to-WB extension) and 4x (NB-to-SWB extension) SSR task.

estimation which leads to audible and annoying artefacts. The work in [169] introduced an asymmetric cost function in the MMSE estimation of the energy ratio (between NB and HB); the cost function penalises the over-estimates of the HB energy more in comparison to the under-estimates. Reduction in artefacts was observed in the extended speech signals with the modified MMSE estimates. Importance of discriminative training of regression DNNs to avoid over-smoothing problem is investigated in [170]. It is suggested to add a discriminative term to the conventional MSE loss function during training in order to force DNNs to preserve the differences between speech sounds such as fricatives and vowels. Such discriminative training when used for GANs and cGANs is found to improve quality of extended speech signals [164].

2.6 Feature selection and memory inclusion for ABE

Few approaches to ABE have focused on investigation of feature selection and memory inclusion to improve estimation performance. They are discussed in brief in the following.

2.6.1 Feature selection

The first investigation of correlation between NB and HB features is reported in [171]. It is suggested that the ABE algorithms should not rely only on MI between NB and HB components and they should exploit the perceptually-relevant properties of speech. The work reported in [172] investigated the correlation of several NB features with HB cepstral coefficients in terms of mutual information (MI) as well as separability[5]. The analysis includes auto-correlation coefficients (ACs), LP coefficients, line spectral frequencies (LSFs), cepstral coefficients and MFCCs. The correlation properties of scalar, energy-based features[6] are also studied to include the information related to voice activity in the front-end features.

[5]While MI between two feature representations $\mathbf{x}$ and $\mathbf{y}$ provides the information gained on $\mathbf{y}$ from the knowledge of $\mathbf{x}$, separability measures the discriminative ability of the feature set $\mathbf{x}$ for a given classification problem.

[6]These features include gradient index, zero crossing rate, pitch period, local kurtosis, spectral centroid, spectral flatness, normalised relative frame energy, etc. These features provide good discriminative properties for distinction of different voice and unvoiced sounds [78, Section 5.3].

It is suggested that the chosen NB features should exhibit not only higher MI but also higher separability in order to improve ABE performance. The work further aims to combine different NB features heuristically based on their MI and separability to extract diverse information. Low-dimensional composite NB feature vectors with maximal compactness are then obtained by employing linear discriminant analysis (LDA). The work reported in [117] employs PCA instead, for dimensionality reduction.

In order to maintain a low complexity for ABE algorithms, the input NB features should be chosen optimally; the length of feature vectors should be low. An application of forward selection approach (proposed in [173]) for a regression DNN is reported in [174] in order to select a minimal input feature set – from a pool of many NB features – that yields a good estimation performance. The use the feature selection approach is also reported for a two-class DNN classifier which is used to discriminate between the sharp fricatives (/s/ and /z/) and other phonemes. The optimal feature set is shown to maintain the similar speech quality in comparison to that obtained with the full set while reducing the computational complexity.

2.6.2 Memory inclusion

ABE approaches exploit inter-frame temporal dependencies of speech signals to improve estimation performance. The contextual information or memory can be incorporated either via front-end memory inclusion or with back-end regression models or both.

Back-end memory inclusion

Several statistical models have inherent capability of modelling temporal information of speech signals. The speech dynamics in the form of temporal correlations between neighbouring speech frames can be captured implicitly via specific back-end regression models. Notable examples are Hidden Markov models (HMMs), temporally-extended Gaussian mixture models [175] and LSTM based RNNs, CRBMs, recurrent temporal RBMs (RTRBMs); all are capable of capturing memory.

Front-end memory inclusion

Other ABE approaches which employ GMMs or DNNs capture contextual information via front-end NB features. Memory can be incorporated by concatenation of the instantaneous static features with either dynamic delta features or static features extracted from neighbouring speech frames [86, 144, 145]. Such inclusion of memory leads to higher-dimensional features thereby increasing complexity of the regression model (refer to Section 1.6).

Drawing upon the work to optimise front-end feature extraction reported in [172], the first attempt to quantify the importance of front-end memory inclusion is reported in [85, 103, 104]. The work demonstrates the benefit of using memory in the form of delta features with a standard GMMR under the constraint of fixed dimensionality. The superior class-separability properties of MFCCs (as reported in [172]) are further exploited to improve cross-band correlations between NB and HB dynamic (static+delta) MFCC features.

2.7 Evaluation of speech quality

The evaluation of ABE performance is a sensitive problem in itself. Evaluation is often performed using estimates of speech quality, involving a comparison of extended WB speech signals to the original NB and WB speech signals. According to [176, Section 2.1], the content (text) of a speech signal has a strong influence on how its acoustic form is perceived by a listener. A human listener establishes a relationship between the content and the form of speech to give a judgment, referred to as speech quality. The perceived quality of speech is highly subjective, reflecting many dimensions such as naturalness, clarity, brightness and pleasantness [177]. Reliable evaluation of speech quality is an important factor and thus, an active field of research. Individual human listeners have their own perceptions of the speech quality and therefore, subjective tests should involve a substantial number of participants to obtain, on average, a reliable estimate.

The most reliable and meaningful approach to ABE assessment involves subjective listening tests. Such tests, however, when conducted over multiple sessions, often of substantial duration, may not yield meaningful results. This is because the listeners' ratings are highly dependent on psychological factors such as motivation, emotional state and fatigue. Obtaining reliable quality estimates via subjective

listening tests is thus time-consuming, expensive and cumbersome. Accurate prediction of speech quality that is perceived by user is important in the design of speech communication systems. Objective assessment measures thus have been developed in order to estimate automatically quality from speech signals. These measures provide a quick and inexpensive tool for speech quality evaluation and provide convenience during development of speech coding algorithms and communication systems [19]. Objective measures approximate the human perception mechanism in order to estimate speech quality. However, speech quality estimates are not as reliable as fully fledged subjective listening tests which remain the only reliable ground truth [178, Section 15.1.2].

Subjective assessment

Subjective tests involve the active participation of human listeners and the recording of their personal opinions. These opinions are typically referred to as opinion scores and reflect the perception of the quality of speech signals under test. The average quality ratings obtained from a pool of listeners is obtained thus giving speech quality estimates in terms of mean opinion score (MOS). Listening-only or conversational methods have been developed for the subjective assessment of telephony communication systems quality [179]. Conversational test setups have a more realistic nature in the sense that quality measure derived from them reflect actual service for listeners, therefore giving more reliable or meaningful measurements. However, such tests are generally infeasible during the initial development phase. In this case listening-opinion tests are thus conducted as a feasible option. are used instead.

Conversation-opinion tests: Conversation-opinion test methodology is explained in ITU-T Rec. P.800 [180] and ITU-T Rec. P.805 [179]. The tests are designed in order to assess the effect on the speech quality because of impairments caused during a telephone conversation; the impairments such as delay, packet loss, echo, clipping, noise, interruptions. The aim is to move a step closer to a real conversation over a telephone system where two subjects, who participate in the conversation, are placed in two sound proof rooms and rate the speech or voice quality. The telephone users may consider various aspects such as intelligibility, loudness, listening effort, or naturalness of the conversation during assessment [181]. The arithmatic mean over the subjective quality judgements of all the test subjects, gives MOS-Conversational Quality Score(MOS-CQS) as defined in ITU-T

Rec. P.800.1 [182].

In comparison to the listening-only tests, the conversational tests are complex and difficult to design and conduct [183, Section 2.2.3.2]. For further details, an overview of conversational tests can be found in [176] whereas relationship between listening and conversational speech quality is presented in [184].

Listening-only tests: In subjective listening-only tests (LOTs), a pair of pre-recorded and processed speech signals of shorter duration is presented to human listeners who are asked to give their personal opinion about the quality on a pre-defined scale [181]. Such tests are directed i.e. the perception of the subject is affected by listening test design factors and rating procedures [185]. In listening tests, subjects focus only on the acoustic form of the signal - that includes perturbations caused in signal characteristics due to codec, packet loss, transmission channel noise - rather than on its content [186, Section 9.5.2]. Therefore, LOTs do not provide as realistic measurements as conversational tests, however, they are of comparatively shorter duration and are less expensive and thus, allow the testing of more systems and conditions in certain duration.

The most popular used LOT in ITU-T Rec. P.800 [179] is the Absolute Category Rating (ACR) method. In this test, the quality of speech samples is rated on a 5-point listening-quality(LQ) scale, as defined in ITU-T Rec. P.800.1 [179]. The discrete rating scale ranges from 1 to 5 corresponding to bad, poor, fair, good and excellent speech quality. The arithmetic mean of all these scores represent the speech quality in terms of MOS subjective listening quality (MOS-LQS), as defined in ITU-T Rec. P800.1 [182]. For good quality systems, the ACR method has low sensitivity in terms of its distinguishing capacity.

Alternately, the Degradation Category Rating (DCR) method, degradations caused due to the system under test are compared in comparison with a high quality reference on the five-point annoyance or degradation category scale. The scale scores (very annoying, annoying, slightly annoying, audible but not annoying, inaudible) degradations in a processed speech sample relative to an unprocessed (reference) sample, in a range from 1 to 5 (respectively). The reference speech sample is always presented first, followed by the same but processed sample. The quantity measured using these scores is referred as degradation MOS (DMOS). The DCR method with these modifications over ACR test, offers higher sensitivity in evaluation of good quality speech and it is suitable particularly when the degradations are small [179].

The comparison Category Rating (CCR) approach, a variation of the DCR test, measures quality rather than degradation of a processed speech sample. A single test involves a comparison between the processed sample and a reference. Listeners rate the test signal on a scale of -3 (much worse), -2 (slightly worse), -1 (worse), 0 (about the same), 1 (slightly better), 2 (better), 3 (much better). In contrast to DCR test, the order in which the speech samples are presented to a listener is shuffled for half of the test pairs [179] and quality of the second sample is judged compared to that of the first. CCR methods can thus be used to evaluate both the degradation or the improvement in the quality of speech, measured in terms of comparison MOS (CMOS)[7].

Objective assessment

Objective or instrumental metrics exploit signal characteristics for automatic evaluation of speech quality. Some of these are *distance* metrics which calculate the amount of distortion created in a processed signal in comparison to a reference signal. Notable examples include the segmental signal-to-noise ratio (segSNR), mean square error, or spectral distance measures (such as root mean square log spectral distance (RMS-LSD), COSH distortion). However, according to investigations in [187], these measures typically exhibit moderate correlation with subjective quality scores. Therefore, other measures try to mimick the human perception mechanism in order to judge speech quality as it would be rated by human listeners [13], e.g., perceptual evaluation of speech quality (PESQ) [188] and its WB extension [189], POLQA (a successor of PESQ which supports for quality evaluation of FB signals) [190], a quality measure for artificially bandwidth-extended (QABE) signals [191].

2.7.1 Assessement of different ABE algorithms

As discussed before, the assessment of different ABE algorithms is still an open issue. The results of a formal listening tests in three different languages (American English, Mandarin Chinese, Russian) are reported in [192,193]. After application of ABE to AMR-NB coded signals, quality of speech is shown to improve significantly. However, different variants of a *same* ABE algorithm were evaluated. The work reported in [194] evaluated performance of five different ABE algorithms in terms

[7]Before averaging the opinion scores (to obtain CMOS), sign of the scores for half of the test pairs are reversed in order to normalise the order of presentation.

of WB-PESQ, POLQA and ACR subjective listening tests. No statistical difference was found between G.711-coded NB speech signals and their bandwidth-extended versions. WB-PESQ was shown to exhibit significant Pearson correlation (0.93) with the results of subjective listening tests in comparison to POLQA (0.87). The objective measures failed to predict correct rank order of the ABE algorithms under assessment.

In [195], four different algorithms are evaluated. Both WB-PESQ and POLQA failed to predict rank orders correctly as given by the ACR subjective listening tests and it is concluded that no objective measure is fully capable of replacing subjective assessments. Subjective test results showed correlation of 0.82 for WB-PESQ and 0.75 for POLQA. While one ABE algorithm was found to be of equivalent speech quality as AMR-NB codec in ACR test, it was rated with significantly better quality in CCR test; subjective ratings also showed inconsistency. It is suggested that ACR listening tests are more suitable for quality assessment when a NB call is followed by a WB call whereas CCR tests better represent a handover scenario where the call switches between NB and WB modes. In [196], the quality of AMR-NB coded signals extended by three different ABE algorithms and those processed with different standard speech codecs is evaluated; background noise condition is also considered. The results showed different results that POLQA correlates better than WB-PESQ with the subjective quality ratings. Evaluation of six different ABE algorithms in four different languages (American English, Korean, Chinese and German) is reported in [197]. While no ABE algorithm gave consistent improvements in all the languages, every approach failed to give statistically significant improvements over AMR-NB signals in Chinese.

The failure of WB-PESQ and POLQA measures for speech quality assessment is perhaps because they are not designed to test ABE conditions. A QABE quality measure specifically designed to assess ABE performance is proposed in [191] which is shown to outperform WB-PESQ and POLQA in terms of three metrics used to assess the instrumental measures. The analyses on quality assessment of ABE methods show that a careful design of a subjective listening test setup is important; the instrumental measures of speech quality such as WB-PESQ and POLQA are not suitable to assess different ABE algorithms, particularly to predict their rank orders; subjective as well as instrumental assessment of speech quality is still an open topic in ABE research.

2.8 Approaches to super-wide bandwidth extension (SWBE)

The extensive body of research in the ABE literature targets mostly the extension of NB speech signals to WB speech signals. In this case there is substantial potential to improve speech quality because significant information between the NB limit of 4kHz and the WB limit of 8kHz can be recovered reliably using ABE. SWB speech signals extend the limit to 16kHz. Super-wide bandwidth extension (SWBE) approaches can then be employed to recover missing high-frequency (HF) components between 8kHz and 16kHz using the available low-frequency (LF) components between 0.05-8kHz.

2.8.1 SWBE for audio signals (speech and music)

Few approaches perform SWBE for audio (speech as well as music) signals via *non-linear prediction* methods which exploit principles of audio signals and the characteristics of the human ear. First, LF MDCT coefficients are modelled using a time-series; second, the phase spectrum of LF MDCT coefficients is then reconstructed via so-called phase space reconstruction (PSR) methods; and finally the HF MDCT coefficients are recovered using non-linear prediction techniques. The approach in [198] uses chaotic prediction theory[8] for PSR and the sub-band energies of the estimated HF MDCT coefficients are normalised with respect to that of LF MDCT coefficients in order to reduce large prediction errors. In [199, 200], a radial basis function (RBF) neural network is employed for non-linear estimation of HF MDCT coefficients. While voltera series based non-linear prediction technique is adopted to recover HF MDCT coefficients in [201], their energies are adjusted via the use of GMM and codebook mapping. A PSR technique from [202] is adopted for SWBE in [203] wherein LF MDCT coefficients of WB audio are converted to a multi-dimensional space and the HF MDCT coefficients reconstructed by a non-linear prediction model.

The modulated lapped transform (MLT) based PSR approaches for SWBE are reported in [56, 204]. While a non-linear prediction method based on nearest-

[8] According to the chaotic characteristics of audio spectrum, the phase spectrum of the MDCT coefficients for LF components can be reconstructed in order to predict the HF MDCT coefficients [199].

neighbour matching (NNM) is proposed to estimate HF MLT coefficients of WB audio signals, the HF sub-band energies are estimated via GMMs [204] or HMMs [56, 205].

A source-filter model based SWBE approach is presented in [206] combining ensemble echo state networks (ESNs) with HMMs for estimation of HF spectral envelopes. In [207] temporal information from audio signals is extracted in the form of temporal smoothing cepstral coefficient (TSCC) features. The TSCC features when used with conventional GMMR are shown to provide higher MI than the MFCCs and improved speech quality.

An approach reported in [208], referred to as efficient high-frequency bandwidth extension (EHBE), performs a non-linear operation on audio signals in the time-domain in order to generate missing HF components. Full-wave rectification[9] is employed in order to generate the missing HF components from those available in the highest octave of WB signal.

2.8.2 SWBE for speech only

Most SWBE approaches reported above are designed for audio signals. A small number of attempts, e.g. [209,210], have been made to improve SWBE performance while focusing on properties of only speech signals. The approach presented in [210] performs direct manipulation of DFT coefficients corresponding to 6 to 7kHz band via a pitch-scaling operation to obtain HF spectral content. The use of a conditional codebook mapping (CCM) method for SWBE is reported in [211] to overcome shortcomings of conventional codebook mapping (CM) methods. The work in [209] employs a DNN regression model for SWBE of WB signals coded with the EVS codec.

2.9 Summary

This chapter presents a thorough review of existing solutions to NB-to-WB and WB-to-SWB artificial bandwidth extension. The ABE algorithms are mainly categorised into: non-model based approaches, approaches based on the source-filter model,

[9]Full-wave rectification is an efficient method of harmonic generation in which the resulting spectrum consists of even harmonics of the fundamental frequency [95, Section 2.3.2.2].

approaches based on direct modelling of spectral coefficients or approaches which operate on raw speech waveforms (end-to-end approaches). The existing ABE approaches are also reviewed according to other perspectives such as those utilise perceptually motivated cost functions; feature selection and memory inclusion for bandwidth extension. Evaluation of speech quality is critical for assessment of ABE algorithms; review of typical subjective and objective methods is presented; the works which focus on evaluation of different ABE algorithms are also discussed.

Chapter 3

Baseline, databases and metrics

This chapter describes the baseline ABE algorithm used throughout this thesis. Adapted from [78], it is based upon a classical source-filter model of speech production whereby a narrowband speech signal is modelled by two distinct components: a vocal tract filter and an excitation source. Both components are extended in bandwidth separately before a wideband speech signal is resynthesised. The baseline algorithm differs to that presented in [78] in the choice of narrowband and highband features as well as in the choice of the statistical regression model. Narrowband features are log Mel filter energy coefficients. Highband features take the form of linear prediction coefficients. A conventional Gaussian mixture model based mapping or regression technique is employed for statistical estimation of missing highband components.

Also presented in this chapter are the databases used for training, validation and testing of the ABE algorithms developed throughout the thesis. Finally, objective metrics and subjective listening tests for ABE performance assessments are presented. Particular attention is given to information theoretic analysis via mutual information as a measure of correlation between narrowband and highband features.

3.1 ABE algorithm

The baseline ABE algorithm is illustrated in Fig. 3.1. It comprises three distinct blocks:

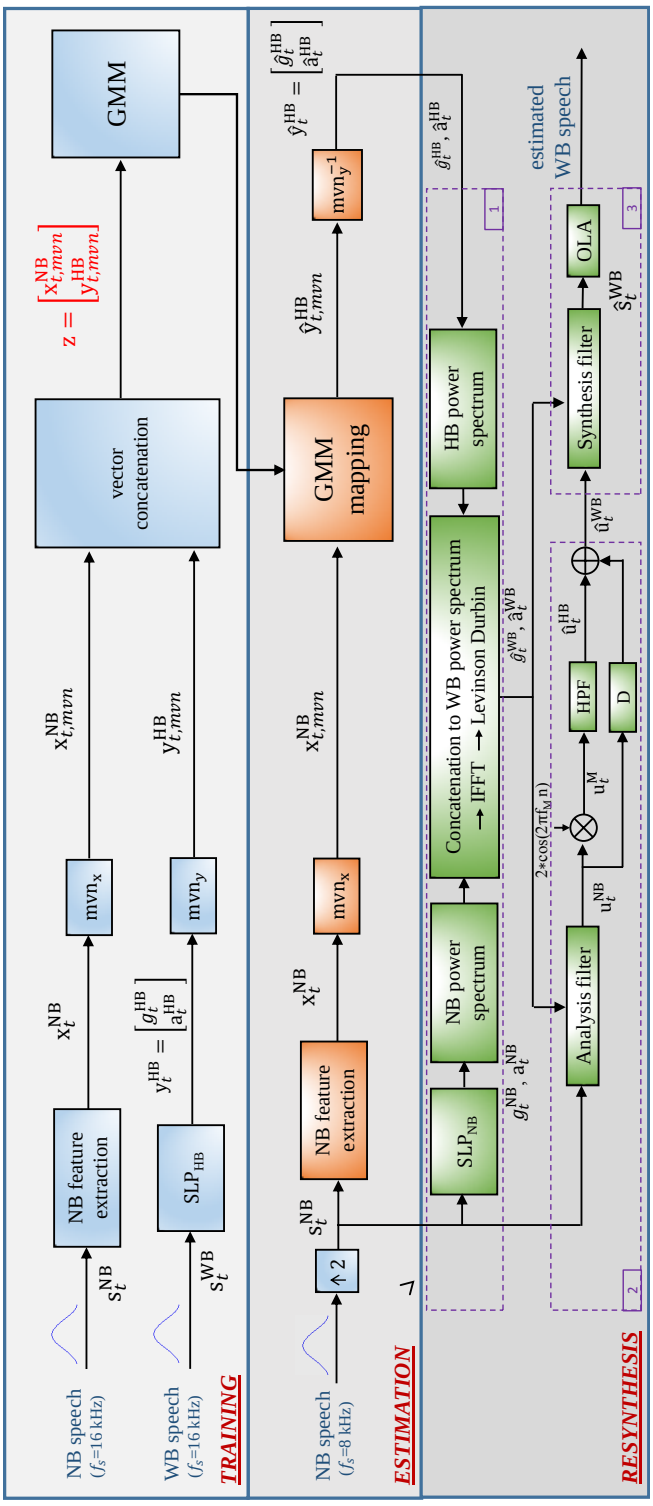


Figure 3.1: A block diagram of the baseline ABE system. A modified version of the ABE system presented in [78]. $\mathbf{s}_t^{\mathrm{NB}}$ denotes a NB speech frame at a sampling rate of 16kHz.

- **Training** (top block) uses parallel NB and WB[1] speech signals for GMM modelling.
- **Estimation** (middle block) of missing HB LP coefficients is performed from NB speech parametrised by some form of features.
- **Resynthesis** (bottom block) is performed using original NB speech and estimated HB LP coefficients.

Each step corresponding to the three blocks of Fig. 3.1 are explained in detail in the following.

3.1.1 Training

The training process is illustrated to the top of Fig. 3.1. Parallel NB and WB signals (both at a sampling rate of 16kHz)[2] are processed frame-by-frame. Frames are represented by $\mathbf{s}_t^{\mathrm{NB}}$ and $\mathbf{s}_t^{\mathrm{WB}}$, where t denotes the frame index. The framing operation is performed with sliding Hann windows of 20ms duration and 50% overlap. A 1024-point fast Fourier transform (FFT) is used to perform discrete Fourier Transform (DFT) operations.

Feature extraction is performed to obtain NB and HB features. NB features ($\mathbf{x}_t^{\mathrm{NB}}$ - top line in training block) with 10 coefficients are extracted from the input NB speech frame $\mathbf{s}_t^{\mathrm{NB}}$. HB features ($\mathbf{y}_t^{\mathrm{HB}}$ - bottom line in training block) consist of 9 LP coefficients $\mathbf{a}_t^{\mathrm{HB}}$ and a gain coefficient g_t^{HB} extracted by applying selective linear prediction (SLP)[3] to the HB frequency components of parallel WB speech frame $\mathbf{s}_t^{\mathrm{WB}}$. Both feature sets are mean and variance normalised ($\mathrm{mvn}_{\mathbf{x}}$ and $\mathrm{mvn}_{\mathbf{y}}$) giving $\mathbf{x}_{t,mvn}^{\mathrm{NB}}$ and $\mathbf{y}_{t,mvn}^{\mathrm{HB}}$. Feature extraction is followed by the statistical modelling of NB and HB features vectors which involves the fitting of a GMM to joint vectors $\mathbf{z} = [\mathbf{x}_{t,mvn}^{\mathrm{NB}}\ \mathbf{y}_{t,mvn}^{\mathrm{HB}}]^T$. This is achieved by maximising the likelihood of the data using the expectation-maximisation (EM) algorithm [214, Section 9.2.2]. Thus, the probability distribution of the vectors $\mathbf{z}$ is given by a mixture of M multi-variate

[1] The data pre-processing steps applied to generate parallel NB and WB speech data are discussed in Section 3.3.1.

[2] NB speech signals are upsampled to 16kHz before training.

[3] SLP [212] is a spectral modelling technique that fits an all-pole or auto-regressive model to a specific frequency region of a signal. A detailed explanation of SLP analysis can be found in [213, Section 6.4], [78, Section 4.1.2].

Gaussians as follows:

$$p(\mathbf{z};\Theta) = \sum_{m=1}^{M} \pi_m \mathcal{N}(\mathbf{z};\mu_m,\Sigma_m) \tag{3.1}$$

$$= \sum_{m=1}^{M} p(\lambda_m)\mathcal{N}(\mathbf{z};\mu_m,\Sigma_m) \tag{3.2}$$

where the m^{th} multi-variate Gaussian distribution $\mathcal{N}(\mathbf{z};\mu_m,\Sigma_m)$ is represented by the mean vector μ_m and covariance matrix Σ_m. Given a discrete random variable Λ that takes values $\{\lambda_m\}_{m\epsilon\{1,\ldots,M\}}$, the parameters π_m represent mixing coefficients or weights such that, $\sum_{m=1}^{M}\pi_m = 1$ and $\pi_m = p(\lambda_m) \geq 0, \forall m$. $p(\lambda_m)$ can be viewed as the prior probability that the observation $\mathbf{z}$ is generated by the m^{th} Gaussian component.

The GMM is randomly initialised with 128 Gaussian densities modelled using full-covariance matrices[4]. The parameters $\Theta = \{\pi_m,\mu_m,\Sigma_m\}_{m\epsilon\{1,\ldots,M\}}$ are estimated using 100 iterations of the EM algorithm. These parameters are then used for estimation of HB frequency components.

3.1.2 Estimation

Estimation of HB components (middle block of Fig. 3.1) begins with input NB speech signals which are again processed frame-by-frame. Upsampling is performed to produce NB speech frames $\mathbf{s}_t^{\mathrm{NB}}$ at 16kHz. NB features $\mathbf{x}_t^{\mathrm{NB}}$ are extracted from NB speech frames $\mathbf{s}_t^{\mathrm{NB}}$. Mean-variance normalisation ($\mathrm{mvn_x}$) is applied using means and variances obtained from training data to produce features $\mathbf{x}_{t,mvn}^{\mathrm{NB}}$. HB features $\hat{\mathbf{y}}_{t,mvn}^{\mathrm{HB}}$ are then estimated according to the minimum mean-square error (MMSE) criterion. The latter is applied to minimise $E[\|\mathbf{y}_{t,mvn}^{\mathrm{HB}} - \hat{\mathbf{y}}_{t,mvn}^{\mathrm{HB}}\|^2]$ which represents the MSE between the true HB features $\mathbf{y}_{t,mvn}^{\mathrm{HB}}$ and their estimates $\hat{\mathbf{y}}_{t,mvn}^{\mathrm{HB}}$. For NB features $\mathbf{x}_{t,mvn}^{\mathrm{NB}}$, the MMSE estimate of HB features $\hat{\mathbf{y}}_{t,mvn}^{\mathrm{HB}}$ is thus given by:

$$\hat{\mathbf{y}}_{t,mvn}^{\mathrm{HB}} = E[\mathbf{y}_{t,mvn}^{\mathrm{HB}}|\mathbf{x}_{t,mvn}^{\mathrm{NB}}] \tag{3.3}$$

[4]In order to achieve the same level of ABE performance, the GMMs with diagonal-covariance matrices are more computationally expensive than with full-covariance matrices [15, Section 3.3.3]. This motivates the use of full-covariance GMMs in our work.

By introducing a discrete random variable Λ that takes values $\{\lambda_m\}_{m\epsilon\{1,\dots,M\}}$ such that $p(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{M} p(\mathbf{x},\mathbf{y},\lambda_m)$, Eq. 3.3 can be re-written as:

$$\hat{\mathbf{y}}^{\mathrm{HB}}_{t,mvn} = \sum_{m=1}^{M} p(\lambda_m|\mathbf{x}^{\mathrm{NB}}_{t,mvn}) E[\mathbf{y}^{\mathrm{HB}}_{t,mvn}|\mathbf{x}^{\mathrm{NB}}_{t,mvn},\lambda_m] \tag{3.4}$$

Variables $\{\lambda_m\}_{m\epsilon\{1,\dots,M\}}$ represent the parameters of the joint probability density function (PDF) $p(\mathbf{x},\mathbf{y})$ which is modelled using a GMM with M-Gaussian components.

If a random vector $\mathbf{z} = [\mathbf{x}\ \mathbf{y}]^T$ has a multivariate Gaussian PDF with mean vector $\mu = [\mu^{\mathbf{x}}\ \mu^{\mathbf{y}}]^T$ and covariance matrix $\Sigma = \begin{bmatrix} \Sigma^{\mathbf{xx}}_m & \Sigma^{\mathbf{xy}}_m \\ \Sigma^{\mathbf{yx}}_m & \Sigma^{\mathbf{yy}}_m \end{bmatrix}$ then the conditional PDF $p(\mathbf{y}|\mathbf{x})$ is also Gaussian [215, Theorem 10.1], [214, Section 2.3.1] with a mean vector given by:

$$E(\mathbf{y}|\mathbf{x}) = \mu^{\mathbf{y}} + \Sigma^{\mathbf{yx}}\Sigma^{\mathbf{xx}^{-1}}(\mathbf{x} - \mu^{\mathbf{x}}) \tag{3.5}$$

The marginal distribution $p(\mathbf{x})$ is also Gaussian, i.e:

$$p(\mathbf{x}) \sim \mathcal{N}(\mathbf{x}; \mu^{\mathbf{x}}, \Sigma^{\mathbf{xx}}) \tag{3.6}$$

Therefore, according to Bayes' theorem and Eq. 3.6, the first term $p(\lambda_m|\hat{\mathbf{x}}^{\mathrm{HB}}_{t,mvn})$ of Eq. 3.4 can be written as:

$$\begin{aligned} p(\lambda_m|\mathbf{x}^{\mathrm{HB}}_{t,mvn}) &= \frac{p(\lambda_m)p(\mathbf{x}^{\mathrm{NB}}_{t,mvn}|\lambda_m)}{\sum_{n=0}^{M} p(\lambda_n)p(\mathbf{x}^{\mathrm{NB}}_{t,mvn}|\lambda_n)} \\ &= \frac{\pi_m\mathcal{N}(\mathbf{x}^{\mathrm{NB}}_{t,mvn}; \mu^{\mathbf{x}}_m, \Sigma^{\mathbf{xx}}_m)}{\sum_{n=0}^{M} \pi_m\mathcal{N}(\mathbf{x}^{\mathrm{NB}}_{t,mvn}; \mu^{\mathbf{x}}_m, \Sigma^{\mathbf{xx}}_m)} \end{aligned} \tag{3.7}$$

Eq. 3.7 represents the posterior probability of λ_m given NB features $\mathbf{x}^{\mathrm{NB}}_{t,mvn}$.

Similarly, using Eq. 3.5, the second term of Eq. 3.4 can be written as:

$$E[\mathbf{y}^{\mathrm{HB}}_{t,mvn}|\mathbf{x}^{\mathrm{NB}}_{t,mvn},\lambda_m] = \mu^{\mathbf{y}}_m + \Sigma^{\mathbf{yx}}_m\Sigma^{\mathbf{xx}^{-1}}_m(\mathbf{x}^{\mathrm{NB}}_{t,mvn} - \mu^{\mathbf{x}}_m) \tag{3.8}$$

Finally, combining Eqs. 3.7 and 3.8 and using parameters π_m, $\mu_m = \begin{bmatrix} \mu^{\mathbf{x}}_m \\ \mu^{\mathbf{y}}_m \end{bmatrix}$ and $\Sigma_m =$

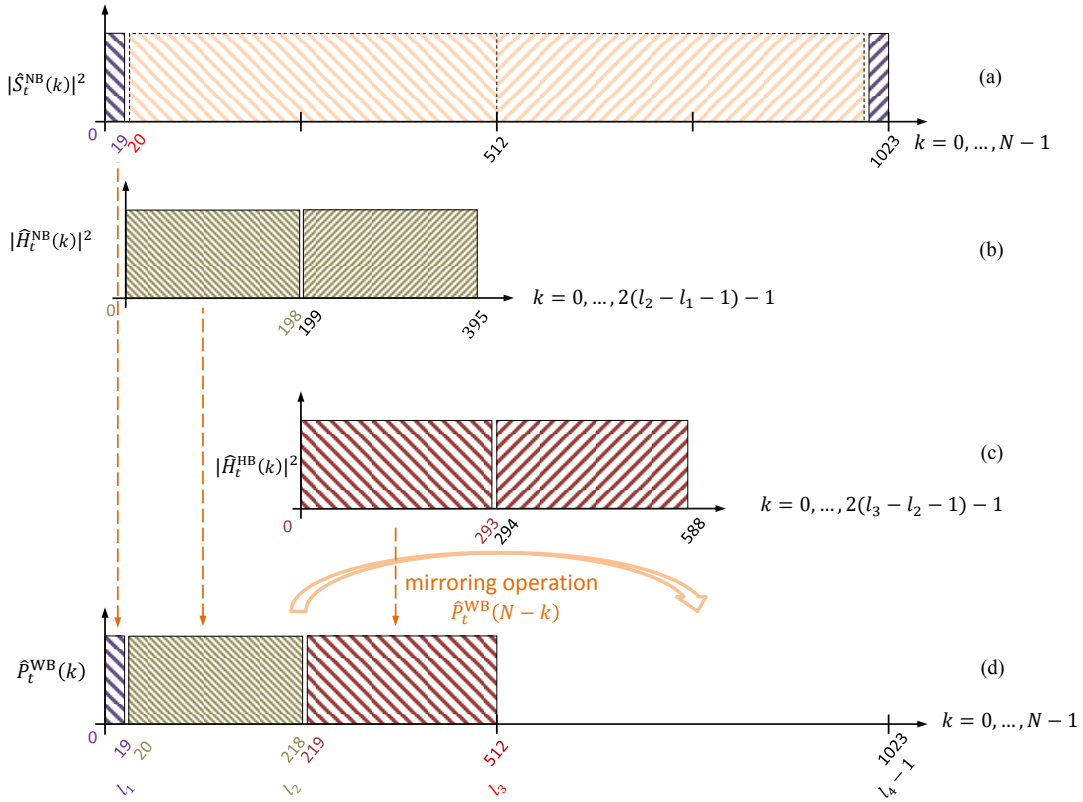


Figure 3.2: Illustration of concatenation of lowband (LB), narrowband (NB) and estimated highband (HB) power spectra to obtain the estimated wideband (WB) power spectrum $P_t^{\mathrm{WB}}(k)$ calculated according to Eq. 3.10 for 1024-point FFT.

$\begin{bmatrix} \Sigma_m^{\mathbf{xx}} & \Sigma_m^{\mathbf{xy}} \\ \Sigma_m^{\mathbf{yx}} & \Sigma_m^{\mathbf{yy}} \end{bmatrix}$ learned during GMM training (Section 3.1.1), HB features $\hat{\mathbf{y}}_{t,mvn}^{\mathrm{HB}}$ are estimated according to[5]:

$$\hat{\mathbf{y}}_{t,mvn}^{\mathrm{HB}} = \sum_{m=1}^{M} \frac{\pi_m \mathcal{N}(\mathbf{x}_{t,mvn}^{\mathrm{NB}}; \mu_m^{\mathbf{x}}, \Sigma_m^{\mathbf{xx}})}{\sum_{n=0}^{M} \pi_n \mathcal{N}(\mathbf{x}_{t,mvn}^{\mathrm{NB}}; \mu_n^{\mathbf{x}}, \Sigma_n^{\mathbf{xx}})} [\mu_m^{\mathbf{y}} + \Sigma_m^{\mathbf{yx}} \Sigma_m^{\mathbf{xx}^{-1}} (\mathbf{x}_{t,mvn}^{\mathrm{NB}} - \mu_m^{\mathbf{x}})] \tag{3.9}$$

Using means and variances obtained from the training data, inverse mean and variance normalisation (mvn_y^{-1}) is then applied to $\hat{\mathbf{y}}_{t,mvn}^{\mathrm{HB}}$ to estimate HB features $\hat{\mathbf{y}}_t^{\mathrm{HB}}$, thus giving estimated HB LP coefficients $\hat{\mathbf{a}}_t^{\mathrm{HB}}$ and gain $\hat{g}_t^{\mathrm{HB}}$.

[5]Refer to [15, Section 3.3.1] for a detailed derivation of Eqs. 3.4–3.9.

3.1.3 Resynthesis

After estimation of HB spectral envelope parameters for an input NB speech frame, resynthesis of the extended WB signal is performed according to the three distinct steps illustrated by the numbered sub-blocks to the bottom of Fig. 3.1.

Step 1 – WB spectral envelope estimation:

In order to resynthesise bandwidth-extended WB speech frames, LP coefficients $\hat{\mathbf{a}}_t^{\mathrm{WB}}$ and gain coefficient $\hat{g}_t^{\mathrm{WB}}$ corresponding to the WB spectral envelope are first calculated from the NB speech frame $\mathbf{s}_t^{\mathrm{NB}}$ and estimated HB parameters $\hat{\mathbf{a}}_t^{\mathrm{HB}}$ and $\hat{g}_t^{\mathrm{HB}}$. These are determined from the auto-correlation coefficients (ACs) obtained from application of the inverse fast Fourier transform (IFFT) to the WB power spectrum $\hat{P}_t^{\mathrm{WB}}(k)$, followed by application of the Levinson-Durbin recursion.

Let $\hat{S}_t^{\mathrm{NB}}(k)$ represent the FFT of input NB speech frame $\mathbf{s}_t^{\mathrm{NB}}$. The WB power spectrum $\hat{P}_t^{\mathrm{WB}}(k)$ is obtained by concatenation of lowband and narrowband power spectra (both extracted from NB speech frame $\mathbf{s}_t^{\mathrm{NB}}$) with the estimated HB power spectrum (extracted using the estimated WB LP parameters $\hat{\mathbf{a}}_t^{\mathrm{HB}}$ and $\hat{g}_t^{\mathrm{HB}}$) according to:

$$\hat{P}_t^{\mathrm{WB}}(k) = \begin{cases} \frac{1}{N}|\hat{S}_t^{\mathrm{NB}}(k)|^2 & \text{for } 0 \leq k \leq l_1 \text{ (lowband power spectrum)} \\ \frac{1}{N}|\hat{H}_t^{\mathrm{NB}}(k-l_1-1)|^2 & \text{for } l_1 < k \leq l_2 \text{ (narrowband power spectrum)} \\ \frac{1}{N}|\hat{H}_t^{\mathrm{HB}}(k-l_2-1)|^2 & \text{for } l_2 < k \leq l_3 \text{ (highband power spectrum)} \\ P_t^{\mathrm{WB}}(N-k) & \text{for } l_3 < k \leq l_4 \end{cases} \tag{3.10}$$

where

- l_1, l_2, l_3 and l_4 represent the discrete DFT bins (or indices) corresponding to frequencies of 300, 3400 and 8000 and 16000 Hz. For a 1024-point FFT, according to $f_{\mathrm{Hz}} = \frac{Fs}{N}(k-1)$, these values are $l_1 = 19$, $l_2 = 218$, $l_3 = 512$, $l_4 = 1024$[6];

[6]The DFT bins l_1 and l_2 correspond to exact frequencies of 296.875Hz and 3406.3Hz respectively.

- The transfer functions,

$$\hat{H}_t^{\mathrm{NB}}(k) = \hat{H}_t^{\mathrm{NB}}(z)\Big|_{z=e^{j\frac{2\pi}{L_1}k}} \text{ for } k = 0, \ldots, L_1 - 1 \text{ with } L_1 = 2(l_2 - l_1 + 1),$$

$$\hat{H}_t^{\mathrm{HB}}(k) = \hat{H}_t^{\mathrm{HB}}(z)\Big|_{z=e^{j\frac{2\pi}{L2}k}} \text{ for } k = 0, \ldots, L_2 - 1 \text{ with } L_2 = 2(l_3 - l_2 + 1)$$

represent NB and estimated HB synthesis filters respectively defined according to:

$$\hat{H}_t^{\mathrm{NB}}(z) = \frac{\hat{g}_t^{\mathrm{NB}}}{1 + \sum_{i=1}^{p} \hat{a}_t^{\mathrm{NB}}(i) z^{-i}} \text{ and}$$

$$\hat{H}_t^{\mathrm{HB}}(z) = \frac{\hat{g}_t^{\mathrm{HB}}}{1 + \sum_{i=1}^{p} \hat{a}_t^{\mathrm{HB}}(i) z^{-i}} \ .$$

The concatenation procedure according to Eq. 3.10 is illustrated in Fig. 3.2. The one-sided WB power spectrum, shown in Fig. 3.2(d)), is obtained by concatenation of: (1) LB components[7] (shown by blue lines) extracted from power spectrum $|\hat{S}_t^{\mathrm{NB}}(k)|^2$, shown in Fig. 3.2(a); (2) NB components (shown by grey lines) extracted from power spectrum $|\hat{H}_t^{\mathrm{NB}}(k)|^2$, shown in Fig. 3.2(b); (3) HB components (shown by red lines) extracted from power spectrum $|\hat{H}_t^{\mathrm{HB}}(k)|^2$, shown in Fig. 3.2(c). The double-sided power spectrum[8] $\hat{P}_t^{\mathrm{WB}}(k)$ is then obtained using a mirroring operation.

Step 2 – WB excitation estimation:

In order to obtain the extended WB excitation, first, the NB speech frame $\hat{\mathbf{s}}_k^{\mathrm{NB}}$ (at 16kHz) is filtered using a LP analysis filter, $A_{\mathrm{NB}}(z) = \frac{1}{H_{\mathrm{NB}}(z)}$, to obtain the NB excitation $\mathbf{u}_k^{\mathrm{NB}}$ (at 16kHz). The NB excitation is then extended via spectral translation[9] with a modulation frequency f_m of 6.8kHz.

The modulated excitation $u_t^{\mathrm{M}}(n)$[10] and the corresponding spectrum $U_t^{\mathrm{M}}(f)$ are

[7]Note that the power spectrum can be calculated in two ways, either by performing the FFT on the speech frame directly or from the LP spectral envelope of the speech frame. Therefore the lowband components of $P_t^{\mathrm{WB}}(k)$ (in the frequency range 0-300Hz) can also be calculated using the lowband LP spectral envelope obtained via SLP.

[8]The power spectrum $P_t^{\mathrm{WB}}(k)$ in the frequency range 0-3.4kHz can be obtained in one step (without separate calculation of its lowband (0-300Hz) and narrowband (0.3-3.4kHz) components as in Eq. 3.10). However, the later option can be easily modified if low-bandwidth extension needs to be employed.

[9]Spectral folding leads to a spectral gap in the extended WB speech frame whereas spectral translation with modulation frequency $f_m = 3.4$kHz generates an extended frame that is band limited to 6.8kHz. Choice of $f_m = 6.8$kHz avoids these problems.

[10]Note that $u_t^{\mathrm{M}}(n)$ denote samples of the vector $\mathbf{u}_k^{\mathrm{NB}}$ for given n.

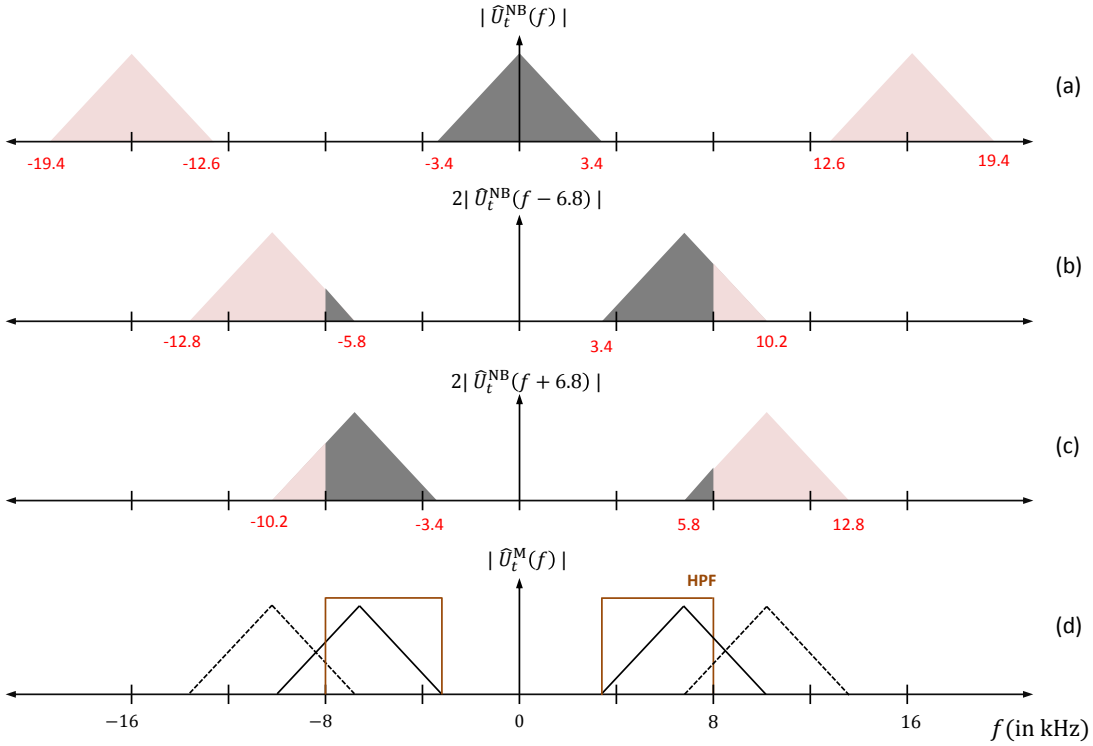


Figure 3.3: Illustration of excitation extension via spectral translation with modulation frequency $f_m = 6.8$kHz. Plot (a) represents the magnitude spectrum $|\hat{U}_t^{\mathrm{NB}}(f)|$ of a narrowband (NB) speech frame $\hat{\mathbf{s}}_k^{\mathrm{NB}}$ (at 16kHz) with a bandwidth of 3.4kHz. Plots (b) and (c) illustrate the translated copies of $|\hat{U}_t^{\mathrm{NB}}(f)|$ after modulation with a cosine signal of frequency 6.8kHz. Plot (d) shows the magnitude spectrum of the resulting modulated frame to which a HPF is applied to extract the highband (HB) excitation components.

given by:

$$u_t^{\mathrm{M}}(n) = 2u_t^{\mathrm{NB}}(n)\cos(2\pi f_{\mathrm{M}} n), \text{ and}$$

$$U_t^{\mathrm{M}}(f) = U_t^{\mathrm{NB}}(f - f_{\mathrm{M}}) + U_t^{\mathrm{NB}}(f + f_{\mathrm{M}}).$$

An illustration of spectral translation with $f_m = 6.8$kHz is presented in Fig. 3.3. The spectrum $U_t^{\mathrm{NB}}(f)$ of the NB excitation frame (at 16kHz) that is bandlimited to 3.4kHz is shown in Fig. 3.3(a). Two translated copies $U_t^{\mathrm{NB}}(f - f_m)$ and $U_t^{\mathrm{NB}}(f + f_m)$ of the spectrum $U_t^{\mathrm{NB}}(f)$ are generated after multiplication of $u_t^{\mathrm{NB}}(n)$ in the time domain with a cosine signal $\cos(2\pi f_{\mathrm{M}} n)$. These are shown in Fig. 3.3(c) and (d). The spectrum of the modulated frame $U_t^{\mathrm{M}}(f)$ is the sum of the resulting shifted spectra (see Fig. 3.3(d)). The spectrum $U_t^{\mathrm{M}}(f)$ (for $f_m = 6.8$kHz) within the frequency

range 3.4-8kHz is, therefore, a sum of two aliased frequency components. The frequency components of the spectrum $U_t^{\mathrm{NB}}(f)$ from -3.4 to 1.2kHz are translated to the 3.4-8kHz band whereas those from 12.6 to 13.8kHz are translated to the 5.8-8kHz band. This leads to some overlap between frequency components in the frequency range 5.8-8kHz. Spectral translation thus introduces some distortion in the extended band. However, such distortion at frequencies above 3.4kHz is typically inaudible and does not significantly degrade the quality of extended WB speech. This is assuming that spectral envelope estimation works reasonably well [16]. The spectrum $U_t^{\mathrm{M}}(f)$ is filtered by a HPF with a cut-off frequency of 3.4kHz to extract HB excitation $\hat{\mathbf{u}}_t^{\mathrm{HB}}$ which is then added to appropriately delayed (D) NB excitation $\mathbf{u}_t^{\mathrm{NB}}$ to give the extended WB excitation $\hat{\mathbf{u}}_t^{\mathrm{WB}}$.

Step 3 – Time domain resynthesis:

In the final step, the estimated WB excitation $\hat{\mathbf{u}}_t^{\mathrm{WB}}$ is filtered using a synthesis filter defined by $\hat{g}_t^{\mathrm{WB}}$ and $\hat{\mathbf{a}}_t^{\mathrm{WB}}$ in order to resynthesise the extended WB speech frame $\hat{\mathbf{s}}_t^{\mathrm{WB}}$. The extended WB speech signal is then obtained with an overlap and add (OLA) method.

In most real-time speech applications including ABE, the speech signal is processed frame-by-frame, where each frame is processed with the application of an appropriate window function. The desired modifications or transformations in the speech signal are achieved via an analysis-modification-synthesis operation where *analysis* corresponds to the transformation from the time to the frequency-domain, e.g. via STFT, and *synthesis* involves the reconstruction of the corresponding modified time-domain signal. In case of no modifications, the analysis-synthesis operation should give *perfect reconstruction* which is possible only if the analysis window $a[n]$ and the synthesis window $s[n]$ with length M for hop size or frame shift L satisfy:

$$\sum_l a[n - lL]s[n - lL] = 1 \ \forall n \in \mathbb{Z} \tag{3.11}$$

This property is referred to as the *overlap-add* (OLA) constraint which is necessary for perfect reconstruction [216, Section 12.1.1].

If there is no synthesis window, then Eq. 3.11 is written as:

$$\sum_l a[n - lL] = \sum_l w_{PR}[n - lL] = 1 \ \forall n \in \mathbb{Z} \tag{3.12}$$

where w_{PR} represents a window function that satisfies the perfect reconstruction property.

In practice, the sequence of overlapping window functions (e.g., Hann or Hamming) sum up, according to Eq 3.12, to a constant value $K \neq 1$. Therefore, given a window function w_K, w_{PR} can be designed as follows:

$$w_{PR} = \frac{w_K}{K} \tag{3.13}$$

where K is given by [217, Section 5.3.1]

$$K = \frac{1}{L} \sum_{m=0}^{M-1} w_K[m].$$

When the same analysis and synthesis windows are used, we have:

$$a[n] = s[n] = \sqrt{w_{PR}} = \sqrt{\frac{w_K}{K}}$$

In this work, w_K is the Hann window.

3.2 Databases

Standard databases are chosen for the training and evaluation of the baseline ABE algorithm. The same databases are also used for other experiments reported in the thesis. They are all discussed below.

3.2.1 TIMIT

The TIMIT database [218] consists of 6300 utterances recorded at a sampling rate of 16kHz. They were produced by 630 speakers from 8 major dialect regions of the United states, all of whom contribute 10 utterances each. The text material of the TIMIT speech corpus consists of:

- 2 dialect (referred to as SA) sentences – spoken by each speaker thereby producing 1260 utterances,

- 450 phonetically balanced (referred to as SX) sentences – each sentence spoken by 7 different speakers thus producing 3150 utterances and

- 1890 phonetically-diverse (referred to as SI) sentences – each sentence is spoken only by one speaker thereby producing 1890 utterances.

The TIMIT corpus is typically divided into two partitions: a training set and a complete test set. The SA dialect sentences are usually removed from both subsets. The training set consists of 3696 utterances spoken by 462 speakers. The complete test set consists of 1344 utterances spoken by 168 speakers. The complete test set consists of a subset which represents core test subset. It consists of 192 unique utterances spoken by 24 speakers (2 male and 1 female speaker from each dialect).

3.2.2 TSP speech database

The TSP speech database [219] consists of 1378 utterances, recorded at a sampling rate of 48kHz. It is collected from 12 male and 12 female speakers. The text material is a subset of the Harvard sentences [220] which are grouped into 72 lists of 10 sentences each. 60 utterances (from 6 lists) are spoken by each speaker. Speech data is recorded mostly in Canadian English with a small portion covering other dialects of English[11].

3.2.3 CMU-Arctic database

The CMU Arctic database [221] is a set of single-speaker speech databases, recorded at a sampling rate of 32kHz. Each consists of 1132 phonetically balanced utterances collected from each of the three (two male and one female) English speakers. Each speech signal is recorded with an additional, parallel electroglottograph (EGG) or laryngograph signal. EGG signals capture glottal activity during production of the corresponding speech signal. Additional information which includes phonetic labels and pitch markers is also available for every speech file. The database is used widely in speech synthesis research [222].

[11]Some speech files are missing or they were overwritten during data collection, leading to 1378 files.

3.2.4 3GPP database

The 3GPP database provides test signals which are commonly used for the objective evaluation of speech quality in telephonometry. 4 phonetically balanced utterances collected from 4 English speakers with a sampling rate of 48kHz were chosen from the 3GPP database, details of which can be found in annexure B and clause 7.3 of ITU-T recommendation P.501 [223].

3.3 Data pre-processing and distribution

ABE training, validation and testing is performed using parallel NB and WB speech signals obtained from speech databases recorded at sampling rates of 16kHz or higher.

3.3.1 Data pre-processing

Data pre-processing steps are illustrated in Fig. 3.4. First, WB signals $\mathbf{x}_{\text{wb}}$ are obtained by downsampling[12] to 16kHz if the sampling rate is higher than 16kHz (e.g. TSP speech database). Using software provided as a part of the ITU-T software tool library (STL)[13] [225] WB speech signals are processed in order to simulate the effect of a telephone channel. Mobile station input (MSIN) characteristics are simulated using a highpass filter with an approximate cut-off frequency of 195 Hz [226] followed by level adjustment to an active speech level of -26 dBov (dBov represents dBs relative to the overload point of a recording system [227]). NB signals $\mathbf{x}_{\text{nb}}$ are then obtained by downsampling to 8kHz followed by lowpass filtering with a cut-off frequency of 3.4kHz. Delays due to filtering are adjusted so as to obtain time-aligned, parallel NB and WB speech signals.

[12]Downsampling was performed using the ResampAudio tool contained in the AFsp package [224].

[13]The software tool library (STL) can be found at:
`https://github.com/openitu/STL` or
`https://www.itu.int/rec/T-REC-G.191-200509-S/en` (Last accessed : March, 2019)

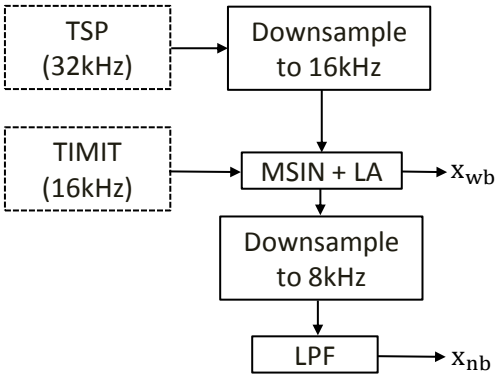


Figure 3.4: Data pre-processing protocol used for ABE. LA = level alignment to -26 dBov. MSIN = mobile station input filtering.

3.3.2 Training, validation and test data

ABE algorithms presented in Chapters 4, 5 and 6 used the TIMIT and the TSP speech databases for training, validation and testing according to the following:

- The work presented in Chapter 4 used the TIMIT training set for statistical modelling via GMM (Section 3.1.1) and the complete test set was used for evaluation.

- In order to increase the quantity of data necessary for the training of neural networks, the work presented in Chapters 5 and 6 used training data pooled from the TIMIT training set and 1152 utterances from the complete test set (excluding the core test subset). Network optimisation and validation is performed using the TIMIT core test subset. The acoustically-different TSP speech database was used for testing.

3.4 Performance assessment

Performance improvement of ABE approaches proposed in this thesis work in comparison to the baseline ABE algorithm is measured via subjective as well as objective assessment techniques. MI is employed as a information theoretic tool to quantify the correlation between NB and HB features.

3.4.1 Subjective assessment

A form of comparison category rating (CCR) approach, a type of listening-only test, was used for all work reported here for subjective assessment of speech quality. Refer to Section 2.7 for further details.

3.4.2 Objective assessment metrics

Objective measures of speech quality used in the work reported in this thesis include: (i) root mean square log spectral distance (RMS-LSD), (ii) COSH distortion and (iii) perceptual evaluation of speech quality (PESQ). They are described in the following.

Root mean square log spectral distance (RMS-LSD):

The log spectral distance (LSD), also referred to in the literature as the log spectral distortion or deviation, is a distance metric that is defined by the difference between the spectra of two signals on a logarithmic scale. The spectrum of a signal can be calculated using either the FT or an all-pole smoothed LP spectrum. The latter allows closer examination of the distortion measure [228, Section 4.5.1].

The RMS-LSD between two power spectra $P(f) = g^2/|A(f)|^2$ and $\hat{P}(f) = \hat{g}^2/|\hat{A}(f)|^2$ is defined as the L_2 norm of the log spectral distance $V(f) = \ln\left(\frac{P(f)}{\hat{P}(f)}\right)$ as follows:

$$d_{\text{RMS-LSD}}(P, \hat{P}) = \left[\frac{1}{\Delta F}\int_{\Delta F} |V(f)|^2 df\right]^{\frac{1}{2}} \tag{3.14}$$

where $\triangle F$ is the frequency range of the estimated HB components over which the distance measure is calculated. $A(f)$ and $\hat{A}(f)$ represent the LP inverse filters of the original and estimated WB speech frames; g and $\hat{g}$ are their respective LP gains. Eq. 3.14 should be multiplied with factor $10/\ln(10) = 4.34$ to obtain the L_2 measure in decibels (dB). In practice, the integral in Eq. 3.14 with normalisation factor $\triangle F$ is approximated using the sample mean.

Since the perception of signal loudness is approximately logarithmic, the $d_{\text{RMS-LSD}}$ is a perceptually relevant distortion measure for the subjective assessment of sound difference. However, it weights both positive or negative logarithmic differences $|V(f)|$ equally along the frequency axis and therefore, does not take

into account the perceptual masking effect [228, Section 4.5.1].

COSH distortion:

COSH distortion is defined as:

$$\begin{aligned} d_{\text{COSH}}(P,\hat{P}) &= \frac{1}{2}\left[d_{\text{IS}}(P,\hat{P}) + d_{\text{IS}}(\hat{P},P)\right] \\ &= \frac{1}{\triangle F}\int_{\triangle F}\left[\cosh(V(f)) - 1\right] df \end{aligned} \tag{3.15}$$

where the Ikatura-Saito(IS) distortion (d_{IS}) is given by:

$$\begin{aligned} d_{\text{IS}}(P,\hat{P}) &= \frac{1}{\triangle F}\int_{\triangle F}\left[e^{V(f)} - V(f) - 1\right]^2 df \\ &= \frac{1}{\triangle F}\int_{\triangle F}\left[\frac{P(f)}{\hat{P(f)}} - \ln\frac{P(f)}{\hat{P}(f)} - 1\right]^2 df \end{aligned} \tag{3.16}$$

Eqs. 3.15 and 3.16 must also be multiplied with a factor of 4.34 to convert d_{COSH} and d_{IS} measures to units of dB.

The COSH distortion metric weights the larger log spectral differences more than the RMS-LSD distortion metric. Heavier weighting of larger deviations in the spectrum, especially around formant regions, is important in speech processing [229] and therefore the COSH measure, a symmetric version of the IS distortion, has a perceptual relevance.

Perceptual evaluation of speech quality (PESQ):

In order to evaluate end-to-end speech quality of narrowband speech codecs and telephone networks, ITU-T has standardized an objective measure in ITU-T Rec. P.862 [188]. The measure is referred to as PESQ. It is based on a perceptual model that includes transformations of both original and degraded signals to an internal representation that takes into account perceptual frequency (Bark) and loudness (Sone). It tries to match psychophysical representations of audio signals in the human auditory system. Additionally, PESQ takes into account the filtering operations and delays involved in the communication systems along with distortions created by the channels and low bit-rate codecs. For 22 unknown ITU benchmark experiments and 8 other validation experiments (that were not

used for PESQ development), the PESQ score was found to exhibit the average correlation of 0.935 with subjective scores [188]. The raw scores provided by the perceptual model of PESQ[14] are mapped to the MOS-objective listening quality (denoted by MOS-LQO) estimates according to a mapping function standardized in ITU-T Rec. P.862.1 [230]. The MOS-LQO estimates are comparable to the subjective scores (MOS-LQS) obtained using ACR listening tests.

A WB extension to the PESQ model, referred to as WB-PESQ, is presented in ITU-T Rec. P.862.2 [189] for the assessment of WB telephone networks and codecs. Quality estimates obtained using the WB-PESQ model represent, or are equivalent to, mean opinion scores obtained when a listener uses WB headphones. Throughout this thesis the WB-PESQ is employed to obtain objective MOS estimates for artificially extended WB signals and are denoted by (MOS-LQO$_{\text{WB}}$).

As reported in [194, 195] MOS-LQO$_{\text{WB}}$ estimates fail to give reliable estimates of speech quality. In particular, rank orders of *different* ABE algorithms under evaluation were not predicted reliably. ABE systems under investigation in this work are different variants of the *same* ABE algorithm. They differ only in use of front-end features to the input of GMM regression (GMMR) model, with other processing blocks remaining unchanged. Therefore, it is assumed that speech quality estimates MOS-LQO$_{\text{WB}}$ obtained using WB-PESQ provide a meaningful tool in our experimental setup.

3.4.3 Mutual information assessment

ABE algorithms estimate missing HB frequency components from available NB components based on an assumption that the NB and HB frequency components of a speech signal exhibit correlation. In the ABE literature, for given NB and HB feature sets, the reliability of estimation performance is often measured in terms of mutual information (MI) between the HB and the NB components of a speech frame. MI quantifies the benefit of one variable for the estimation of the other. It reflects both linear and non-linear dependencies between two random variables [78, Section 4.3.3]. This is in contrast to linear correlation which only

[14]The PESQ scores defined in ITU-T Rec. P.862 provide raw scores in the range of -0.5 to 4.5. These raw scores are mapped to the objective estimates MOS-LQO in the range of 1.02 to 4.56 that correspond to the subjective scores of ACR listening-only tests. The mapping function is trained on a large corpus of test samples collected from voice over Internet Protocol (VoIP), wireless applications.

takes into account the linear dependencies. This section describes estimation of MI via GMM modelling as a standard information theoretic approach to quantify the correlation between NB and HB components.

The mutual information $I(\mathbf{x};\mathbf{y})$ between two continuous random variables $\mathbf{x}$ and $\mathbf{y}$ with the joint PDF $p(\mathbf{x},\mathbf{y})$ and the marginal PDFs $p(\mathbf{x})$ and $p(\mathbf{y})$ is defined as the relative entropy or Kullback-Leibler distance (D_{KL}) between their joint distribution $p(\mathbf{x},\mathbf{y})$ and the product distribution $p(\mathbf{x})p(\mathbf{y})$ [231]:

$$I(\mathbf{x};\mathbf{y}) = \int\int p(\mathbf{x},\mathbf{y}) \log_2 \left(\frac{p(\mathbf{x},\mathbf{y})}{p(\mathbf{x})p(\mathbf{y})} \right) d\mathbf{x}d\mathbf{y} \tag{3.17}$$

$$= D_{KL}(p(\mathbf{x},\mathbf{y}), p(\mathbf{x})p(\mathbf{y})) \tag{3.18}$$

The true distributions $p(\mathbf{x})$, $p(\mathbf{y})$ and $p(\mathbf{x},\mathbf{y})$ are unknown in practice and usually approximated with a GMM. Therefore, if $p(\mathbf{x},\mathbf{y})$ takes the form of a Gaussian mixture model (GMM), then Eq. 3.17 can be written as an expectation approximated by the sample mean over L data vectors $(\mathbf{x}_t,\mathbf{y}_t)$ as follows:

$$I(\mathbf{x};\mathbf{y}) \approx \frac{1}{L}\sum_{t=1}^{L} \log_2 \left(\frac{p(\mathbf{x}_t,\mathbf{y}_t)}{p(\mathbf{x}_t)p(\mathbf{y}_t)} \right) \tag{3.19}$$

Eq. 3.19 can be used to estimate the MI between NB and HB components of speech parametrised with features $\mathbf{x}$ and $\mathbf{y}$ [171, 172] respectively.

In Chapter 4, MI is estimated using the TIMIT dataset (excluding dialect (SA) sentences). First, the WB speech signals are processed according to the steps explained in Section 3.3.1 and then they are processed with some form of feature extraction to give NB and HB features $\mathbf{x}$ and $\mathbf{y}$ respectively. Mean and variance normalisation (MVN) is applied to all features vectors. All signals are processed in frames of 20ms duration with 10ms overlap to give approximately 1.52×10^6 feature vectors. All speech frames are windowed using square root Hann window (refer to OLA processing described in Section 3.1.3). A 128 component, full-covariance GMM is learned on data vectors to obtain estimates of the densities used in Eq. 3.19. In Chapter 5, the MI assessment is performed using TSP speech database over approximately 3.19×10^5 feature vectors.

Chapter 4

ABE with explicit memory inclusion

Artificial bandwidth extension (ABE) algorithms have been developed to improve speech quality when wideband devices are used in conjunction with narrowband devices or infrastructure. This chapter introduces the concept of explicit memory and how it can be utilised efficiently to improve ABE performance. It draws upon past work which shows how memory can be included via delta features under the constraint of fixed dimensionality. While contextual information or memory for ABE is beneficial, a quantitative analysis of the relative benefit of *explicit* memory inclusion is presented. Its potential is investigated via the use of a standard information theoretic approach. Findings are validated through objective and subjective assessments of an ABE system which uses memory with only negligible increases to latency and computational complexity. Listening tests results are reported which show that narrowband signals whose bandwidth is artificially extended with, rather than without the inclusion of memory, are of consistently improved quality.

Section 4.1 discusses memory inclusion for ABE. Section 4.2 briefly describes the past work which includes memory via the use of dynamic/delta features. Section 4.3 reports an investigation of the benefit to ABE of using different degrees of contextual information or explicit memory in the form of static features obtained from neighbouring speech frames. Section 4.4 describes modifications in the baseline ABE algorithm (presented in the previous chapter) in order to include memory using a dimensionality reduction transform. Section 4.5 discusses the experimental setup with objective and subjective assessment results. An information theoretic analysis is also presented. It takes the form of the mutual information between NB and HB features.

4.1 Memory inclusion for ABE

ABE algorithms exploit the correlation between NB and HB frequency components of speech. This correlation is learned via the statistical modelling of NB and HB components obtained from WB training data. For ABE methods based on classical source-filter model, the HB component is usually parametrised with some form of linear prediction (LP) coefficients whereas the NB component can be parameterised by a variety of static and/or dynamic spectral estimates. Use of contextual information or memory, obtained from neighbouring frames, is common to speech processing applications, including ABE. Speech is afterall a signal with dynamic temporal and spectral properties.

In addition to being captured through front-end features, dynamic information, or *memory* can also be captured with specific back-end regression models that can model inter-frame dependencies (refer to Section 2.6.2). Front-end memory is captured in the form of either well-known delta coefficients, commonly referred to as dynamic features, or static features extracted from neighbouring speech frames. The later form, is referred as explicit memory throughout this work. The first attempt to quantify the importance of front-end memory inclusion is reported in [85, 103, 104]. The work quantifies the importance of memory inclusion in the form of delta features without affecting the complexity of a standard regression model. Memory inclusion is performed either by replacing or appending higher order static feature coefficients with lower order dynamic delta coefficients. This is achieved for either only NB features or both NB and HB features. In the latter scenario, the work in [15] reports a significant improvement in MI, however, the lower order delta features are non-invertible and should be discarded during reconstruction of HB components, thus resulting in a loss of information. Therefore, the inclusion of memory necessitates the loss of informative higher-order static HB features in order to accommodate dynamic delta features.

The use of higher number of neighbouring speech frames in order to extract this extra informations leads to increased latency (because the use of future frames introduce look-ahead delay) which is not suitable for ABE, which is a real-time application. Additionally, it increases the computational complexity of the regression model. While a body of the previous work on memory inclusion via delta features points towards the importance of dynamic information to ABE, it raises the questions of what degree of contextual or explicit memory information is of

benefit and how can it be harnessed without increasing latency and computational complexity. Therefore, a quantitative analysis of explicit memory and its inclusion to improve performance of a fixed ABE algorithm is focus of the research work presented in this chapter.

4.2 Brief overview of memory inclusion for ABE via delta features: Past work

This section describes previous work which reported the benefit to ABE of memory inclusion via delta features. Memory inclusion using delta features can be implemented:

- by appending delta coefficients to the existing static coefficients for NB and/or HB features (referred to as scenario A), this increases the complexity of the associated regression model due to increase in dimensionality of the features;
- by substituting a subset of higher-order static coefficients with the delta coefficients of the remaining lower-order static coefficients (referred to as scenario S), in this case, the dimensionality of the resulting feature set is preserved.

The features resulting from such memory inclusion are referred to as dynamic (static+delta) features.

4.2.1 Memory inclusion scenarios

The memory inclusion via delta features can be performed for both NB and/or HB features. This can be done in four ways represented by scenarios A1, A2, S1 and S2, which are explained in the following [15, 4.4.3.1]. Given d_{nb}-dimensional NB features $\mathbf{x}^{\text{NB}}$, d_{hb}-dimensional HB features $\mathbf{y}^{\text{HB}}$ and $\boldsymbol{\Delta}\mathbf{x}^{\text{NB}}$ and $\boldsymbol{\Delta}\mathbf{y}^{\text{HB}}$, their respective, corresponding delta features, the resulting dynamic spaces for the 4 possible memory inclusion cases are represented by the following features:

Scenario A1: $\mathbf{x}_{\Delta}^{\text{NB}} = [\mathbf{x}^{\text{NB}}\ \Delta\mathbf{x}^{\text{NB}}]^T \in \mathcal{R}^{2\times d_{\text{nb}}}$ and $\mathbf{y}^{\text{HB}} \in \mathcal{R}^{d_{\text{hb}}}$

Scenario A2: $\mathbf{x}_{\Delta}^{\text{NB}} = [\mathbf{x}^{\text{NB}}\ \Delta\mathbf{x}^{\text{NB}}]^T \in \mathcal{R}^{2\times d_{\text{nb}}}$ and $\mathbf{y}_{\Delta}^{\text{HB}} = [\mathbf{y}^{\text{HB}}\ \Delta\mathbf{y}^{\text{HB}}]^T \in \mathcal{R}^{2\times d_{\text{hb}}}$

Scenario S1 : $\mathbf{x}_{\Delta}^{\mathrm{NB}} = [\mathbf{x}_{\mathrm{lower}}^{\mathrm{NB}}\ \Delta\mathbf{x}_{\mathrm{lower}}^{\mathrm{NB}}]^T \in \mathcal{R}^{d_{\mathrm{nb}}}$ and $\mathbf{y}^{\mathrm{HB}} \in \mathcal{R}^{d_{\mathrm{hb}}}$

Scenario S2 : $\mathbf{x}_{\Delta}^{\mathrm{NB}} = [\mathbf{x}_{\mathrm{lower}}^{\mathrm{NB}}\ \Delta\mathbf{x}_{\mathrm{lower}}^{\mathrm{NB}}]^T \in \mathcal{R}^{d_{\mathrm{nb}}}$ and $\mathbf{y}_{\Delta}^{\mathrm{HB}} = [\mathbf{y}_{\mathrm{lower}}^{\mathrm{HB}}\ \Delta\mathbf{y}_{\mathrm{lower}}^{\mathrm{HB}}]^T \in \mathcal{R}^{d_{\mathrm{hb}}}$

where lower coefficients correspond to the first few coefficients of the feature vectors $\mathbf{x}^{\mathrm{NB}}$ and $\mathbf{y}^{\mathrm{HB}}$. Features $\mathbf{x}_{\Delta}^{\mathrm{NB}}$ and $\mathbf{x}_{\Delta}^{\mathrm{HB}}$ represent dynamic NB and HB features respectively.

While A1 and S1 cases involve dynamic features only for NB, they are obtained for both NB and HB in cases A2 and S2. A brief overview of results obtained using these memory inclusion scenarios are discussed in sections 4.2.3 and 4.2.4.

4.2.2 Highband certainty

In order to quantify the usefulness of memory inclusion via delta features (represented by four memory inclusion cases A1, A2, S1, S2), in the work presented in [15], the correlation between NB features $\mathbf{x}^{\mathrm{NB}}$ and HB features $\mathbf{y}^{\mathrm{HB}}$ is quantified in terms of an information-theoretic measure of highband certainty (as proposed in [171]). It is given by the ratio of the mutual information between $\mathbf{x}^{\mathrm{NB}}$ and $\mathbf{y}^{\mathrm{HB}}$ to the discrete entropy of the HB representation $\mathbf{y}^{\mathrm{HB}}$:

$$C(\mathbf{y}^{\mathrm{HB}}|\mathbf{x}^{\mathrm{NB}}) = \frac{I(\mathbf{x}^{\mathrm{NB}};\mathbf{y}^{\mathrm{HB}})}{H(\mathbf{y}^{\mathrm{HB}})}$$

In the cases, A1 and S1, of incorporating memory into NB features only, the change in HB certainty is given by [15, 4.4.3.1]:

$$\begin{aligned}\Delta C_1 &= C(\mathbf{y}^{\mathrm{HB}}|\mathbf{x}_{\Delta}^{\mathrm{NB}}) - C(\mathbf{y}^{\mathrm{HB}}|\mathbf{x}^{\mathrm{NB}}) \\ &= \frac{I(\mathbf{x}_{\Delta}^{\mathrm{NB}};\mathbf{y}^{\mathrm{HB}})}{H(\mathbf{y}^{\mathrm{HB}})} - \frac{I(\mathbf{x}^{\mathrm{NB}};\mathbf{y}^{\mathrm{HB}})}{H(\mathbf{y}^{\mathrm{HB}})} \\ &= \frac{1}{H(\mathbf{y}^{\mathrm{HB}})}[I(\mathbf{x}_{\Delta}^{\mathrm{NB}};\mathbf{y}^{\mathrm{HB}}) - I(\mathbf{x}^{\mathrm{NB}};\mathbf{y}^{\mathrm{HB}})] \end{aligned} \tag{4.1}$$

It means that in case of memory inclusion for NB features only, the change in certainty is dependent only on the change in MI.

In second scenario of memory inclusion in both NB and HB, the change in HB certainty is much more complex (in this case the change also depends upon the

change in the entropy of the HB itself) and given by:

$$\begin{aligned}\Delta C_2 &= C(\mathbf{y_\Delta}^{\mathrm{HB}}|\mathbf{x}_\Delta^{\mathrm{NB}}) - C(\mathbf{y}^{\mathrm{HB}}|\mathbf{x}^{\mathrm{NB}}) \\ &= \frac{I(\mathbf{x}_\Delta^{\mathrm{NB}};\mathbf{y}_\Delta^{\mathrm{HB}})}{H(\mathbf{y}_\Delta^{\mathrm{HB}})} - \frac{I(\mathbf{x}^{\mathrm{NB}};\mathbf{y}^{\mathrm{HB}})}{H(\mathbf{y}^{\mathrm{HB}})}\end{aligned} \tag{4.2}$$

4.2.3 Analysis and results

The importance of memory inclusion via delta features was studied and analysed in [15, Section 4.3.3.2] in terms of its effect on the change in the high band certainty. The analysis was presented for two feature representations, namely, line spectral frequencies (LSFs) and Mel-frequency cepstral coefficients (MFCCs). The key results and conclusions (relevant to the work presented in this thesis) are discussed in the following.

- The improvement in HB certainty gain (ΔC_1) relative to the static HB certainty $C(\mathbf{y}^{\mathrm{HB}}|\mathbf{x}^{\mathrm{NB}})$, i.e., $\frac{\Delta C_1}{C(\mathbf{y}^{\mathrm{HB}}|\mathbf{x}^{\mathrm{NB}})}$ is very little in A1 and S1 cases. Case A1 showed modest improvement of $\approx 2.3\%$ and $\approx 5.0\%$ for LSFs and MFCCs respectively.

  Case S1 showed no improvement at all, thereby signifying the fact that *NB delta features contain less information about the static HB than do the higher-order NB static features they replace.*

- The dynamic (static+delta) features of both NB and HB are highly correlated with each other translating into the significant increase in certainty about the dynamic representation $\mathbf{y}_\Delta^{\mathrm{HB}}$ of HB given the dynamic representation $\mathbf{x}_\Delta^{\mathrm{NB}}$ of NB. This certainty is denoted by $C(\mathbf{y_\Delta}^{\mathrm{HB}}|\mathbf{x}_\Delta^{\mathrm{NB}})$. Case A2 shows $\approx 115\%$ and $\approx 99\%$ of relative improvements in the certainty gains for LSFs and MFCCs respectively. However, this improvement is achieved at the cost of increased feature dimensionality.

  In case S2 (in which the dimensionality of the features is preserved), the relative improvements fall to $\approx$ 10% and $\approx$ 78% for LSFs and MFCCs respectively. It was concluded that the MFCCs show superior correlation properties in comparison to LSFs in the context of memory inclusion via delta features under the fixed dimensionality constraint; which exhibit relatively higher certainty gains in case S2.

4.2.4 Discussion

There are some drawbacks associated with such type of memory inclusion:

- With memory inclusion via delta features under the constraint of fixed dimensionality, the dynamic HB representation takes form $\mathbf{y}_{\Delta}^{\mathrm{HB}} = [\mathbf{y}_{\mathrm{lower}}^{\mathrm{HB}}\ \Delta\mathbf{y}_{\mathrm{lower}}^{\mathrm{HB}}]$. As the delta features $\Delta\mathbf{y}_{\mathrm{lower}}^{\mathrm{NB}}$ are non-invertible, only static HB coefficients $\mathbf{y}_{\mathrm{lower}}^{\mathrm{HB}}$ are used during reconstruction of HB speech components. In contrast, all coefficients $\mathbf{y}^{\mathrm{HB}}$ are used during reconstruction when memory is not included.

  This leads to the reduced spectral resolution for HB representation, spectra now represented by fewer coefficients, especially if the HB representation is not MFCC. In case of MFCCs, if a higher number of Mel scale filters is used during MFCC extraction, the spectrum with higher resolution can still be generated using fewer, lower-order coefficients. This is because the extraction of MFCCs involves a DCT operation which attempts to compress Mel-filter energies into fewer coefficients. The truncation operation thus does not result in considerable loss of spectral information.

- Without increasing the dimensionality and therefore subsequent complexity of the ABE algorithm, only MFCC features showed significant improvement in HB certainty. This would suggest a corresponding improvement in ABE estimation performance. However, during resynthesis of bandwidth-extended speech, the achieved HB certainty gains are offset by artefacts involved inversion of HB MFCCs. This is because MFCC extraction involves lossy (non-invertible) operations such as use of the magnitude of complex spectrum, Mel-scale filter bank binning and removal/truncation of higher-order cepstral coefficients [85].

- In memory inclusion scenario S2, relative improvements[1] of $\approx 78\%$ in the HB certainty $C(\mathbf{y}_{\Delta}^{\mathrm{HB}}|\mathbf{x}_{\Delta}^{\mathrm{NB}})$ (especially for MFCC features) showed promising potential for improvements in estimation performance of HB dynamic (static+delta) features $\mathbf{y}_{\Delta}^{\mathrm{HB}}$ given NB dynamic (static+delta) features $\mathbf{x}_{\Delta}^{\mathrm{NB}}$. While delta coefficients $\Delta\mathbf{y}_{\mathrm{lower}}^{\mathrm{HB}}$ are non-invertible, only static features $\mathbf{y}_{\mathrm{lower}}^{\mathrm{HB}}$ are invertible and useful during reconstruction. Any improvement in HB certainty $C(\mathbf{y}_{\Delta}^{\mathrm{HB}}|\mathbf{x}_{\Delta}^{\mathrm{NB}})$ (i.e., the improved cross-band correlation between

[1] See Section 4.2.3.

the dynamic (static+delta) representations $\mathbf{y}_{\Delta}^{\mathrm{HB}}$ and $\mathbf{x}_{\Delta}^{\mathrm{NB}}$) is thus useful for ABE performance only if leads the improved correlation between dynamic (static+delta) NB features $\mathbf{x}_{\Delta}^{\mathrm{NB}}$ and static HB features $\mathbf{y}_{\mathrm{lower}}^{\mathrm{HB}}$. This means that the HB certainty $C(\mathbf{y}_{\mathrm{lower}}^{\mathrm{HB}}|\mathbf{x}_{\Delta}^{\mathrm{NB}})_{\mathrm{S2}}$[2], i.e., the certainty about the static HB features $\mathbf{y}_{\mathrm{lower}}^{\mathrm{HB}}$ given the dynamic NB features $\mathbf{x}_{\Delta}^{\mathrm{NB}}$ should be increased.

Analysis in [15, Section 5.3.3.2] showed that the certainty $C(\mathbf{y}_{\mathrm{lower}}^{\mathrm{HB}}|\mathbf{x}_{\Delta}^{\mathrm{NB}})_{\mathrm{S2}}$ was lower than the certainty $C(\mathbf{y}^{\mathrm{HB}}|\mathbf{x}^{\mathrm{NB}})$ obtained with an equivalent, memoryless baseline. Therefore, further optimisations were performed in order to obtain optimal NB and HB feature dimensionality which brought modest improvements over the baseline. Additionally, the relative improvement in certainty $C(\mathbf{y}_{\mathrm{lower}}^{\mathrm{HB}}|\mathbf{x}_{\Delta}^{\mathrm{NB}})_{\mathrm{S2}}$ was only a small fraction of the certainty gain $C(\mathbf{y}_{\mathbf{\Delta}}{}^{\mathrm{HB}}|\mathbf{x}_{\Delta}^{\mathrm{NB}})$ (that was obtained for MFCCs in scenario S2). This confirmed that, while delta features help to improve cross-band correlation between dynamic features $\mathbf{x}_{\Delta}^{\mathrm{NB}}$ and $\mathbf{y}_{\Delta}^{\mathrm{HB}}$, their non-invertibility puts restrictions on achieving the equivalent improvements in correlation between features $\mathbf{x}_{\Delta}^{\mathrm{NB}}$ and $\mathbf{y}_{\mathrm{lower}}^{\mathrm{HB}}$. This results in no or very little improvements in estimation of the static HB features $\mathbf{y}_{\mathrm{lower}}^{\mathrm{HB}}$. This was further confirmed by modest improvements in performance of the memoryless ABE system with front-end memory inclusion [15, Section 5.3.4].

4.3 Assessing the benefit of explicit memory to ABE

Speech signals are quasi-periodic in nature. During speech production, the configuration of the vocal tract and the nature of its source vary with time. Even though they are time-varying, speech signals can be considered stationary over short durations, typically 20-30ms. This is because it is assumed that the vocal tract changes its characteristics relatively slowly over such short-time interval. Transfer function of the vocal tract filter thus can be assumed to be fixed or nearly fixed. Therefore, speech signals are processed in terms of short duration frames after applying appropriately chosen sliding windows[3]. Windows slide at a frame

[2] $C(\mathbf{y}^{\mathrm{HB}}|\mathbf{x}_{\Delta}^{\mathrm{NB}})_{\mathrm{S2}}$ represents the HB certainty calculated in scenario S2 which involves modelling of joint-dynamic feature space $[\mathbf{x}_{\Delta}^{\mathrm{NB}}\ \mathbf{y}_{\Delta}^{\mathrm{HB}}]^T$ using a GMM. This is in contrast to the certainty estimations in scenarios A1 and S1 that involve GMM modelling of joint space $[\mathbf{x}_{\Delta}^{\mathrm{NB}}\ \mathbf{y}^{\mathrm{HB}}]^T$.

[3] Based on their specific shape, different windows provide different time and frequency resolution properties. They are characterised by different main lobe and side lobe structure, e.g., the

interval (typically 5-10ms) that is sufficient to follow the changing events in speech signals [10, Section 5.2].

ABE algorithms estimate HB frequency components from available NB components. As the dynamics of the speech signal vary slowly, the HB frequency components of a speech frame exhibit a significant correlation with the NB frequency components in surrounding speech frames. In order to verify this hypothesis, the quantification and study of the correlation via the standard MI measure[4] is presented in this section.

4.3.1 Analysis

The estimation and analysis of MI requires a choice of features. Approaches to ABE use different features for NB and HB spectral content. Due to the ease in time domain reconstruction, the most widely used HB features are LP coefficients [81], line spectral frequencies (LSFs) [83] and cepstral coefficients [16]. Few algorithms use Mel-frequency cepstral coefficients (MFCCs) [85]. Study reported in [172] investigated the selection of front-end NB features based on information theoretic measures. It is suggested that the features offering maximal MI and class separability offer potential to improve ABE performance.

For analysis of the MI reported in this chapter, three feature representations are chosen for NB frequency components. They include log-Mel filter energy (logMFE) coefficients, LP coefficients and autocorrelation coefficients (ACs). LP coefficients are chosen for HB features. Each WB speech frame obtained from the entire TIMIT dataset (excluding dialect (SA) sentences) is processed (refer to Section 3.3.1 for data pre-processing details) in order to extract NB and HB features respectively, which form the vector pairs $(\mathbf{x}_t, \mathbf{y}_t)$ for MI estimation according to Eq. 3.19 (Section 3.4.3). Mean and variance normalisation (MVN) is applied to all feature vectors before MI estimation.

rectangular window has a narrower main lobe structure than the Hamming window, but the higher side lobe structure.

[4]Note that the work presented in [103, 171] suggested that in context of ABE, the MI, $I(\mathbf{x}^{\mathrm{NB}}; \mathbf{y}^{\mathrm{HB}})$ alone is not sufficient rather the highband certainty, $\frac{I(\mathbf{x}^{\mathrm{NB}}; \mathbf{y}^{\mathrm{HB}})}{H(\mathbf{y}^{\mathrm{HB}})}$, gives a more relevant measure that quantifies cross-band dependence. This ratio quantifies the certainty about the chosen HB parametrisation given a NB parametrisation. The work throughout this thesis employs only LP coefficients for HB feature representations, thus, making the MI measure equivalent to the HB certainty.

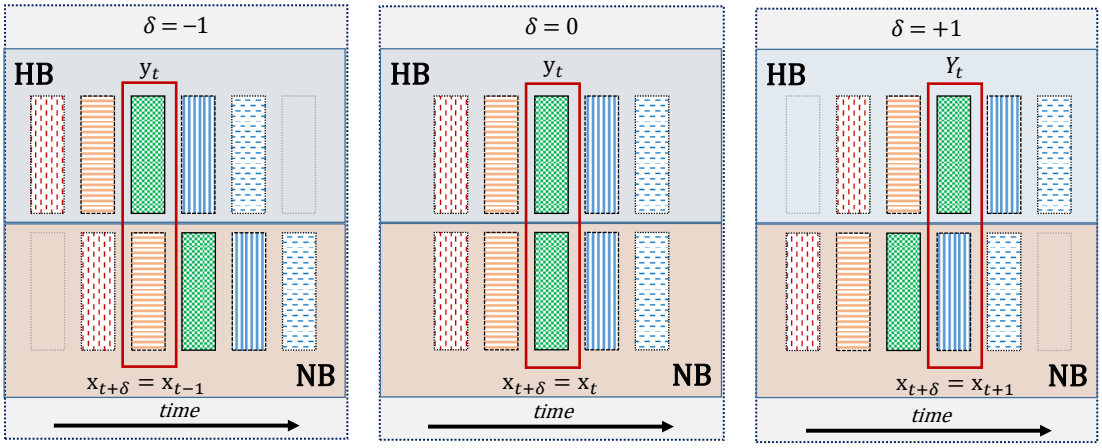


Figure 4.1: An Illustration of mutual information (MI) estimation with contextual information from neighbouring frames. Vertical bars represent NB (bottom) and HB (top) feature vectors. Red boxes represent the pair of NB ($\mathbf{x}_{t+\delta}, \delta = -1, 0, 1$) and HB ($\mathbf{y}_t$) components used for MI calculations.

Each WB speech frame is processed to extract 200-dimensional power spectrum (PS) coefficients P_t^{NB} corresponding to the NB frequency range. logMFE features are then calculated by applying a Mel filter bank (MFB) to P_t^{NB}. The MFB consists of 10 filters with triangular frequency responses in the frequency range 0.3-3.4kHz. The filter centre frequencies are linearly spaced according to the Mel-scale[5]. NB LP coefficient features of 10 dimensions including the gain parameter are obtained through SLP [212]. Conventional AC features consist of the first 10 normalised auto-correlation coefficients obtained by applying the IFFT to P_t^{NB}. Similarly 10-dimensional HB features are extracted through an application of SLP to HB components, also giving 9 LP coefficients and a gain parameter. Note that the LP gain obtained from LP analysis is an important property of a spectral envelope and it is related to the power of the residual error – which acts as an excitation to the vocal tract during production of a particular speech sound – and, therefore, it is included in the LP features.

Eq. 3.19 is then used with a GMM of 128 components to estimate the MI between instantaneous HB features and NB features at different time instances. This procedure is illustrated in Fig. 4.1 where $\mathbf{y_t}$ is the instantaneous HB component at time t and where $\mathbf{x}_{t+\delta}$ is the NB component at time $t + \delta$ where $\delta \in \mathbb{Z}$.

[5]Mel-scale converts frequency from linear scale (f) to logarithmic scale according to: $m(f) = 2595 \log_{10}(1+f/700)$. The mapping is approximately linear in frequency unto 1kHz and logarithmic at higher frequencies.

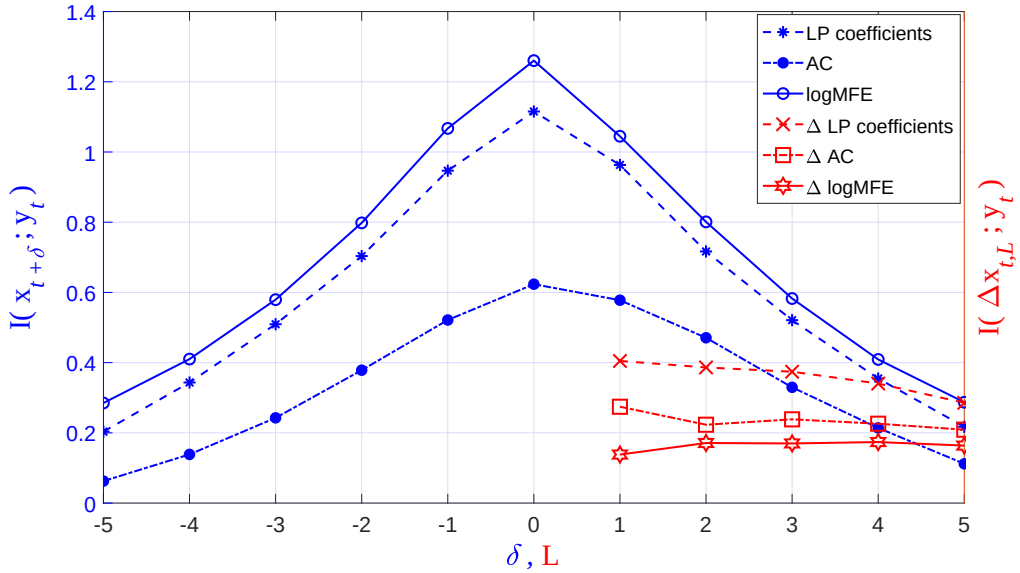


Figure 4.2: An illustration of the variation in mutual information (MI) between static highband (HB) features $\mathbf{y}_t$ and static narrowband (NB) features $\mathbf{x}_{t+\delta}$, (blue profiles) extracted from neighbouring frames and delta features $\Delta\mathbf{x}_{t,L}$ (red profiles).

Typical phonetic events span in the order of 50ms [11, Section 6.10.1]. However, some shorter sounds like plosives, stop bursts, stop onsets, releases and phone boundaries invoke more rapid spectral changes. Therefore, in this analysis we consider 1 to 5 neighbouring speech frames (i.e., 1 to 5 frames from past and future) in order to cover phonetic events within 50ms on either side of t. Another reason to confine our analysis to not more than 5 neighbouring speech frames is to avoid problem of increased latency due to higher number of future (or look ahead) frames. Next section illustrates how on average, for a given speech frame, NB features obtained from neighbouring speech frames are correlated with instantaneous HB features.

4.3.2 Findings

Blue profiles in Fig. 4.2 show the MI (vertical axis) between instantaneous HB features $\mathbf{y}_t$ and NB features $\mathbf{x}_{t+\delta}$ for $\delta \in [-5, +5]$ (horizontal axis). The three profiles correspond to logMFE, LP coefficients and AC features. As expected, for all three profiles, the MI is greatest for $\delta = 0$ for which NB and HB features are extracted from the same frame. For $\delta \neq 0$, the MI is symmetrically lower. The highest MI is obtained with logMFE coefficients and for $\delta = 1, 2$ the MI falls by

17% and 36% respectively relative to that obtained for $\delta = 0$.

Fig. 4.2 also shows the MI between static HB and dynamic or delta NB features (red profiles). Delta features or coefficients $\Delta\mathbf{x}_{t,L}$ are extracted for a frame with index t by a first-order regression, known as a time-derivative. They are calculated using linearly weighted differences between L neighbouring static feature vectors on either side of t according to:

$$\Delta\mathbf{x}_{t,L} = \frac{\sum_{l=1}^{L} l.(\mathbf{x}_{t+l} - \mathbf{x}_{t-l})}{2\sum_{l=1}^{L} l^2} \tag{4.3}$$

where $L \in [1, 5]$ (same horizontal axis in Fig. 4.2) is the number of static frames considered either side of t. The MI between static HB and delta NB features is considerably less than for static NB features. This observation corroborates the findings reported in [15, Section 5.3.3.1], namely that NB delta features are of little use to ABE; they provide comparatively little information about static HB features.

This same finding suggests that ABE algorithms should use explicit *memory*, i.e. static features extracted from neighbouring frames, instead of dynamic information captured in delta features. The two research hypotheses under investigation in this thesis are thus that:

- the inclusion of memory for ABE via front-end NB features in such a way that it should help better modelling of phonetic events or sequences, and should thus improve estimation of the HB features, thereby, giving bandwidth extended speech of enhanced quality, and
- crucial to this work, however, is that the inclusion of such additional information should not have prohibitive impacts on latency or computational complexity.

Since the aim of the work presented in this chapter is to investigate the contribution of memory to ABE and since the highest level of MI is obtained with logMFE features, they are used as NB representations for all subsequent experiments reported in this chapter. However, it is expected that the hypothesis remains valid for *any* other front-end feature as confirmed through the MI assessments illustrated in Fig. 4.2, namely the hypothesis that neighbouring static NB features exhibit correlation with instantaneous static HB features irrespective of the NB

representation. In addition, the use of energy based coefficients such as zero crossing rate, gradient index, normalised relative frame energy, local kurtosis and spectral centroid, e.g. as used in [16], is not considered here; it is assumed that their use will further enhance the performance of *any* ABE system.

4.3.3 Need for dimensionality reduction

The results presented above show that NB features obtained from neighbouring speech frames exhibit considerable correlation (measured in terms of MI) with the instantaneous HB features. This correlation decreases gradually with increase in δ on either side of t (see Fig. 4.2). The question we address here is how this extra information can be incorporated in the context of ABE. One trivial approach to explicit memory inclusion is through the concatenation of instantaneous NB features with NB static features obtained from neighbouring speech frames to obtain higher dimensional NB composite vectors (also referred to as a supervectors) that can be used for estimation of HB components.

These supervectors are further processed with a dimensionality reduction transform in order to obtain a compact, lower-dimensional NB feature vectors. This is because of the following two reasons. (1) In practice, from an information theoretic viewpoint, the MI between HB feature vectors and NB composite vectors (obtained by combination of different lower-dimensional NB feature vectors) is not the result of the simple addition of the MI between HB feature vectors and individual NB feature vectors, i.e., for given feature vectors $\mathbf{x_1}, \mathbf{x_2}, \mathbf{y}$, $I(\mathbf{x_1}, \mathbf{x_2}; \mathbf{y}) \leq I(\mathbf{x_1}; \mathbf{y}) + I(\mathbf{x_2}; \mathbf{y})$ [78, Section 4.3.3]. (2) Higher-dimensional supervectors increase the complexity of traditional regression models and thus restrict the number of primary features to be added to obtain the supervector. For example, the GMM regression technique can't handle high-dimensional data and suffers from the *curse of dimensionality* where the number of training samples required for reliable estimation of probability densities grows exponentially with the number of features [232].

The well-known and widely used dimesionality reduction transforms in the ABE literature are linear discriminant analysis (LDA) [109,113,172], and principal component analysis (PCA) [117]. LDA transform obtains the lower-dimensional feature vectors while retaining the discriminative power as much as possible. This is achieved through a linear transformation that maximises the ratio of between-class to within-class covariance of the target vector in a projected, lower-

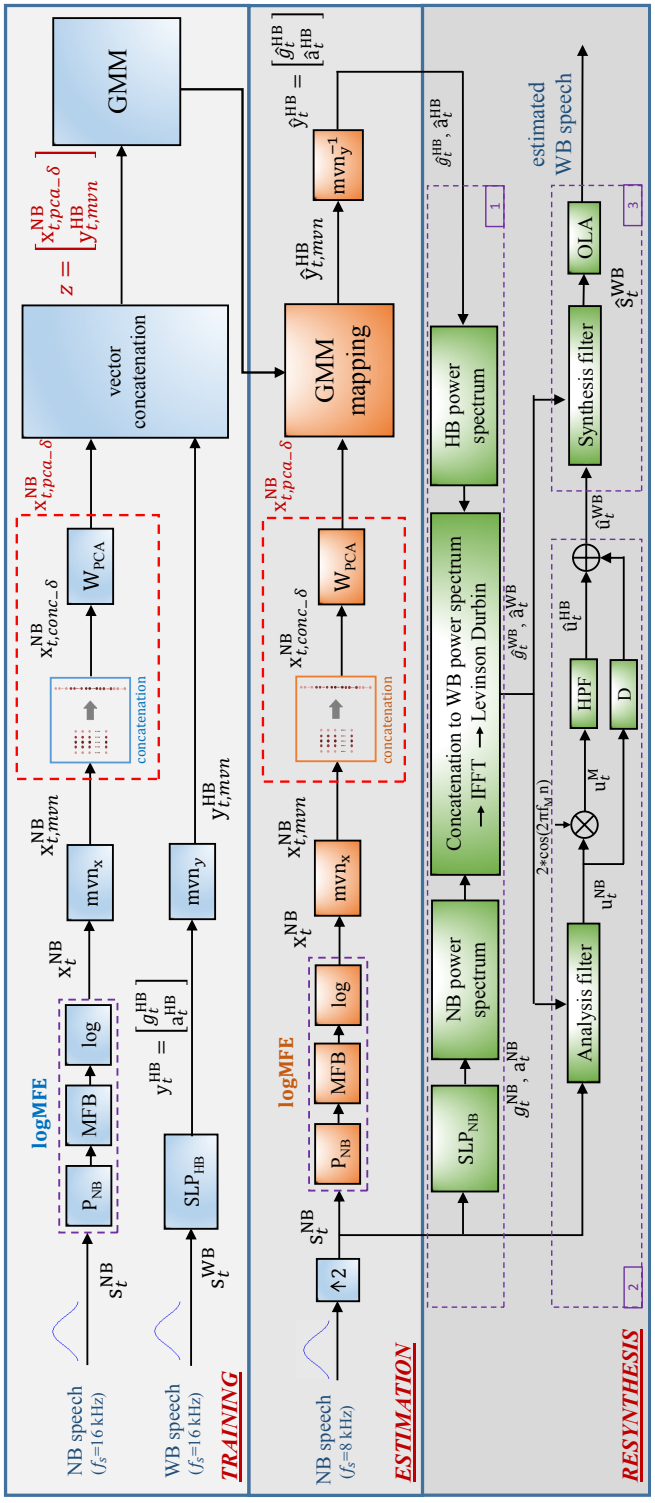


Figure 4.3: A block diagram of the artificial bandwidth extension (ABE) system with explicit memory inclusion.

dimensional space; the ratio also known as class separability. LDA involves offline learning of the transformation matrix using labelled training data, i.e., the class to which each feature vector belongs is known [172]. PCA, also known as the *Karhunen–Loève transform* (KLT), is a linear transformation in which the higher-dimensional data is projected into a lower-dimensional orthogonal space (known as the *principal subspace*) such that the variance of the projected data is maximised [214, Section 12.1]. Unlike LDA, PCA does not need labelled training data to learn the transformation matrix and is therefore the chosen dimensionality reduction technique in this chapter.

4.4 ABE with explicit memory inclusion

The ABE algorithm with memory inclusion is illustrated in Fig. 4.3. It corresponds to the baseline ABE algorithm discussed in Chapter 3 (refer to Fig. 3.1), with the following modifications:

- NB feature extraction is performed by calculation of logMFE features (purple box), and
- memory is included from neighbouring speech frames and a PCA transformation is employed in order to reduce feature dimensionality and computational complexity (red box), for both during training and estimation.

Since the full details are available in Section 3.1, only the key modifications are provided here. Each step corresponding to the three blocks of Fig. 4.3 are explained in *brief* as follows:

4.4.1 Training

Parallel NB and WB signals (both at a sampling rate of 16kHz) are processed frame-by-frame. Each NB speech frame $\mathbf{s}_t^{\mathrm{NB}}$ is processed to extract 10-dimensional logMFE features which are mean-variance normalised to obtain the NB features ($\mathbf{x}_{t,mvn}^{\mathrm{NB}}$ - top line in training block). HB features extracted from WB speech frame $\mathbf{s}_t^{\mathrm{WB}}$ consist of 9 LP coefficients $\mathbf{a}_t^{\mathrm{HB}}$ and a gain coefficient g^{NB} with mean-variance normalisation ($\mathbf{y}_{t,mvn}^{\mathrm{HB}}$ - bottom line in training block).

Memory inclusion: NB features at time t are concatenated with neighbouring

features extracted from δ frames on either side of t thus giving $10 \times (2\delta + 1)$-dimensional features:

$$\mathbf{x}_{t,conc\_\delta} = \left[\mathbf{x}^{\mathrm{NB}}_{t-\delta,mvn}, ..., \mathbf{x}^{\mathrm{NB}}_{t,mvn}, ..., \mathbf{x}^{\mathrm{NB}}_{t+\delta,mvn}\right]^T \tag{4.4}$$

In order that the complexity of subsequent steps is unaffected, principal component analysis (PCA) is applied to reduce $\mathbf{x}_{t,conc\_\delta}$ to 10-dimensional features $\mathbf{x}^{\mathrm{NB}}_{t,pca\_\delta}$. The PCA matrix $\mathbf{W}_{\mathrm{PCA}}$ is learned from training data and retained for use in the estimation step. Finally, a 128-component, full-covariance GMM is learned from the training data using joint vectors $\mathbf{z} = [\mathbf{x}^{\mathrm{NB}}_{t,pca\_\delta}, \mathbf{y}^{\mathrm{HB}}_{t,mvn}]^T$.

4.4.2 Estimation

During estimation, logMFE features are extracted from upsampled NB speech frames $\mathbf{s}^{\mathrm{NB}}_t$ to obtain features $\mathbf{x}^{\mathrm{NB}}_t$ followed by mean-variance normalisation ($\mathrm{mvn}_{\mathrm{x}}$) (using means and variances obtained from training) to produce features $\mathbf{x}^{\mathrm{NB}}_{t,mvn}$. Memory is included according to the same procedure used during training, thereby giving 10-dimensional features $\mathbf{x}^{\mathrm{NB}}_{t,pca\_\delta}$. HB features $\hat{\mathbf{y}}^{\mathrm{HB}}_{t,mvn}$ are then estimated according to:

$$\hat{\mathbf{y}}^{\mathrm{HB}}_{t,mvn} = \sum_{m=1}^{M} \frac{\pi_m \mathcal{N}(\mathbf{x}^{\mathrm{HB}}_{t,pca\_\delta}; \mu^{\mathbf{x}}_m, \Sigma^{\mathbf{xx}}_m)}{\sum_{n=0}^{M} \pi_n \mathcal{N}(\mathbf{x}^{\mathrm{HB}}_{t,pca\_\delta}; \mu^{\mathbf{x}}_n, \Sigma^{\mathbf{xx}}_n)} [\mu^{\mathbf{y}}_m + \Sigma^{\mathbf{yx}}_m \Sigma^{\mathbf{xx}^{-1}}_m (\mathbf{x}^{\mathrm{HB}}_{t,pca\_\delta} - \mu^{\mathbf{x}}_m)] \tag{4.5}$$

where π_m, μ_m, Σ_m are the GMM parameters learned during training. Inverse mean and variance normalisation (mvn^{-1}_y) is then applied to estimate HB LP coefficients $\hat{\mathbf{a}}^{\mathrm{HB}}_t$ and gain $\hat{g}^{\mathrm{HB}}_t$. For further details, refer to Section 3.1.2.

4.4.3 Resynthesis

After estimation of the HB spectral envelope parameters for a input NB speech frame, resynthesis of the extended WB signal is performed in the three distinct steps as explained in Section 3.1.3.

4.5 Experimental setup and results

This section presents an assessment of the ABE algorithm with memory inclusion in comparison to the ABE system without it. While objective and subjective

assessment results are reported in order to demonstrate improvements in quality of bandwidth extended signals, the correlation of new NB lower-dimensional representations with HB features is reported via MI measurements.

ABE experiments were performed on TIMIT dataset (see Section 3.2). While the TIMIT training set is used for GMM training, the complete test set was used for assessment (see Section 3.3.2).

4.5.1 Implementation details and baseline

The ABE algorithm with memory is denoted as M_δ where δ indicates the number of neighbouring speech frames which form the memory. The baseline algorithm which does not utilise any memory is denoted as B_1 ($= \mathrm{M}_0$). Note that the key difference between two ABE systems M_δ and B_1 is that, during training and estimation, they use features $\mathbf{x}^{\mathrm{NB}}_{t,pca\_\delta}$ and $\mathbf{x}^{\mathrm{NB}}_{t}$ features respectively, both being 10-dimensional. Therefore, the performance of the system M_δ is completely attributed to the memory inclusion performed via features $\mathbf{x}^{\mathrm{NB}}_{t,pca\_\delta}$ and not to any other system difference.

For comparison to the past work in [85] which exploits front-end memory inclusion in the form of delta features, assessment includes a second baseline, denoted B_2. System B_2 uses 5-dimensional static features appended with 5-dimensional delta features for both NB and HB parametrisations (logMFE and LP coefficients in the context of our implementation), thereby satisfying the constraint of fixed dimensionality. This is equivalent to the S2 (Section 4.2.1) memory inclusion scenario, where higher order coefficients (6 to 10) of the (10-dimensional) NB and HB features are replaced by the delta coefficients of lower order static coefficients (1 to 5). During resynthesis, delta coefficients from the estimated HB features were eliminated with only the first 5 static HB LP features being used.

All ABE algorithms were implemented with Hann windows of 20ms duration and 10ms overlap, thereby supporting perfect OLA reconstruction (Section 3.1.3). A 1024-point FFT was used for all frequency domain operations.

4.5.2 Objective assessment

Objective assessment is performed by evaluation of speech quality using the two distance metrics $d_{\text{RMS-LSD}}$, d_{COSH} and objective estimates of MOS scores MOS-LQO_{WB}

Table 4.1: Objective assessment results (with mean and standard deviation values). RMS-LSD and d_{COSH} are distance measures (lower values indicate better performance) in dB whereas MOS-LQO$_{\text{WB}}$ values reflect quality (higher values indicate better performance).

ABE method	$d_{\text{RMS-LSD}}$	d_{COSH}	MOS- LQO$_{\text{WB}}$
B_1	9.2 (1.24)	2.4 (0.66)	2.4 (0.40)
B_2	10.1 (1.22)	3.6 (1.20)	2.2 (0.37)
M_1	8.2 (0.95)	2.2 (0.64)	2.8 (0.43)
M_2	**8.1** (0.89)	**2.1** (0.65)	**2.9** (0.42)
M_3	8.2 (0.89)	2.2 (0.68)	2.8 (0.41)

(Section 3.4.2). Objective assessment results are illustrated in Tab. 4.1. While all ABE systems with memory outperform both baselines B_1 and B_2, system M_2, which uses memory contained within two neighbouring frames, performs best. Performance of system M_2 is improved (in comparison to that of system B_1) in terms of distance metrics $d_{\text{RMS-LSD}}$ and d_{COSH} by 0.9dB (9.2 $\rightarrow$ 8.1dB) and 0.3dB (2.4 $\rightarrow$ 2.1dB) showing 15% and 14% of relative improvements respectively. The MOS estimates MOS-LQO$_{\text{WB}}$ are improved from 2.4 to 2.9 signifying 21% of relative improvement. The performance of system M_3 is less, however insignificant, than that of M_2 despite it utilising relatively higher amounts of memory. This is perhaps because of the constraints imposed by the fixed-dimensionality. PCA – the employed dimensionality reduction technique – possibly fails to conserve the extra information obtained from a higher number of static, neighbouring frames into the only 10-dimensional features.

Surprisingly, baseline system B_2 gives worse performance than B_1. This is caused by the inclusion of memory through delta features with the constraint of fixed dimensionality. The latter necessitates the loss of informative higher-order static HB features in order to accommodate dynamic delta features. On account of these findings, costly, time-consuming subjective assessments were performed with systems B_1 and M_2 only.

Table 4.2: Subjective assessment results in terms of CMOS (with corresponding 95% confidence interval (CI_{95})).

Comparison B $\rightarrow$ A	CMOS [CI_{95}]
$M_2 \rightarrow$ NB	0.69 [0.42; 0.96]
$M_2 \rightarrow B_1$	0.51 [0.36; 0.66]
$M_2 \rightarrow$ WB	-0.78 [-0.97; -0.59]

Table 4.3: Mutual information assessment results. $I(\mathbf{x};\mathbf{y})$ denotes the MI between features $\mathbf{x}$ and $\mathbf{y}$.

Comparison	logMFE
$I(\mathbf{x}_t^{\text{NB}};\mathbf{y}_t^{\text{HB}})$ (System B_1)	1.24
$I(\mathbf{x}_{pca\_2}^{\text{NB}};\mathbf{y}_t^{\text{HB}})$ (System M_2)	1.34

4.5.3 Subjective assessment

Subjective assessments were performed using CCR listening tests (Section 3.4.1) in order to compare performance in terms CMOS. Tests were performed by 14 listeners who were asked to compare the quality of 14 pairs of speech signals A and B. They were asked to rate the quality of signal B with respect to A according to the CCR scale. : -3 (much worse), -2 (slightly worse), -1 (worse), 0 (about the same), 1 (slightly better), 2 (better), 3 (much better). The samples were played using DT 770 PRO headphones. Tab. 4.2 shows the CMOS results with corresponding 95% confidence interval (CI_{95}). CMOS results of 0.69 and 0.51 show that bandwidth extended speech produced by system M_2 is preferred to original NB speech and that produced by memoryless system B_1. The bandwidth-extended speech obtained relatively 0.78 CMOS points below the WB speech quality.

4.5.4 Mutual information assessment

The explicit memory inclusion presented in this chapter, from another viewpoint, can be seen as obtaining a 10-dimensional compact NB representation from additional information in the form of static NB features extracted from neighbouring speech frames. These NB features are further utilised by the underlying GMM

regression model to improve estimation of HB features. The benefit is confirmed though the findings of both objective and subjective assessments. These findings are further validated by showing the improvement in mutual information (MI) brought by the inclusion of memory. MI is estimated, between the two features sets extracted from the entire TIMIT dataset (excluding dialect (SA) sentences), according to Eq. 3.19 (Section 3.4.3).

Tab. 4.3 compares MI estimated between features $\mathbf{x}_t^{\mathrm{NB}}$ and $\mathbf{y}_t^{\mathrm{HB}}$ (denoted by $I(\mathbf{x}_t^{\mathrm{NB}}, \mathbf{y}_t^{\mathrm{HB}})$) with that between features $\mathbf{x}_{t,pca\_2}^{\mathrm{NB}}$ and $\mathbf{y}_t^{\mathrm{HB}}$ (denoted by $I(\mathbf{x}_{t,pca\_2}^{\mathrm{NB}}), \mathbf{y}_t^{\mathrm{HB}})$. NB features with memory $\mathbf{x}_{t,pca\_2}^{\mathrm{NB}}$ exhibit 8.1% higher MI relative to features $\mathbf{x}^{\mathrm{NB}}$; the MI improves from 1.24 to 1.34. These results show that the inclusion of memory results in notably higher MI; memory helps to better model missing HB information.

4.5.5 Discussion

Objective, subjective and mutual information improvements

In the past work, and as discussed in Section 4.2, the inclusion of memory via delta features improved HB certainty relatively by 78% (in S2 scenario) for dynamic (static+delta) NB and HB MFCC representations. In other words, reliability of estimation of dynamic HB features was improved significantly. However, the estimated dynamic features consist of static as well as non-invertible delta features. Latter are thus discarded during reconstruction of HB components. The improvement in HB certainty thus did not reflect into corresponding improvement in actual ABE performance.

In the work reported here, however, the new NB representation obtained via explicit memory inclusion exhibits 8% relative improvement in MI and this gain translates to improvements in speech quality of 0.51 CMOS points over memoryless ABE. Improved ABE performance in this case suggests that explicit memory inclusion successfully exploits the correlation between HB features and static NB features obtained from neighbouring frames. These findings are further validated through MI assessment results.

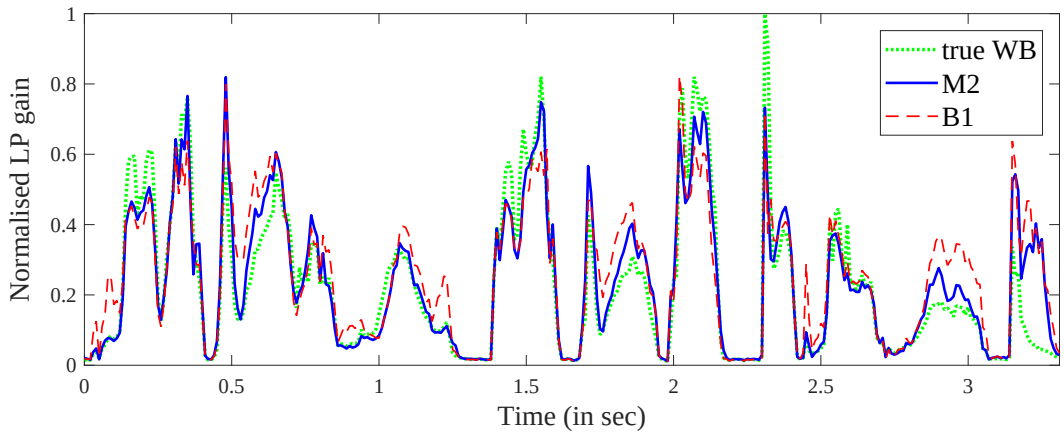


Figure 4.4: A comparison of true wideband (WB) linear prediction (LP) gain $g_{\text{true}}^{\text{WB}}$ to estimated WB LP gain $\hat{g}^{\text{WB}}$ for ABE systems M_2 and B_1. A comparison of corresponding speech spectrograms is shown in Fig 4.5.

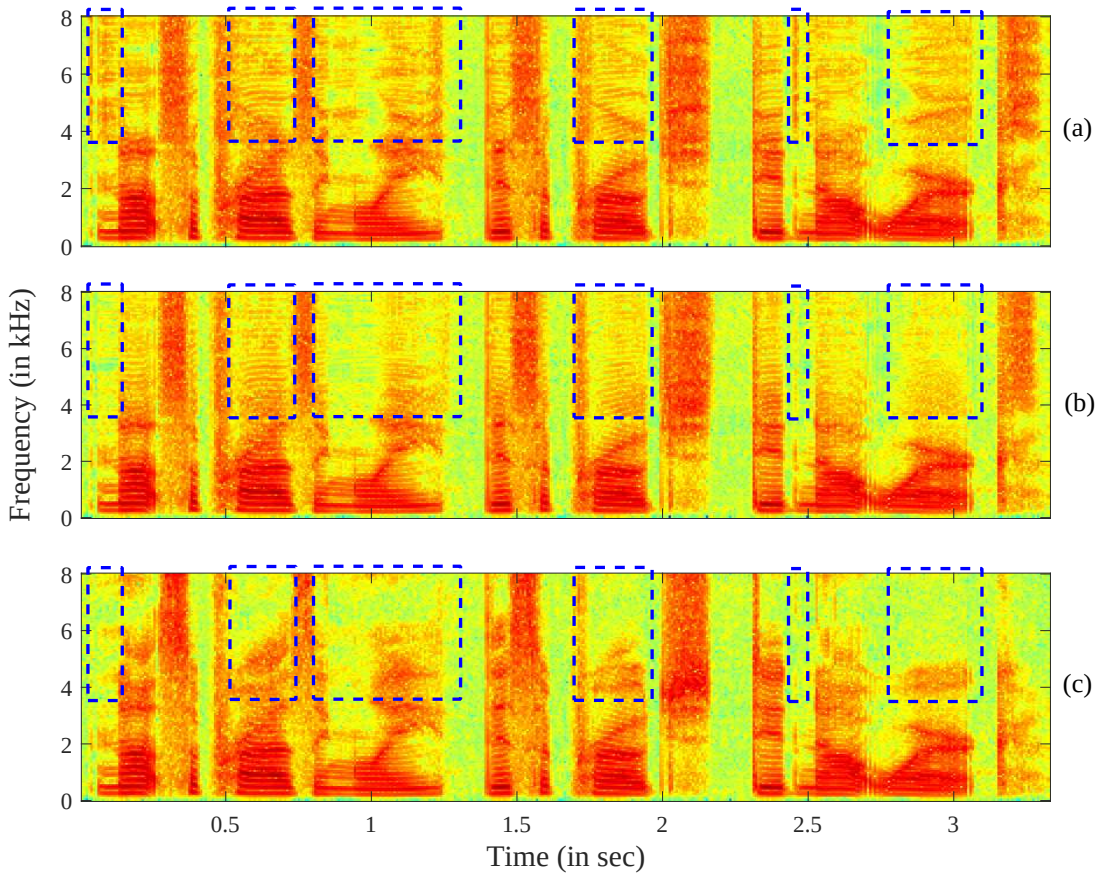


Figure 4.5: A comparison of spectrograms of wideband (WB) speech signals artificially bandwidth-extended using ABE systems (a) B_1 and (b) M_2 to that of (c) original WB speech signal. The comparison is shown for the utterance "Not surprisingly, this approach did not work" from the TIMIT test set.

Latency and complexity

Most ABE algorithms reported in the literature, especially those that exploit the complex, higher dimensional data modelling capacities of DNNs, incorporate higher number of static neighbouring features to provide input to the DNN regression model. The relative performance improvements obtained, however, are perhaps impractical due to the amount of latency (or look-ahead delay) introduced by higher number of future frames. According to [216, Section 18.4], end-to-end transmission delays in a two-way communication system should not exceed 150 ms for effective voice interaction. To meet this requirement the algorithmic delay introduced only by encoder/decoder (codec) should not exceed 20-30 ms[6].

The aforementioned improvements in quality of extended speech signals using explicit memory inclusion are achieved at the modest cost of latency and complexity. The ABE system with memory M_δ involves a look-ahead of δ frames giving rise to an algorithmic delay of δ multiplied by the frame shift. The best performing ABE system M_2 introduces a delay of 20ms which is within the practical limits. In practice, this delay can be further reduced if the ABE framing scheme matches that of the decoder.

The only factor contributing to increased complexity is the weight matrix $\mathbf{W}_{\mathrm{PCA}}$ with size 50×10 ($(2\delta+1)d_{\mathrm{nb}} \times d_{\mathrm{hb}}$) that is involved in computations for dimensionality reduction. This is manageable considering high computing power available with today's smart communication devices.

Improvemnts in gain estimation

Additional informal listening test results showed that the inclusion of memory helps to reduce processing artefacts in extended speech, thereby resulting in enhanced quality. This is perhaps because of the improved estimation of the LP gain g^{WB}. To illustrate this, a comparison of the WB LP gains (estimated using ABE systems M_2 and B_1) $\hat{g}^{\mathrm{WB}}$ and the true WB LP gain $\hat{g}^{\mathrm{WB}}_{\mathrm{true}}$ is shown in Fig. 4.4. The gain values were calculated for all frames over a speech utterance. It can be observed that the trajectory of the estimated gain with memory inclusion follows more closely that of the true gain than in case of gain estimation without memory. Improvements in gain estimation thus help to reduce overestimation or underestimation of energy

[6]Apart from delay introduced by codec, the actual end-to-end delay includes other sources of delay, e.g., delay introduced by transmission network, audio post-processing, etc.

levels of the individual speech sounds thereby helping to reduce perceived processing artefacts, improvements which are confirmed by reductions in RMS-LSD. This can also be confirmed from a comparison of respective spectrograms shown in Fig 4.5.

4.6 Summary

This chapter reports an approach to artificial bandwidth extension that incorporates explicit memory for better estimation of the missing highband speech components. The work builds upon and extends prior work that studied the benefit of capturing font-end memory via dynamic delta features; memory that can be exploited by traditional statistical models such as GMM regression. The key contributions of the work reported in this chapter are as follows.

- A thorough study and analysis of explicit memory that can captured through static features extracted from neighbouring speech frames is performed through information theoretic analyses.
- The effective inclusion of explicit memory in a memoryless ABE system using feature dimensionality reduction (under the constraint of fixed dimensionality) to improve estimation performance is demonstrated. This improvement is achieved via modest increase in complexity and an algorithmic delay of 20ms. The results show that the higher number of future frames should be avoided as they do not provide further improvements, however, lead to the increased latency.
- The potential of this approach is validated though both objective and subjective assessments. The corresponding improvements obtained with new narrowband compact representation are further validated by showing the improvement in mutual information.

The work reported in this chapter employs principal component analysis – a linear, unsupervised approach to dimensionality reduction that preserves the input variation as much as possible. This motivates us to investigate and explore other dimensionality reduction techniques designed to preserve quality rather than feature variance. In particular, deep learning based approach to dimensionality reduction is explored and presented in next chapter.

Chapter 5

ABE with memory inclusion using semi-supervised stacked auto-encoders

The utilisation of contextual information in the form of dynamic features or explicit memory captured from neighbouring frames is common to ABE research, however the use of additional cues augments complexity and can introduce latency. The previous chapter showed that unsupervised, linear dimensionality reduction technique can help to reduce complexity and latency can be reduced with the use of no more than two look-ahead speech frames. This chapter reports a semi-supervised, non-linear approach to dimensionality reduction using a stacked auto-encoder. In further contrast to the work reported in the previous chapter, it directly operates on spectral coefficients, from which low dimensional narrowband features are learned automatically in a data-driven manner. The objective speech quality measures show that the new features can be used with a standard regression model to improve ABE estimation performance. Improvements in the mutual information between learned narrowband and highband features are also observed whereas improvements in speech quality are corroborated by informal listening tests.

Section 5.1 describes the two unsupervised dimensionality reduction techniques, namely principal component analysis and stacked auto-encoders. Section 5.2 the explains need of supervision for regression tasks such as ABE and how semi-supervised stacked auto-encoder can be used to improve ABE performance under the constraint of fixed dimensionality. Section 5.3 describes experiments, training and optimisation details whereas assessment results are presented in Section 5.4. Summary of the chapter is presented in Section 5.5.

5.1 Unsupervised dimensionality reduction

The work in the previous chapter presents a quantitative analysis of the benefit of explicit memory inclusion in a fixed ABE solution. That work builds upon previous investigations of front-end feature extraction for ABE [172] and of memory inclusion via delta features under the constraint of fixed dimensionality [15]. Principal component analysis (PCA) is employed in order to incorporate memory without increasing feature dimensionality; regression complexity is unaffected. While retaining the amount of variation in the input NB features as much as possible, PCA being an unsupervised approach to dimensionality reduction does not take into account the output HB features. The hypothesis of the research presented in this chapter is that supervised or semi-supervised and non-linear dimensionality reduction techniques may offer potential to learn lower dimensional representations tailored specifically to ABE, thereby giving better performance.

The two unsupervised techniques for dimensionality reduction, namely PCA and stacked auto-encoders (SAEs) are explained in brief in the following.

5.1.1 Principal component analysis

PCA, also known as the *Karhunen–Loève transform* (KLT), is a linear transformation that is defined as an orthogonal projection of higher dimensional data into a lower dimensional space. This is achieved by maximising the variance of the input data in the projected space, also known as the principal subspace. The variables of the input data set are transformed to a new set of uncorrelated variables, also referred to as principal components, which are ordered so that first few preserve most of the variation in the original input data [233, Chapter 1]. PCA is widely used for different applications such as dimensionality reduction or lossy compression for higher dimensional data, feature extraction and data visualisation [214, Section 12.1]. As an unsupervised, linear approach to dimensionality reduction, PCA aims only to produce a low dimensional NB representation which retains as much as possible the variation in the original representation. For a specific task such ABE considered in this thesis – where 10 dimensional HB features are estimated using a standard regression model from the 10-dimensional compact NB features extracted from higher dimensional NB data with memory – the new features may not be optimised. This is because the compact NB features may not contain information related to HB features.

5.1.2 Stacked auto-encoders

With the increased computational power of graphics processing units (GPUs) and availability of training data, the numerous research topics are influenced by the success of deep learning techniques. Deep neural networks (DNNs) are well-known for their ability to model highly complex, non-linear functions from training data. The most commonly used non-linear technique to dimensionality reduction in the deep learning literature is that of stacked auto-encoders (SAEs)[1]. SAEs have been widely studied and investigated in the last decade. In contrast to linear PCA, SAEs offer a non-linear solution to dimensionality reduction or feature extraction. They have been applied to many speech processing tasks, e.g., phoneme/speech recognition [148, 234, 235], speech synthesis [236], spoofing detection for automatic speaker recognition [237], speech compression [238] and voice conversion [239, 240]. Common to most of these examples is the use of SAEs to learn so-called bottleneck features, namely compact feature representations tailored to pattern recognition and classification.

Auto-encoders

An auto-encoder (AE) is an artificial neural network that is widely used for the learning of higher-level data representations. An AE consists of an encoder and a decoder as shown in Fig. 5.1(a). The encoder $f_\theta(\cdot)$ maps an input vector $\mathbf{x} \in \mathcal{R}^d$ to a hidden representation $\mathbf{h} \in \mathcal{R}^{d_h}$ according to:

$$\mathbf{h} = f_\theta(\mathbf{x}) = s(\mathbf{W}\mathbf{x} + \mathbf{b}) \tag{5.1}$$

where $\theta = \{\mathbf{W}, \mathbf{b}\}$ is the parameter set of weight matrix $\mathbf{W}$ (of dimension $d \times d_h$) and bias vector $\mathbf{b}$ (of dimension $d_h \times 1$). The function $s(\cdot)$ is a non-linear transformation (or an activation function). The encoder is followed by a decoder $g_{\theta'}(\cdot)$ which aims to reconstruct the original input $\mathbf{x}$ from the learned representation $\mathbf{y}$ according to:

$$\hat{\mathbf{x}} = g_{\theta'}(\mathbf{h}) = s'(\mathbf{W}'\mathbf{h} + \mathbf{b}') \tag{5.2}$$

where $\theta' = \{\mathbf{W}', \mathbf{b}'\}$ and $s'(\cdot)$ is either a linear or a non-linear transformation depending on the nature of input $\mathbf{x}$. For real-valued inputs, parameters $\{\theta, \theta'\}$ are

[1]SAEs are alternatively referred to as deep auto-encoders (DAEs). The literature also refers to stacked auto-associators (SAAs).

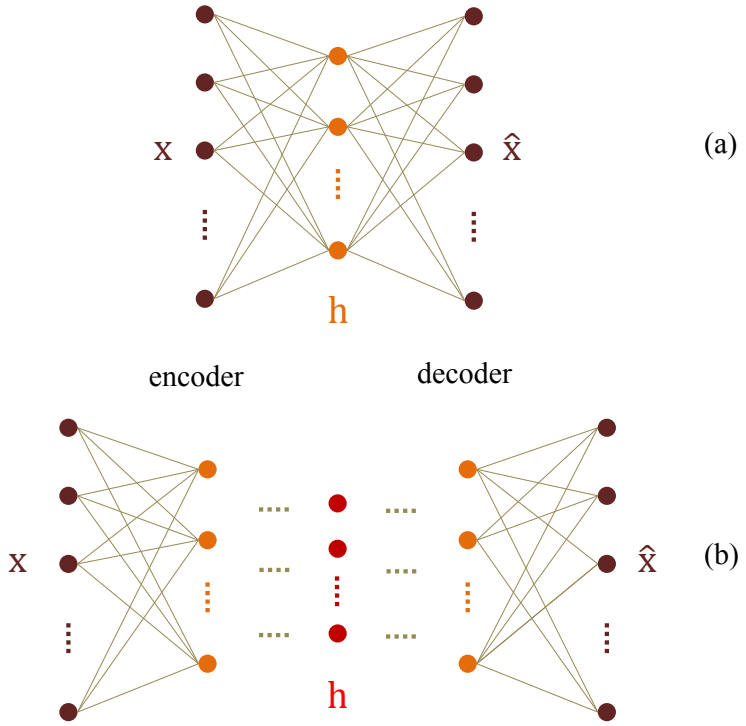


Figure 5.1: The architecture of (a) an auto-encoder (AE) and (b) stacked (deep) auto-encoder (SAE).

optimised according to a mean squared error (MSE) objective loss function which reflects the difference between the input and the reconstructed output.

When the dimension of the hidden representation $\mathbf{h}$ is lower than that of the input $\mathbf{x}$, i.e., $d_h < d$, then the AE is referred to as undercomplete. This form of structure is used for dimensionality reduction or feature extraction [241, Section 14.1]. With linear activations and a squared error loss function, an AE learns to span the same subspace as PCA [242]. A non-linear AE may also perform PCA like transformations if the inputs to the non-linear hidden activations stay in the linear region of the sigmoid function [243]. However, when the layer inputs stay away from the linear range of the sigmoid activation function, then AEs can learn a nonlinear generalisation of PCA [244].

Stacked auto-encoders

The modelling of complex and high-dimensional data distributions using shallow neural networks would require a very large number of parameters in comparison

to deeper networks. DNNs are formed by composing multiple levels of non-linear operations via the use of more than one hidden layer with non-linear activation functions [245, 246]. In comparison to the shallow architectures, they are hence inherently capable of learning highly non-linear and complex functions, efficiently. The depth of an AE can be increased by stacking multiple layers of encoders and decoders, thereby forming a stacked auto-encoder (SAE) as illustrated in Fig. 5.1(b).

Evolution of training strategies in deep learning

DNNs have a tendency to get stuck in local minima as their size increases. Different techniques to improve the optimisation of DNNs have been proposed. These methods seem to improve gradient-descent optimisation, thereby encouraging DNNs towards global minima and better generalisation. A selection of such techniques that are most relevant to the work presented in this chapter, is discussed in below.

Weight initialisation: The training of SAEs (or DNNs) becomes increasingly difficult because of the difficulty in finding global minima during optimisation. With large initial weights, SAEs usually find local minima whereas with small initial weights the gradients in the initial layers become very small thus making training of SAEs infeasible. Therefore, initialisation of the network weights becomes an important factor for efficient optimisation [247].

Some form of pre-training is usually employed to initialise network weights. Popular solutions include weight initialisation via the layer-by-layer learning of deeper networks, also known as unsupervised pre-training, using restricted Boltzmann machines (RBMs) [247] and denoising AEs [248]. Individual layers are stacked after pre-training and subsequently fine-tuned according to a particular task of interest. Such greedy layer-wise pretraining helps optimisation through the better initialisation of hidden layers; hidden layers then represent more meaningful representations of the input. This helps in improved generalisation [249]. Pretraining has been shown to act as a regulariser [250, 251].

Further studies [252] have shown that the use of logistic sigmoid and tanh activation functions drive top/last hidden layers of a randomly initialised deep network into a saturation. This prevents the backward flow of the gradients and therefore, the lower/initial layers are prevented from learning useful representations.

An alternative empirical approach to weight initialisation, also known as normalised initialisation[2] is proposed in [252]. It is designed to maintain the same variance for all activation outputs and back-propagated gradients across all layers upon initialisation. While this method theoretically assumes that the network operates in the linear regime of its activation function, the technique is reported to perform reasonably well with hyperbolic tangent (tanh) and softsign [253] non-linearities.

The study reported in [254] derives the improved weight initialisation technique for extremely deep networks (with 30 hidden layers) with rectified linear units (ReLUs). The derivation takes into account the non-linear behaviour of the rectified activation units.

Rectifier linear units (ReLUs): Rectified (or hinge) activation units, also known as rectifiers, are a better model of biological neurons and give equivalent or better performance than the logistic sigmoid or tanh activations, even without the need of unsupervised pretraining [255]. However, a ReLU is activated only when its input is above 0. This can cause a problem in some cases during gradient-based optimisation when a small number of inactive units may never be learned reliably due to zero gradients. The use of ReLUs is thus adapted slightly in order to allow small, non-zero gradients during back-propagation when the unit is saturated or inactive. The ReLUs or their variants (e.g., leaky ReLUs (LReLUs) [256], parametric ReLUs (PReLUs) [254], exponential linear units(ELUs) [257]) have all shown improved results on tasks such as image recognition and text classification [255], image classification [258] and speech recognition [256, 259].

Dropout: Dropout [260] is a technique to improve the performance of DNNs by reducing overfitting. It involves randomised dropping out of a unit (hidden or visible) in each layer (except the output layer) of a DNN with some fixed probability p (that is independent of the other units) for each training sample in a mini-batch optimisation procedure. The hidden unit cannot then rely on the presence of the other units in the network during training. This reduces the so-called co-adaptation among hidden units so that they are forced to learn more robust features. From another point of view, dropout is an approximate and efficient model averaging technique where the huge number of different architectures of a DNN are trained in a reasonable time and all the networks share the same weights for the hidden units that are present. During testing, outgoing weights of each unit are multiplied

[2]In deep learning literature, alternatively it is also referred as Xavier or Glorot initialisation.

by the scaling factor $(1 - p)$ where p represents the probability with which the unit was dropped during training. Dropout is also considered as a regularisation technique that is similar to the addition of noise to hidden units as in the case of denoising AEs, the noise in this case being generated by the dropping of the units with a certain probability [261].

Batch-normalisation: The distribution of inputs to each hidden layer of a DNN keep changing during training because of the change in the parameters of the previous layer leading to a problem of *internal covariance shift*. The effect leads to the saturation of the neurons which, in turn, slows down convergence making the optimisation inefficient. This problem is tackled by normalising the inputs to each layer for each mini-batch where the normalisation parameters for each layer are learned during training. This mechanism is referred to as batch-normalisation (BN) [262]. It helps to achieve a stable distribution of the activation values throughout training thereby reducing the chances of the network getting stuck in the saturation regime of non-linearity. BN thus also makes the optimisation process less sensitive to the higher learning rates and initial weight initialisation. It also acts as a regulariser thereby making the use of dropout unnecessary [262].

With incorporation of some methods[3] (e.g.: use of rectifiers and weight initialisation techniques to avoid saturation regime and local minima during optimisation; techniques such as dropout and batch-normalisation that help to reduce overfitting) during DNN training, the pretraining is often no longer needed as it does not bring any significant improvements [241, Section 15.1].

5.2 ABE using semi-supervised stacked auto-encoders

This section reports the use of SAEs for non-linear dimensionality reduction in ABE, specifically the use of SAEs trained in a semi-supervised manner. The objectives are to (i) harness memory in a compact, low dimensional representation in order to improve the reliability of estimated HB components and (ii) to learn NB features directly from spectral coefficients instead of hand-crafted features. The merit of both contributions is assessed through objective assessment, an information

[3]Numerous techniques for weight initialisation and regularisation for better optimisation of DNNs are reported in the literature. Only few techniques are discussed here in brief and explored in our experiments. This is because it is well known that no particular technique gives consistent results and the choice is mostly dependent on the task of interest. The tuning of hyperparameters thus becomes important.

theoretic approach and informal listening tests.

The ABE algorithm presented in the previous chapter uses unsupervised, linear dimensionality reduction, i.e., PCA, so that the complexity of the standard regression model learned in training and used in estimation, remains unchanged as a result of memory inclusion. It is used as a baseline algorithm. We seek to further improve ABE performance using a SAE trained in a semi-supervised manner.

5.2.1 Semi-supervised stacked auto-encoders

With a reconstruction-based objective loss function, SAEs are trained to maximise the lower bound on the mutual information between the input $\mathbf{x}$ and the learned representation $\mathbf{h}$; the significant information about the input $\mathbf{x}$ is retained in $\mathbf{h}$. However, such training criterion may not necessarily yield the most *useful* representation because SAEs can simply learn a trivial identity mapping between the input and the reconstructed output, rather than a meaningful, high-level representation [248]. Additionally, being unsupervised, features extracted from the bottleneck layer of a conventional SAE are not expressly designed for classification or regression; they will likely be suboptimal in this respect.

Supervised (or discriminative) fine-tuning of DNNs after unsupervised pretraining for robust feature extraction in recognition or classification tasks is important. The partially-supervised pre-training of AEs was shown in [249] to be beneficial, especially for regression tasks. Specifically, a mixed training criterion was used to pretrain each layer that is a combination of both supervised and unsupervised objectives. The unsupervised objective thus helps to model or reconstruct the input whereas the supervised objective helps to predict the target.

Drawing upon this work, we have explored the semi-supervised training of SAEs in order to learn compact representations designed specifically for regression modelling and ABE. The resulting semi-supervised SAE (SSAE) architecture with 2 output layers is illustrated in Fig. 5.2. While one output layer is learned to reconstruct the input NB features (AE output) as with a conventional SAE, the other output layer is learned to estimate the missing HB features (regression output). This is achieved though a joint objective loss function given by:

$$L_{total} = c * L_{reg} + (1 - c) * L_{ae}$$

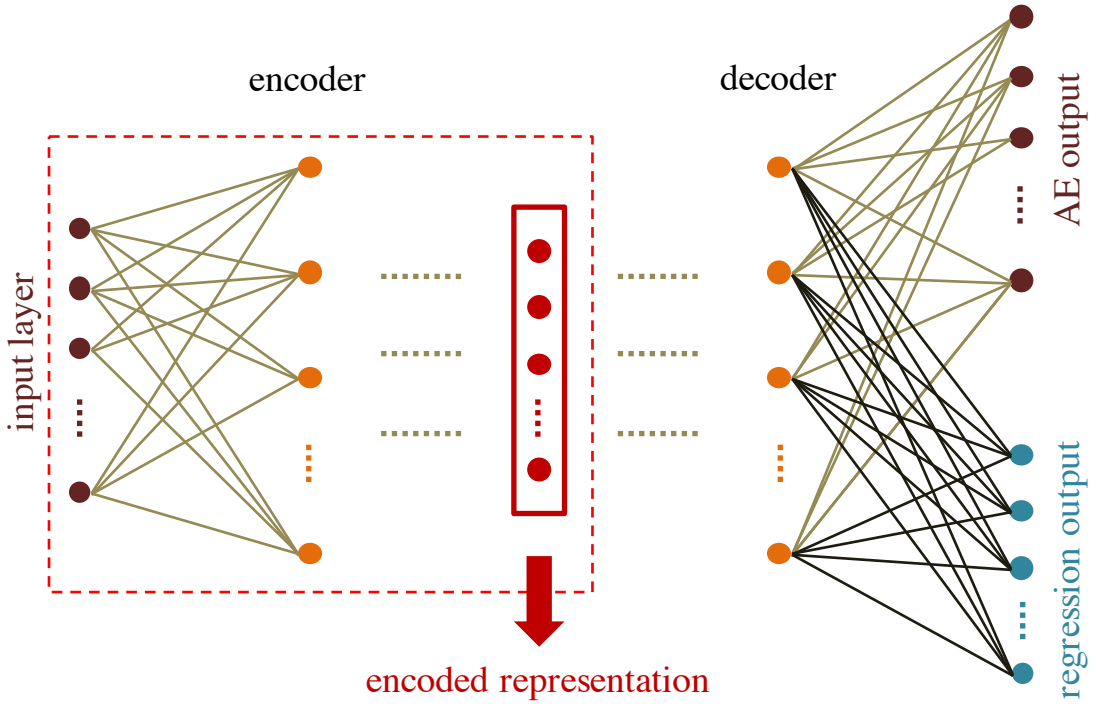


Figure 5.2: A semi-supervised stacked auto-encoder (SSAE).

where L_{reg} and L_{ae} are the objective loss functions for regression and AE outputs respectively and where $c \in [0, 1]$ weights the contribution of individual losses. As the input NB and output HB features are real-valued, optimisation is performed with a MSE objective loss function.

5.2.2 Application to ABE

After training, the AE output layer of the SSAE can be discarded and the resulting feedforward architecture can also be used to estimate the HB components directly from the regression layer. A similar CNN based architecture designed to regularise the mapping of short i-vectors to long i-vectors for a speaker verification task is reported in [263]. The focus here is different, i.e., to regularise/supervise dimensionality reduction so that it preserves information critical to ABE, information that is further exploited by an otherwise standard regression model. Therefore, after SSAE training, the decoder is discarded and only the *encoder* (red box in Fig. 5.2) is retained.

In order to investigate the merit of the SSAE-based approach to dimensionality

reduction and ABE, the weight matrix $\mathbf{W}_{\text{PCA}}$ in Fig. 4.3 (red boxes) is replaced by the SSAE *encoder*. Extracted low dimensional features are then mean and variance normalised[4]. GMM training and estimation are performed in the same manner described in Section 4.4.

Also reported here in this chapter is a variation on this approach whereby the low dimensional NB representation is derived directly from NB log power spectrum (LPS) coefficients instead of logMFE features. This is achieved quite simply by replacing logMFE features with LPS coefficients given by P_t^{NB} (refer to Section 4.3.1 for details related to the calculation of P_t^{NB}).

5.3 Experimental setup

The ABE algorithm with explicit memory inclusion and dimensionality reduction using PCA presented in Chapter 4 is used as a baseline. Experiments are designed to compare the performance of the baseline ABE system using PCA dimensionality reduction $\text{M}_{\text{PCA\_2}}$ to that of the same system using SSAE dimensionality reduction $\text{M}_{\text{SSAE\_2}}$. Systems $\text{M}_{\text{PCA\_2}}$ and $\text{M}_{\text{SSAE\_2}}$ use 10-dimensional features $\mathbf{x}^{\text{NB}}_{t,pca\_2}$ and $\mathbf{x}^{\text{NB}}_{t,ssae\_2,mvn}$ respectively, obtained using static features extracted from 2 neighbouring speech frames. This section describes the details of the configuration, training and optimisation of the SSAE architecture.

5.3.1 SSAE training, configuration and optimisation

Training

The SSAE was implemented with the Keras toolkit [264]. Consistent with the work presented in the previous chapter, features $\mathbf{x}^{\text{NB}}_{t,conc\_2}$ at time t (obtained from the concatenation of features extracted from 2 preceding and 2 proceeding frames) are fed to the input of the SSAE. Whereas the AE output is the same as the input, the regression output is set to HB features $\mathbf{y}^{\text{HB}}_{t,mvn}$. So as to improve the rate of convergence to global minima, the SSAE is initialised according to the approach described in [254]. With a MSE criterion, optimisation is performed according to the procedure described in [265], referred to as Adam optimisation, with an initial

[4]The mean-variance normalisation (MVN) is found to be critical to ABE performance using SSAE features. Further analysis and results with and without MVN are discussed in detail in the next chapter.

learning rate of 10^{-3} and hyperparameters $\beta_1 = 0.9, \beta_2 = 0.999$ and $\epsilon = 10^{-8}$.

Configuration

We investigated two 6-layer conventional symmetric SSAE structures with different numbers of units in the hidden layers:

1. 512, 256, 10, 256, 512 (denoted as Arch-1);
2. 1024, 512, 10, 512, 1024 (denoted as Arch-2).

The number of units in the middle hidden layer is chosen to be 10 in order to satisfy the constraint of fixed dimensionality imposed by the GMM regression model used for subsequent ABE. The analysis of previously reported DNN-based solutions to ABE have shown that increasing the number of hidden layers or the number of units does not significantly improve estimation performance [86, 124, 147]. Therefore the search space for optimisation is reduced by selecting only two topologies mentioned above with the assumption that they provide the best trade-offs between performance and complexity.

Output layers consists of 50 (AE) and 10 (regression) units. Hidden layers have tanh or ReLU activation units whereas output layers have linear activation units. In order to discourage over-fitting, the use of dropout (dr) [261] and batch-normalisation (bn) [262] techniques are also investigated. The learning rate is reduced by half in the case that the validation loss increases between 2 consecutive epochs. Regression and AE loss weights were both set to c=0.5. Networks are trained for 30 epochs and the model with the lowest validation loss is used for subsequent analysis.

Optimisation

Performance in terms of MSE, on training (T) and validation (V) data, for the two different architectures (Arch-1 and Arch-2) and four different combinations of dropout (dr) and batch-normalisation performed either after (bn-a) or before (bn-b) activation is shown in Table 5.1. Dropout is applied before all hidden layers.

Relatively low values of MSE are achieved without dropout or batch normalisation (configuration A), although performance is poor for Arch-2 with ReLU

Table 5.1: MSE for different SSAE configurations with either ReLU or tanh activation functions, with and without dropout (dr) and batch normalisation (bn) either after (a) or before (b) activation. dr value represents fraction (p) of randomly chosen hidden units being set to 0. Results are illustrated for the each SSAE configuration on training (T) and validation (V) datasets.

		dr	bn	Arch-1		Arch-2	
				ReLU	tanh	ReLU	tanh
T	A	-	-	0.454	0.439	1.000	0.445
	B	-	a	**0.434**	**0.436**	**0.428**	0.444
	C	-	b	0.437	0.438	0.436	**0.438**
	D	0.2	-	1.000	0.486	1.000	0.492
V	A	-	-	0.474	0.461	1.012	0.467
	B	-	a	0.460	0.461	0.461	0.467
	C	-	b	**0.459**	**0.459**	**0.460**	**0.460**
	D	0.2	-	1.012	0.504	1.012	0.509

activation. The use of dropout without batch-normalisation (configuration D) results in poorly regularized networks, especially for ReLU activation. Similar observations are reported in [124], namely that the use of dropout with filterbank based features does not improve performance on training and validation data and does not help to reduce overfitting. The use of either form of batch-normalisation without dropout gives consistently low values of MSE, with the best results being obtained with a bn-b configuration (C) on validation data. This configuration is thus used for further experiments related to ABE.

5.3.2 Databases and metrics

ABE performance is assessed with three objective metrics and informal listening tests. The correlation between NB features (obtained using SSAE and PCA) and HB features is measured via mutual information (MI).

ABE experiments were performed on TIMIT and TSP speech datasets described in Sections 3.2.1 and 3.2.2 respectively. While the TIMIT dataset is used for training

and optimisation of SSAE parameters, the TSP speech dataset is used for testing (Section 3.3.2).

5.4 Results

This section presents improvements in ABE performance using SSAE-based dimensionality reduction approach in comparison to PCA. Results are reported for both the configurations, Arch-1C and Arch-2C. The assessment of speech quality is performed via three objective speech quality measures and informal listening tests. MI results are also reported.

5.4.1 Speech quality assessment

The objective metrics include the two distance metrics $d_{\text{RMS-LSD}}$, d_{COSH} and objective estimates of MOS scores MOS-LQO$_{\text{WB}}$ (Section 3.4.2). Objective performance obtained for the testing set and for both the baseline $\text{M}_{\text{PCA\_2}}$ and SSAE-based approach $\text{M}_{\text{SSAE\_2}}$ to ABE are illustrated in Table 5.2. With only one exception (i.e., the case of $d_{\text{RMS-LSD}}$ for Arch-2C with ReLUs), spectral distortion metric results show lower values for SSAE than for the baseline. MOS-LQO$_{\text{WB}}$ scores for SSAE systems are consistently higher. The Arch-2C SSAE system with a tanh activation performs best; the distance metrics $d_{\text{RMS-LSD}}$, d_{COSH} are improved by 3.1% (7.34 $\rightarrow$ 7.11 dB) and 5.9% (1.52 $\rightarrow$ 1.43 dB) relative to the baseline system. The MOS-LQO$_{\text{WB}}$ results are improved by 0.11 points (2.96 $\rightarrow$ 3.11) leading to 3.7% of relative improvement. Unfortunately, though, despite convincing improvements in objective performance metrics, informal listening tests showed little discernible differences between the quality of speech signals produced by the baseline and SSAE systems.

Objective performance measures for the two best performing SSAE configurations, Arch-1C and Arch-2C both with tanh activations, trained using LPS inputs instead of logMFE features are illustrated in Table 5.3. Distortion measures are consistently lower, whereas MOS-LQO$_{\text{WB}}$ scores are consistently higher than results for all other SSAE-based systems. Specifically, the best SSAE system (Arch-2C with tanh activation) improves ABE performance in terms of $d_{\text{RMS-LSD}}$, d_{COSH} and MOS-LQO$_{\text{WB}}$ measures by 6.3% (7.34 $\rightarrow$ 6.88 dB), 13.4% (1.52 $\rightarrow$ 1.34 dB) and 7.1% (2.96 $\rightarrow$ 3.17) relative in comparison to the baseline system. In contrast

Table 5.2: Objective performance metric results (with mean and standard deviation values) for ABE system $\mathrm{M}_{\mathrm{SSAE\_2}}$. $d_{\text{RMS-LSD}}$ and d_{COSH} are spectral distortion measures in dB (lower values indicate better performance) whereas $\text{MOS-LQO}_{\text{WB}}$ values reflect quality (higher values indicate better performance).

Objective metrics	Arch-1C		Arch-2C		Baseline
	ReLU	tanh	ReLU	tanh	
$d_{\text{RMS-LSD}}$	7.28 (0.70)	7.12 (0.68)	7.38 (0.69)	**7.11** (0.67)	7.34 (0.70)
d_{COSH}	1.48 (0.35)	1.44 (0.36)	1.49 (0.35)	**1.43** (0.34)	1.52 (0.38)
$\text{MOS-LQO}_{\text{WB}}$	2.99 (0.34)	3.06 (0.34)	2.99 (0.34)	**3.07** (0.34)	2.96 (0.34)

Table 5.3: Objective assessment results for ABE system $\mathrm{M}_{\mathrm{SSAE\_2}}$ using log power spectrum (LPS) inputs in place of log-Mel filter energy (logMFE).

SSAE configuration	$d_{\text{RMS-LSD}}$	d_{COSH}	$\text{MOS-LQO}_{\text{WB}}$
Arch-1C, tanh	6.90 (0.63)	1.37 (0.34)	3.16 (0.33)
Arch-2C, tanh	**6.88** (0.62)	**1.34** (0.33)	**3.17** (0.33)

Table 5.4: Mutual information assessment results. $I(\mathbf{x};\mathbf{y})$ denotes the MI between features $\mathbf{x}$ and $\mathbf{y}$.

$I(\mathbf{x}^{\text{NB}}_{t,pca\_2};\mathbf{y}_t)$, Baseline	1.55
$I(\mathbf{x}^{\text{NB}}_{t,ssae\_2};\mathbf{y}_t)$, Arch-1C (with logMFE input)	1.69
$I(\mathbf{x}^{\text{NB}}_{t,ssae\_2};\mathbf{y}_t)$, Arch-2C (with logMFE input)	1.71
$I(\mathbf{x}^{\text{NB}}_{t,ssae\_2};\mathbf{y}_t)$, Arch-1C (with LPS input)	1.84
$I(\mathbf{x}^{\text{NB}}_{t,ssae\_2};\mathbf{y}_t)$, Arch-2C (with LPS input)	**1.90**

to findings for the SSAE systems that operate using logMFE features, informal listening tests show discernible improvements to speech quality compared to speech produced using the baseline ABE system.

5.4.2 Mutual information assessment

A final set of results aims to further validate the findings of both objective metrics and informal listening tests. This is achieved by observing improvements to the mutual information (MI) between the learned NB representation and true HB representation measured using the testing set. A 128-component full-covariance GMM trained with joint vectors (obtained from the testing set) formed by learned NB and true HB features is used for the MI estimation according to Eq. 3.19 (Section 3.4.3).

MI results presented in Table 5.4 show that the Arch-2C SSAE system with tanh activations trained using LPS inputs gives a relative increase in MI of $\approx 23\%$ ($1.55 \rightarrow 1.90$) over the baseline system. This result corroborates the findings presented above, namely that semi-supervised techniques which operate on log spectral inputs are capable of learning better representations that can be exploited to deliver improved ABE performance.

5.5 Summary

This chapter presents a non-linear, semi-supervised approach to dimensionality reduction for artificial bandwidth extension. The work aims to further improve ABE system performance with explicit memory inclusion that uses PCA – a linear unsupervised approach to dimensionality reduction. The key contribution to this work are as follows:

- The ability of stacked auto-encoders trained in semi-supervised fashion to learn higher-level representations is explored to learn compact narrowband features directly from log spectra. The merit of the approach is demonstrated with different objective metrics and is confirmed by the findings of informal listening tests. The benefit of the approach is confirmed by information theoretic analysis.

- The narrowband feature representation is learned automatically from log spectral coefficients in a data-driven manner. The learned low-dimensional features are further used by a standard regression model without augmenting complexity. Therefore, artificial bandwidth extension is presented as a feature learning problem where the compact NB representation can be learned from higher dimensional spectral data resulting from memory inclusion. The aim is

to learn front-end NB features that can improve HB estimation performance under the constraint of fixed dimensionality.

The work presented in this chapter points towards a number of directions for future work that should investigate potential spectral modelling transforms and their further optimisation to learn features for ABE. The investigation of the combination of semi-supervised auto-encoders with unsupervised or partially supervised pre-training methods is also of interest. These directions may offer even greater potential to improve the quality of artificially bandwidth-extended speech. Generative models such as variational auto-encoders can also be explored for improved feature learning. This idea is the subject of research presented in the next chapter.

Chapter 6

Latent representation learning for ABE

Artificial bandwidth extension (ABE) algorithms improve speech quality when wideband devices are used with narrowband devices or infrastructure. Most ABE solutions employ some form of memory, implying high-dimensional feature representations that increase both latency and complexity. The work presented in previous chapters showed that dimensionality reduction techniques can be employed to preserve efficiency. These entail the extraction of compact, low-dimensional narrowband representations that are then used with a standard regression model to estimate high-band components. The previous chapter showed that some form of supervision is crucial to the optimisation of dimensionality reduction techniques for ABE. In extending this work, this chapter reports the first application of conditional variational auto-encoders (CVAEs) for supervised dimensionality reduction specifically tailored to ABE. CVAEs, a form of directed, graphical models, are exploited to model higher-dimensional log-spectral data to extract the latent narrowband representations. When compared to results obtained with alternative dimensionality reduction techniques, objective and subjective assessments show that the probabilistic latent representations learned with CVAEs produce bandwidth-extended speech signals of notably better quality.

The remainder of this chapter is organised as follows. Details of VAE and CVAE architectures are presented in Sections 6.1 and 6.2 respectively. Section 6.3 explains the proposed feature extraction scheme using VAEs and CVAEs. Experimental results are presented in Section 6.4. Conclusions are summarised in Section 6.5.

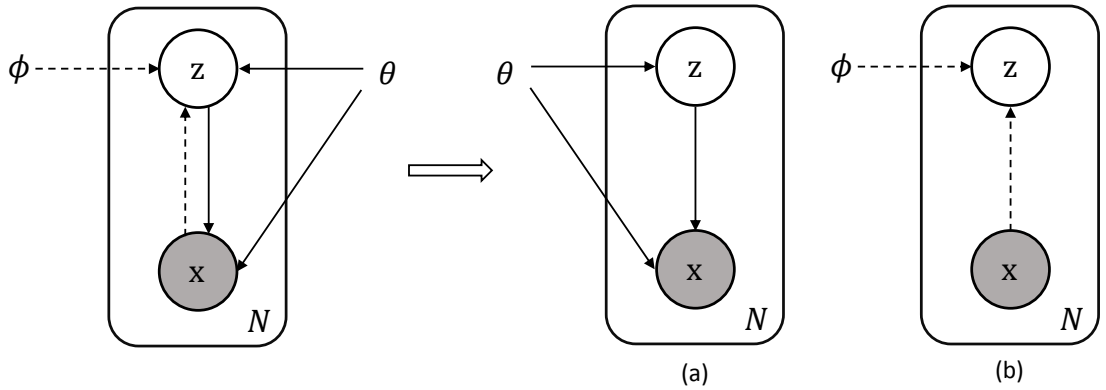


Figure 6.1: A variational auto-encoder (VAE) as a directed graphical model (adapted from [266]). Solid lines represent the generative model $p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{z})p_\theta(\mathbf{x}|\mathbf{z})$ with parameters θ (shown in (a)). Dashed lines represent the inference of the true posterior $p_\theta(\mathbf{z}|\mathbf{x})$ performed via the variational approximation $q_\phi(\mathbf{z}|\mathbf{x})$ with parameters ϕ (shown in (b)). Dashed and solid lines alternately represent *encoding* and *decoding* phases respectively. The shaded node represents the observed variable $\mathbf{x}$. The generative parameters θ and the variational parameters ϕ are jointly learned during optimisation.

6.1 Variational auto-encoders

A variational auto-encoder (VAE) [266] is a generative model $p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{z})p_\theta(\mathbf{x}|\mathbf{z})$ (with parameters θ) which assumes that a dataset $\mathcal{D}$, consisting of N i.i.d. samples or observations of a random variable $\mathbf{x}$, is generated from a underlying continuous unobserved (or latent) variable $\mathbf{z}$. The generative process thus consists of two steps. First, a value $\mathbf{z}^{(i)}$ is generated from some prior distribution $p_\theta(\mathbf{z})$. Second, a value $\mathbf{x}^{(i)}$ is generated from the conditional distribution $p_\theta(\mathbf{x}|\mathbf{z})$. Thus the aim is to optimise the parameters θ such that the generated $\mathbf{x}^{(i)}$ is similar to the samples in the training dataset $\mathcal{D}$ with a high probability. This is achieved by maximising the marginal likelihood $p_\theta(\mathbf{x})$ of each datapoint $\mathbf{x}$ in the training data, given by:

$$\begin{aligned} p_\theta(\mathbf{x}) &= \int p_\theta(\mathbf{x}, \mathbf{z}) d\mathbf{z} \\ &= \int p_\theta(\mathbf{z}) p_\theta(\mathbf{x}|\mathbf{z}) d\mathbf{z} \end{aligned} \tag{6.1}$$

In practice, when the likelihood functions $p_\theta(\mathbf{x}|\mathbf{z})$ are complex (e.g. a neural network with non-linear hidden layers), the integral in Eq. 6.1 is intractable and therefore the marginal likelihood $p_\theta(\mathbf{x})$ cannot be optimised directly w.r.t θ. The

true posterior density $p_\theta(\mathbf{z}|\mathbf{x}) = \frac{p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{z})}{p_\theta(\mathbf{x})}$ is also thus intractable, prohibiting the use of the EM algorithm for optimisation of θ [214, Chapter 10]. To alleviate this problem, VAEs introduce a recognition/inference model $q_\phi(\mathbf{z}|\mathbf{x})$ as an approximation to the posterior $p_\theta(\mathbf{z}|\mathbf{x})$ which allows to obtain a tractable lower bound on the data likelihood, using the so-called *reparameterisation trick*. The data likelihood then can be maximised by optimisation of the lower bound. A representation of VAE in the form of a directed graphical model[1] is shown in Fig. 6.1.

6.1.1 Variational lower bound

For a given inference model $q_\phi(\mathbf{z}|\mathbf{x})$, the marginal likelihood of a single datapoint $\mathbf{x}$ is given by:

$$\log p_\theta(\mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x})] \tag{6.2}$$

$$= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log\left[\frac{p_\theta(\mathbf{x},\mathbf{z})}{p_\theta(\mathbf{z}|\mathbf{x})}\right]\right] \tag{6.3}$$

$$= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log\left[\frac{q_\phi(\mathbf{z}|\mathbf{x})}{q_\phi(\mathbf{z}|\mathbf{x})}\frac{p_\theta(\mathbf{x},\mathbf{z})}{p_\theta(\mathbf{z}|\mathbf{x})}\right]\right] \tag{6.4}$$

$$= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log\left[\frac{q_\phi(\mathbf{z}|\mathbf{x})}{p_\theta(\mathbf{z}|\mathbf{x})}\right]\right] + \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log\left[\frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right]\right] \tag{6.5}$$

$$= D_{KL}[q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x})] + \mathcal{L}(\theta,\phi;\mathbf{x}) \tag{6.6}$$

The first term in Eq. 6.6 represents the Kullback-Leibler (KL) divergence (D_{KL}) between the approximate and true posterior distributions. For simplicity, it is assumed that the approximate and true posteriors are diagonal multivariate Gaussian distributions whose respective parameters θ and ϕ are computed using two different deep neural networks.

Since the KL divergence term is non-negative, the second term $\mathcal{L}(\theta,\phi;\mathbf{x})$ represents a *variational lower bound*, also called the *evidence lower bound* (ELBO), on

[1] Visualisation and analysis of complex joint probability distributions can be provided via their diagrammatic representations referred to as *probabilistic graphical models* (PGMs). Conditional independence properties of joint distributions thus can be easily inspected from their graphical representations. A PGM consists of *nodes*, each of them represents a random variable. The nodes are connected by *links* which represent the probabilistic relationship among the connected nodes. The resulting graph then captures the joint distribution over all of the random variables and its decomposition into the product of various conditional probability distributions [214, Chapter 8].

The type of PGMs in which the links are represented by arrows which have directional significance are called *directed graphical models* (also called *Bayesian networks*).

the marginal likelihood. It is given by:

$$\mathcal{L}(\theta, \phi; \mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}, \mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x})] \tag{6.7}$$
$$\leq \log p_\theta(\mathbf{x}) \tag{6.8}$$

Since the ELBO has an expectation w.r.t. $q_\phi(\mathbf{z}|\mathbf{x})$ (which is a function of ϕ), Eq. 6.7 is differentiable w.r.t θ, but not ϕ, i.e.:

$$\begin{aligned}\nabla_\theta \mathcal{L}(\theta, \phi; \mathbf{x}) &= \nabla_\theta \, \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}, \mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x})] \\ &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\nabla_\theta(\log p_\theta(\mathbf{x}, \mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x}))] \\ &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\nabla_\theta \log p_\theta(\mathbf{x}, \mathbf{z})] \end{aligned} \tag{6.9}$$

$$\begin{aligned}\nabla_\phi \mathcal{L}(\theta, \phi; \mathbf{x}) &= \nabla_\phi \, \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}, \mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x})] \\ &\neq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\nabla_\phi(\log p_\theta(\mathbf{x}, \mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x})] \end{aligned} \tag{6.10}$$

This prohibits the joint optimisation of the ELBO w.r.t. both θ and ϕ. In order to alleviate this problem, VAEs exploit the reparameterisation trick.

6.1.2 Reparameterisation trick

The ELBO can be straightforwardly optimised w.r.t. both θ and ϕ by a change of random variables from $\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})$ to $\epsilon \sim p(\epsilon)$ using a some deterministic differentiable transformation such that:

$$\mathbf{z} = g_\phi(\mathbf{x}, \epsilon) \tag{6.11}$$

This transformation is known as the *reparameterisation trick*. Accordingly, the expectation of the ELBO (Eq. 6.7) w.r.t. $q_\phi(\mathbf{z}|\mathbf{x})$ can now be replaced with one w.r.t. $p(\epsilon)$ such that:

$$\mathcal{L}(\theta, \phi; \mathbf{x}) = \mathbb{E}_{p(\epsilon)}[\log p_\theta(\mathbf{x}, g_\phi(\mathbf{x}, \epsilon)) - \log q_\phi(g_\phi(\mathbf{x}, \epsilon)|\mathbf{x})] \tag{6.12}$$

where $\epsilon \sim p(\epsilon)$. The ELBO $\mathcal{L}(\theta, \phi; \mathbf{x})$ thus takes a form that can be differentiated w.r.t. the parameters ϕ to obtain the gradients $\nabla_\phi \mathcal{L}(\theta, \phi; \mathbf{x})$.

Now, using Monte Carlo estimates for expectation over L samples, the generic stochastic gradient variational Bayes (SGVB) estimator $\tilde{\mathcal{L}}(\theta, \phi; \mathbf{x})$ of the ELBO is

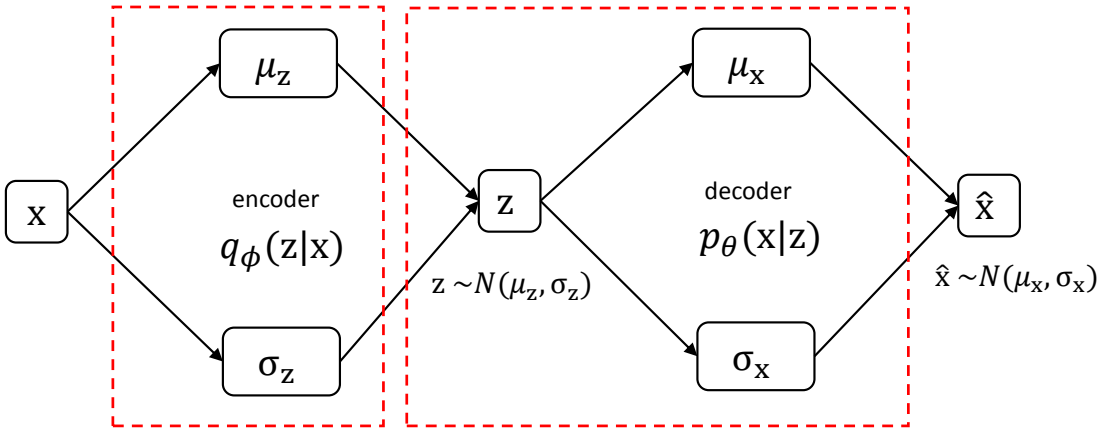


Figure 6.2: An illustration of the variational auto-encoder (VAE) generative model that learns a joint distribution $p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{z})p_\theta(\mathbf{x}|\mathbf{z})$. The latent space (with prior distribution $p(\mathbf{z})$) is inferred using the probabilistic encoder $q_\phi(\mathbf{z}|\mathbf{x})$, that approximates the true but intractable posterior $p_\theta(\mathbf{z}|\mathbf{x})$ of the generative model $p_\theta(\mathbf{x}, \mathbf{z})$. The latent space is mapped back to the input space using the probabilistic decoder $p_\theta(\mathbf{x}|\mathbf{z})$.

given according to:

$$\tilde{\mathcal{L}}(\theta, \phi; \mathbf{x}) = \frac{1}{L}\sum_{l=1}^{L}[\log p_\theta(\mathbf{x}, g_\phi(\mathbf{x}, \epsilon^{(l)})) - \log q_\phi(g_\phi(\mathbf{x}, \epsilon^{(l)})|\mathbf{x})] \tag{6.13}$$

where $\epsilon \sim p(\epsilon)$. Eq. 6.13 can now be jointly optimised w.r.t. the parameters θ and ϕ using stochastic gradient descent.

6.1.3 Relation to conventional auto-encoders

The ELBO given in Eq 6.7 can alternately be re-written as:

$$\mathcal{L}(\theta, \phi; \mathbf{x}) = -D_{KL}[q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})] + \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] \tag{6.14}$$

The first term $D_{KL}(\cdot)$ in Eq. 6.14 acts as a regulariser which represents the KL divergence between the prior $p(\mathbf{z})$ and the approximate posterior $q_\phi(\mathbf{z}|\mathbf{x})$ and can be computed analytically. The second term is the expected negative reconstruction error and must be estimated by sampling. The optimisation of Eq. 6.14 is thus equivalent to that of reconstruction error (as in conventional auto-encoders) with an additional regularisation term.

Thus, as illustrated in Fig. 6.2, VAEs consist of an inference model $q_\phi(\mathbf{z}|\mathbf{x})$ (also referred to as probabilistic encoder) which, for every datapoint $\mathbf{x} \in \mathcal{D}$, estimates the parameters of the posterior distribution over all possible values of the latent variable $\mathbf{z}$ that may have generated the datapoint $\mathbf{x}$. The probabilistic decoder, i.e., $p_\theta(\mathbf{x}|\mathbf{z})$, then produces a distribution over all possible values of $\mathbf{x}$ for a given $\mathbf{z}$.

6.1.4 VAEs for real valued Gaussian data

The VAE framework combines both inference and generative models which are derived using neural networks and provides a simple method for their joint optimisation wherein the lower bound on the marginal log-likelihood of the data is maximised. For real-valued data, the prior $p(\mathbf{z})$, encoder $q_\phi(\mathbf{z}|\mathbf{x})$ and decoder $p_\theta(\mathbf{x}|\mathbf{z})$ of the VAE framework are modelled as follows:

- The prior is assumed to be a centered isotropic multivariate Gaussian with no free parameters, i.e., $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I})$.
- The probabilistic decoder, constructed using a DNN, models the multivariate Gaussian distribution with mean $\boldsymbol{\mu}_\mathbf{x}$ and covariance $\sigma^2\mathbf{I}$ such that $p_\theta(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_\mathbf{x}, \sigma^2\mathbf{I})$. The mean $\boldsymbol{\mu}_\mathbf{x} = \boldsymbol{\mu}(\mathbf{z}; \theta)$ is the output of a DNN $f_\theta(\cdot)$ parameterised by θ whereas σ^2 is a hyperparameter.
- For simplicity, the intractable posterior is assumed to be a multivariate Gaussian distribution with diagonal covariance matrix. Therefore, the variational approximation to the true posterior is given by: $q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_\mathbf{z}, \text{diag}(\boldsymbol{\sigma}^2_\mathbf{z}))$. The mean $\boldsymbol{\mu}_\mathbf{z} = \boldsymbol{\mu}(\mathbf{x}; \phi)$ and variance $\boldsymbol{\sigma}^2_\mathbf{z} = \boldsymbol{\sigma}^2(\mathbf{x}; \phi)$ are obtained using the arbitrary non-linear deterministic transformations ($\boldsymbol{\mu}(\mathbf{x}; \phi)$ and $\boldsymbol{\sigma}^2(\mathbf{x}; \phi)$) with parameters ϕ implemented via a DNN $f_\phi(\cdot)$ that models $q_\phi(\mathbf{z}|\mathbf{x})$.

With the above mentioned approximations, the D_{KL} term in Eq. 6.14 can usually be integrated analytically and approximated by:

$$\begin{aligned} -D_{KL}[q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})] &= -D_{KL}[\mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_\mathbf{z}, \text{diag}(\boldsymbol{\sigma}^2_\mathbf{z}))||\mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I})] \\ &= \frac{1}{2}\sum_{j=1}^{d_\mathbf{z}} \left[1 + \log(\sigma^2_{z_j}) - \mu^2_{z_j} - \sigma^2_{z_j}\right] \end{aligned} \tag{6.15}$$

where $d_\mathbf{z}$ is the dimensionality of the latent variable $\mathbf{z}$ and where μ_{z_j} and σ_{z_j} denote the j^{th} elements of the vectors $\boldsymbol{\mu}_\mathbf{z}$ and $\boldsymbol{\sigma}_\mathbf{z}$ respectively. The second term in Eq. 6.14

is estimated by sampling and is approximated by an expectation over L samples drawn/sampled from the inference network $q_\phi(\mathbf{z}|\mathbf{x})$ according to:

$$\begin{aligned}\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] &= \frac{1}{L}\sum_{l=1}^{L}\log p_\theta(\mathbf{x}|\mathbf{z}^{(l)}) \\ &= \frac{1}{L}\sum_{l=1}^{L}\log \mathcal{N}(\mathbf{x};\boldsymbol{\mu}(\mathbf{z}^{(l)};\theta),\sigma^2\mathbf{I}) \\ &= \frac{1}{L}\sum_{l=1}^{L}\left[C - \frac{\|\mathbf{x}-\boldsymbol{\mu}(\mathbf{z}^{(l)};\theta)\|^2}{\alpha_{\text{vae}}}\right] \end{aligned} \tag{6.16}$$

where $C = \frac{d_\mathbf{x}}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2)$ is a constant that can be ignored during optimisation and $d_\mathbf{x}$ is the dimensionality of $\mathbf{x}$. The sampling operation (that samples $\mathbf{z}$ from the distribution $q_\phi(\mathbf{z}|\mathbf{x})$) is non-differentiable. Accordingly, it is performed via the *reparameterisation trick* such that:

$$\mathbf{z}^{(l)} = g_\phi(\mathbf{x},\epsilon^{(l)}) = \boldsymbol{\mu}(\mathbf{x};\phi) + \epsilon^{(l)}\odot\boldsymbol{\sigma}(\mathbf{x};\phi)$$

where $\epsilon^{(l)} \sim \mathcal{N}(\epsilon;\mathbf{0},\mathbf{I})$. The scalar $\alpha_{\text{vae}} = 2\sigma^2$ can be seen as a weighting factor between the KL-divergence and the reconstruction terms [267]. In practice, $L = 1$ sample is used per datapoint [266].

Combining Eqs. 6.14, 6.15 and 6.16, the SGVB estimator of the ELBO from Eq. 6.13 can also be written as:

$$\tilde{\mathcal{L}}(\theta,\phi;\mathbf{x}) = \frac{1}{2}\sum_{j=1}^{d_\mathbf{z}}\left[1+\log(\sigma_{z_j}^2) - \mu_{z_j}^2 - \sigma_{z_j}^2\right] - \frac{1}{L}\sum_{l=1}^{L}\frac{\|\mathbf{x}-\boldsymbol{\mu}(g_\phi(\mathbf{x},\epsilon^{(l)});\theta)\|^2}{\alpha_{\text{vae}}} \tag{6.17}$$

where $\epsilon^{(l)} \sim p(\epsilon)$. The lower bound $\tilde{\mathcal{L}}(\theta,\phi;\mathbf{x})$ in Eq 6.17 forms the objective function which can be optimized with respect to parameters θ and ϕ using a stochastic gradient descent algorithm.

6.2 Conditional variational auto-encoders

A conditional variational auto-encoder (CVAE) [268, 269] is a conditional generative model[2] (CGM) of the form $p_\theta(\mathbf{y},\mathbf{z}|\mathbf{x}) = p_\theta(\mathbf{z})p_\theta(\mathbf{y}|\mathbf{z},\mathbf{x})$. For a given input observation $\mathbf{x}$, a latent variable $\mathbf{z}$ is drawn from a prior distribution $p_\theta(\mathbf{z})$ from

[2]CVAEs are also referred to as conditional directed graphical models.

which the distribution $p_\theta(\mathbf{y}|\mathbf{z},\mathbf{x})$ generates the output $\mathbf{y}$. The purpose here is to maximise the conditional likelihood $p_\theta(\mathbf{y}|\mathbf{x})$ given by:

$$\begin{aligned} p_\theta(\mathbf{y}|\mathbf{x}) &= \int p_\theta(\mathbf{y},\mathbf{z}|\mathbf{x})d\mathbf{z} \\ &= \int p_\theta(\mathbf{z})p_\theta(\mathbf{y}|\mathbf{z},\mathbf{x})d\mathbf{z} \end{aligned} \tag{6.18}$$

Note that in contrast to VAEs (that maximise the marginal likelihood $p_\theta(\mathbf{x})$), CVAEs maximise the conditional likelihood $p_\theta(\mathbf{y}|\mathbf{x})$. Unfortunately, Eq. 6.18 is often intractable. The posterior $p_\theta(\mathbf{z}|\mathbf{y}) = \frac{p_\theta(\mathbf{y},\mathbf{z}|\mathbf{x})}{p_\theta(\mathbf{y}|\mathbf{x})}$ is therefore also intractable. To deal with the intractability, similar to VAEs, CVAEs also use an approximate posterior $q_\phi(\mathbf{z}|\mathbf{y})$[3]. The conditional likelihood $p_\theta(\mathbf{y}|\mathbf{x})$ of a datapoint $\mathbf{y}$ given a datapoint $\mathbf{x}$ is then written as follows:

$$\begin{aligned} \log p_\theta(\mathbf{y}|\mathbf{x}) &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{y})}[\log p_\theta(\mathbf{y}|\mathbf{x})] \\ &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{y})}\left[\log\left[\frac{p_\theta(\mathbf{y},\mathbf{z}|\mathbf{x})}{p_\theta(\mathbf{z}|\mathbf{y})}\right]\right] \\ &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{y})}\left[\log\left[\frac{q_\phi(\mathbf{z}|\mathbf{y})}{q_\phi(\mathbf{z}|\mathbf{y})}\frac{p_\theta(\mathbf{y},\mathbf{z}|\mathbf{x})}{p_\theta(\mathbf{z}|\mathbf{y})}\right]\right] \\ &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{y})}\left[\log\left[\frac{q_\phi(\mathbf{z}|\mathbf{y})}{p_\theta(\mathbf{z}|\mathbf{y})}\right]\right] + \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{y})}\left[\log\left[\frac{p_\theta(\mathbf{y},\mathbf{z}|\mathbf{x})}{q_\phi(\mathbf{z}|\mathbf{y})}\right]\right] \\ &= D_{KL}[q_\phi(\mathbf{z}|\mathbf{y})||p_\theta(\mathbf{z}|\mathbf{y})] + \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{y})}\left[\log\left[\frac{p_\theta(\mathbf{z})p_\theta(\mathbf{y}|\mathbf{z},\mathbf{x})}{q_\phi(\mathbf{z}|\mathbf{y})}\right]\right] \end{aligned}$$

Since the D_{KL} term is non-negative, the variational lower bound $\mathcal{L}(\theta,\phi;\mathbf{x},\mathbf{y})$ on the conditional likelihood $p_\theta(\mathbf{y}|\mathbf{x})$ is then given by:

$$\begin{aligned} \mathcal{L}(\theta,\phi;\mathbf{x},\mathbf{y}) &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{y})}\left[\log\left[\frac{p_\theta(\mathbf{z})p_\theta(\mathbf{y}|\mathbf{z},\mathbf{x})}{q_\phi(\mathbf{z}|\mathbf{y})}\right]\right] \\ &= E_{q_\phi(\mathbf{z}|\mathbf{y})}\left[\log\left[\frac{p_\theta(\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{y})}\right]\right] + E_{q_\phi(\mathbf{z}|\mathbf{y})}\left[\log p_\theta(\mathbf{y}|\mathbf{z},\mathbf{x})\right] \\ &= D_{KL}[q_\phi(\mathbf{z}|\mathbf{y})||p_\theta(\mathbf{z})] + E_{q_\phi(\mathbf{z}|\mathbf{y})}\left[\log p_\theta(\mathbf{y}|\mathbf{z},\mathbf{x})\right] \end{aligned} \tag{6.19}$$

[3]In our formulation, it is assumed that the true posterior is dependent only on $\mathbf{y}$, i.e., $p_\theta(\mathbf{z}|\mathbf{x},\mathbf{y}) = \frac{p_\theta(\mathbf{y},\mathbf{z}|\mathbf{x})}{p_\theta(\mathbf{y}|\mathbf{x})} = p_\theta(\mathbf{z}|\mathbf{y})$. The true posterior is then approximated by the inference model $q_\phi(\mathbf{z}|\mathbf{y})$. This is motivated by our task of interest, namely feature extraction/ dimensionality reduction for ABE (discussed in next Section). The graphical representation of the CVAE model is illustrated in Fig 6.3. Other CGM formulations [268, 269] (e.g., $p_\theta(\mathbf{y},\mathbf{z}|\mathbf{x}) = p_\theta(\mathbf{z}|\mathbf{x})p_\theta(\mathbf{y}|\mathbf{z},\mathbf{x})$) include a prior $p_\theta(\mathbf{z}|\mathbf{x})$ and a inference model $q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})$ as an approximation to the true posterior $p_\theta(\mathbf{z}|\mathbf{x},\mathbf{y})$. The choices of models are dependent on the task.

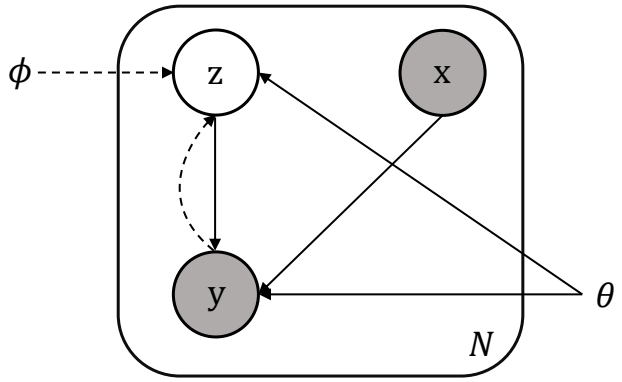


Figure 6.3: A conditional variational auto-encoder (CVAE) model as a conditional directed graphical model. The solid lines represent the generative model $p_\theta(\mathbf{y}, \mathbf{z}|\mathbf{x}) = p_\theta(\mathbf{z})p_\theta(\mathbf{y}|\mathbf{z}, \mathbf{x})$ with parameters θ. The dashed lines represent the inference of the true posterior $p_\theta(\mathbf{z}|\mathbf{y})$ performed via the variational approximation $q_\phi(\mathbf{z}|\mathbf{y})$ with parameters ϕ. The observed variables $\mathbf{x}$ and $\mathbf{y}$ are represented by the shaded nodes.

Here, similar to the VAE formulation, the prior is assumed to be a centred isotropic multivariate Gaussian distribution, i.e., $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I})$. The two conditional distributions $p_\theta(\mathbf{y}|\mathbf{z}, \mathbf{x})$ and $q_\phi(\mathbf{z}|\mathbf{y})$ are also assumed to be multivariate Gaussians given by $p_\theta(\mathbf{y}|\mathbf{z}, \mathbf{x}) = \mathcal{N}(\mathbf{y}; \boldsymbol{\mu}_\mathbf{y}, \sigma^2\mathbf{I})$ and $q_\phi(\mathbf{z}|\mathbf{y}) = \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_\mathbf{z}, \text{diag}(\boldsymbol{\sigma}_\mathbf{z}^2))$. The encoder network $q_\phi(\mathbf{z}|\mathbf{y})$ is modelled with a DNN with two output layers which define the mean $\boldsymbol{\mu}_\mathbf{z} = \boldsymbol{\mu}(\mathbf{y}; \phi)$ and variance $\boldsymbol{\sigma}_\mathbf{z}^\mathbf{2} = \boldsymbol{\sigma}^\mathbf{2}(\mathbf{y}; \phi)$. The decoder network $p_\theta(\mathbf{y}|\mathbf{z}, \mathbf{x})$ is modelled using another DNN whose output layer defines mean parameter $\boldsymbol{\mu}_\mathbf{y} = \boldsymbol{\mu}(\mathbf{z}, \mathbf{x}; \theta)$. Similar to the VAE framework, the first term of Eq. 6.19 can be calculated analytically whereas the second term is equivalent to the negative reconstruction error.

Finally, the SGVB estimator of the variational lower bound of Eq. 6.19 is given by:

$$\tilde{\mathcal{L}}(\theta, \phi; \mathbf{x}, \mathbf{y}) = \frac{1}{2}\sum_{j=1}^{d_\mathbf{z}} \left[1 + \log(\sigma_{z_j}^2) - \mu_{z_j}^2 - \sigma_{z_j}^2\right] - \frac{1}{L}\sum_{l=1}^{L} \frac{\|\mathbf{y} - \boldsymbol{\mu}(\mathbf{z}^{(l)}, \mathbf{x}; \theta)\|^2}{\alpha_{\text{cvae}}} \quad (6.20)$$

The expectation in the second term is calculated by a sampling operation that is performed via the *reparameterisation trick* such that:

$$\mathbf{z}^{(l)} = g_\phi(\mathbf{y}, \epsilon^{(l)}) = \boldsymbol{\mu}(\mathbf{y}; \phi)) + \epsilon^{(l)} \odot \boldsymbol{\sigma}(\mathbf{y}; \phi)$$

where $\epsilon^{(l)} \sim p(\epsilon) = \mathcal{N}(\epsilon; \mathbf{0}, \mathbf{I})$. The scalar $\alpha_{\text{cvae}} = 2\sigma^2$ acts as a weighting factor

between the KL-divergence and the reconstruction error.

6.3 Application to ABE

In this section we discuss how VAEs and CVAEs can be used for ABE.

6.3.1 Motivation

ABE algorithms exploit contextual information, or *memory*, to improve ABE performance. While the use of memory improves ABE performance, it implies the use of higher dimensional features and, therefore, more complex and computationally demanding ABE regression models. This is undesirable given that ABE is often required to function on battery-powered devices.

The work presented in Chapter 4 proposed an approach to include memory in the form of static features from neighbouring speech frames. Principal component analysis was used for dimensionality reduction and to preserve efficiency. The subsequent work in Chapter 5 showed that memory in the form of log spectral coefficients can be used to learn a compact, low dimensional NB feature representation for ABE using semi-supervised stacked auto-encoders (SSAEs). This section aims to explore the use of generative modelling techniques such as VAEs and CVAEs to further improve ABE performance. The goal is to model the distribution of higher-dimensional spectral data (that includes memory) and to extract higher-level, lower-dimensional features that improve the reliability of the ABE regression model, without affecting complexity. Essentially, we seek a form of dimensionality reduction that is tailored specifically to ABE.

Deep PGMs such as VAEs and CVAEs are capable of modelling complex data distributions. In contrast to bottleneck features learned by SAEs and SSAEs, the latent representation is probabilistic and can be used to generate new data. Inspired by their successful use in image processing [268,269,270], they have become increasingly popular in numerous fields of speech processing, e.g., speech modelling and transformation [271,272], voice conversion [273], speech synthesis [274], speech enhancement for voice activity detection [275], emotion recognition [276] and audio source separation [277].

CVAEs generate data via the combination of latent and so-called conditioning

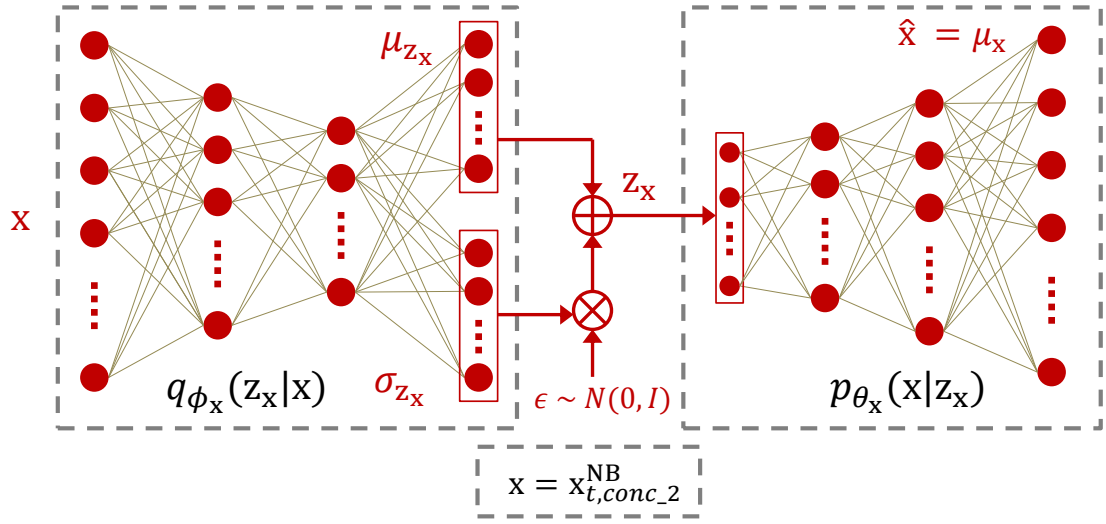
variables. The idea behind exploiting CVAEs for ABE in the framework of feature extraction is that the conditioning variable can be optimised via an auxiliary neural network in order to learn higher-level NB features, features that are tailored to the estimation of missing HB components in an ABE task. The novel contributions of this work are: (i) the first application of VAEs and CVAEs to dimensionality reduction for regression tasks such as ABE; (ii) the combination of CVAE with a probabilistic encoder in the form of an auxiliary neural network which derives the conditioning variable; (iii) an approach to their joint optimisation; (iv) their application to extract probabilistic NB latent representations for estimation of missing HB data in an otherwise standard ABE framework and (v) use of the proposed approach to deliver substantially improved ABE performance.

6.3.2 Extracting latent representations

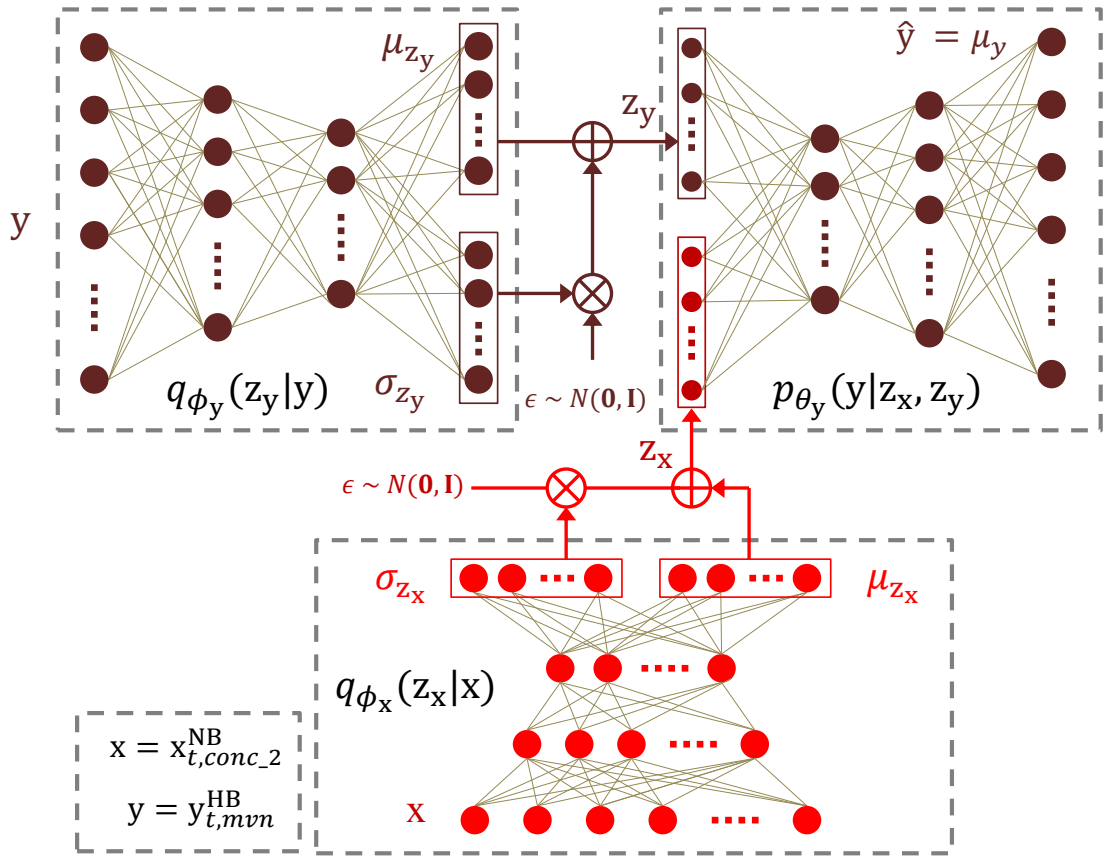
This section describes the proposed scheme to jointly optimise VAEs and CVAEs in order to learn latent representations tailored to ABE. The scheme is illustrated in Fig. 6.4. NB and HB features are extracted from parallel NB and WB training utterances, using frames of 20ms duration with 10ms overlap. 200-dimensional NB LPS coefficients P_t^{NB} obtained from 2 neighbouring frames are concatenated (after MVN) according to Eq. 4.4. The resulting input data $\mathbf{x} = \mathbf{x}_{t,conc\_2}^{\text{NB}}$ thus consists of 1000-dimensional NB features with memory. Refer to Section 4.3.1 for details relating to the calculation of P_t^{NB}. The output data $\mathbf{y} = \mathbf{y}_{t,mvn}^{\text{HB}}$ consists of 9 LP coefficients and a gain parameter (with MVN) extracted from parallel HB data via selective linear prediction (SLP).

First, the VAE is trained whereby the encoder $q_{\phi_\mathbf{x}}(\mathbf{z_x}|\mathbf{x})$ (left network of Fig. 6.4(a)) is fed with input data $\mathbf{x}$ in order to predict the mean $\boldsymbol{\mu}_{\mathbf{z_x}} = \boldsymbol{\mu}(\mathbf{x}; \theta_\mathbf{x})$ and log-variance $\log(\boldsymbol{\sigma}^2_{\mathbf{z_x}}) = \boldsymbol{\sigma}^2(\mathbf{x}; \theta_\mathbf{x})$ that represent the posterior distribution $q_{\phi_\mathbf{x}}(\mathbf{z_x}|\mathbf{x})$. A corresponding decoder $p_{\theta_\mathbf{x}}(\mathbf{x}|\mathbf{z_x})$ (right network in Fig. 6.4(a)) is fed with input $\mathbf{z_x} \sim q_{\phi_\mathbf{x}}(\mathbf{z_x}|\mathbf{x})$ in order to predict the mean $\boldsymbol{\mu}_\mathbf{x} = \boldsymbol{\mu}(\mathbf{z_x}; \phi_\mathbf{x})$ of the distribution $p_{\theta_\mathbf{x}}(\mathbf{x}|\mathbf{z_x})$. The latent variable $\mathbf{z_x}$ is sampled using $q_{\phi_\mathbf{x}}(\mathbf{z_x}|\mathbf{x})$ via the *reparameterisation trick* (see Section 6.1.2 and 6.1.4). Note that, at this stage (referred to as Stage-I), the NB representation $\mathbf{z_x}$ is learned without any supervision from the HB data. The VAE decoder is then discarded. The encoder $q_{\phi_\mathbf{x}}(\mathbf{z_x}|\mathbf{x})$ is then used as the conditioning variable of the CVAE (as shown in Fig. 6.4(b)).

The CVAE is then trained to model the distribution of the HB data $\mathbf{y}$ as



(a)



(b)

Figure 6.4: A feature extraction scheme using (a) VAE and (b) CVAE.

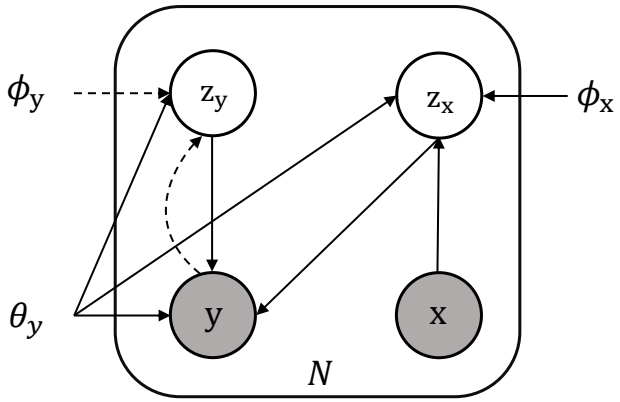


Figure 6.5: The proposed CVAE scheme as a conditional directed graphical model. The solid lines represent the generative model $p_\theta(\mathbf{y}, \mathbf{z_y}|\mathbf{z_x}) = p_\theta(\mathbf{z_y})p_\theta(\mathbf{y}|\mathbf{z_x}, \mathbf{z_y})$ with parameters θ. The dashed lines represent the inference of the true posterior $p_\theta(\mathbf{z_y}|\mathbf{y})$ performed via the variational approximation $q_\phi(\mathbf{z_y}|\mathbf{y})$ with parameters ϕ. The observed variables $\mathbf{x}$ and $\mathbf{y}$ are represented by the shaded nodes.

follows. $\mathbf{y}$ is fed to the encoder $q_{\phi_\mathbf{y}}(\mathbf{z_y}|\mathbf{y})$ (top-left network in Fig. 6.4(b)) in order to predict the mean $\boldsymbol{\mu}_{\mathbf{z_y}} = \boldsymbol{\mu}(\mathbf{y}; \theta_\mathbf{y})$ and log-variance $\log(\boldsymbol{\sigma}^2_{\mathbf{z_y}}) = \boldsymbol{\sigma}^2(\mathbf{y}; \theta_\mathbf{y})$ of the approximate posterior distribution $q_{\phi_\mathbf{y}}(\mathbf{z_y}|\mathbf{y})$. The predicted parameters are then used to obtain the latent representation $\mathbf{z_y} \sim q_{\phi_\mathbf{y}}(\mathbf{z_y}|\mathbf{y})$ of the output variable $\mathbf{y}$ via the *reparameterisation trick*. Next, the latent variable $\mathbf{z_x} \sim q_{\phi_\mathbf{x}}(\mathbf{z_x}|\mathbf{x})$ is used as the CVAE conditioning variable. After concatenation, $\mathbf{z_x}$ and $\mathbf{z_y}$ are fed to the decoder $p_{\theta_\mathbf{y}}(\mathbf{y}|\mathbf{z_x}, \mathbf{z_y})$ (top-right network in Fig. 6.4(b)) in order to predict the mean $\boldsymbol{\mu}_\mathbf{y} = \boldsymbol{\mu}(\mathbf{z_x}, \mathbf{z_y}; \theta_\mathbf{y})$ of the output variable $\mathbf{y}$. Finally, the entire network is trained to learn parameters $\phi_\mathbf{x}$, $\phi_\mathbf{y}$ and $\theta_\mathbf{y}$ jointly. The graphical representation of the proposed CVAE scheme is illustrated in Fig. 6.5. From Eq. 6.19 and 6.20, the equivalent variational lower bound under optimisation is given by:

$$\begin{aligned}
\log p_{\theta_\mathbf{y}}(\mathbf{y}|\mathbf{z_x}) &\geq \mathcal{L}(\theta_\mathbf{y}, \phi_\mathbf{y}, \phi_\mathbf{x}; \mathbf{z_x}, \mathbf{y}) \\
&= -D_{KL}[q_{\phi_\mathbf{y}}(\mathbf{z_y}|\mathbf{y})||p_{\theta_y}(\mathbf{z_y})] + E_{q_{\phi_\mathbf{y}}(\mathbf{z_y}|\mathbf{y})}\left[\log p_{\theta_\mathbf{y}}(\mathbf{y}|\mathbf{z_x}, \mathbf{z_y})\right] \\
&= \frac{1}{2}\sum_{j=1}^{d_{\mathbf{z_y}}}\left[1 + \log(\sigma^2_{z_{y_j}}) - \mu^2_{z_{y_j}} - \sigma^2_{z_{y_j}}\right] - \frac{1}{L}\sum_{l=1}^{L}\frac{\|\mathbf{y} - \boldsymbol{\mu}(\mathbf{z_y}^{(l)}, \mathbf{z_x}; \theta)\|^2}{\alpha_{\text{cvae}}}
\end{aligned} \tag{6.21}$$

where $\mathbf{z_y}^{(l)} = g_{\phi_\mathbf{y}}(\mathbf{y}, \epsilon^{(l)}) = \boldsymbol{\mu}(\mathbf{y}; \phi_\mathbf{y})) + \epsilon^{(l)} \odot \boldsymbol{\sigma}(\mathbf{y}; \phi_\mathbf{y})$ and $\epsilon^{(l)} \sim p(\epsilon) = \mathcal{N}(\epsilon; \mathbf{0}, \mathbf{I})$, $d_{\mathbf{z_y}}$ is the dimensionality of the latent variable $\mathbf{z_y}$, $\mu_{z_{y_j}}$ and $\sigma_{z_{y_j}}$ denote the j^{th} element of the vectors $\boldsymbol{\mu}_{\mathbf{z_y}}$ and $\boldsymbol{\sigma}_{\mathbf{z_y}}$ respectively.

It is expected that, during optimisation of Eq. 6.21, parameters $\phi_{\mathbf{x}}$ of the encoder $q_{\phi_{\mathbf{x}}}(\mathbf{z_x}|\mathbf{x})$ are updated so that the proposed CVAE framework encodes the higher-dimensional NB data $\mathbf{x}$ into the latent representation $\mathbf{z_x}$ that aids the reconstruction of the HB data $\mathbf{y}$. From another perspective, at this stage (referred to as Stage-II), the lower-dimensional representation $\mathbf{z_x}$ is learned using some form of supervision from the HB data. It therefore encodes some useful information about the HB representation.

Finally, the encoder $q_{\phi_{\mathbf{x}}}(\mathbf{z_x}|\mathbf{x})$ (signified by the red components in Fig. 6.4(b)) is used, instead of the weight matrix $\mathbf{W}_{\text{PCA}}$ in Fig. 4.3 (red boxes), to estimate the latent representation $\mathbf{z_x}$ for every $\mathbf{x}$. The GMM training is then performed using joint vectors $\mathbf{z_x}$ and $\mathbf{y}$ followed by the estimation in the same manner as described in Section 4.4. Note that the latent representation $\mathbf{z_x}$ extracted using the encoder $q_{\phi_{\mathbf{x}}}(\mathbf{z_x}|\mathbf{x})$ at training stages I and II represent the features extracted using VAE and CVAE architectures.

6.3.3 Direct estimation using CVAE-DNN

The networks $q_{\phi_{\mathbf{x}}}(\mathbf{z_x}|\mathbf{x})$ and $p_{\theta_{\mathbf{y}}}(\mathbf{y}|\mathbf{z_x}, \mathbf{z_y})$ of the proposed CVAE architecture shown in Fig. 6.4(b) together form a DNN (referred to as a CVAE-DNN), with two stochastic layers $\mathbf{z_x}$ and $\mathbf{z_y}$. It can itself be used for ABE. In this case, there is some discrepancy between *training* and *testing* phases since the HB output vector $\mathbf{y}$ is available during training but not during testing or estimation. The schematic of CVAE-DNN is illustrated in Fig. 6.6.

While during the *training* phase the inference model $q_{\phi_{\mathbf{y}}}(\mathbf{z_y}|\mathbf{y})$ is used to sample the latent variables $\mathbf{z_y}$ (for reconstruction of $\mathbf{y}$ using the decoder $p_{\theta_{\mathbf{y}}}(\mathbf{y}|\mathbf{z_x}, \mathbf{z_y})$), the CVAE uses the prior distribution $p_{\theta_y}(\mathbf{z_y}) = \mathcal{N}(\mathbf{z_y}; \mathbf{0}, \mathbf{I})$ (for prediction of $\mathbf{y}$) during the *testing* phase. The training and testing phases thus also correspond to *reconstruction* and *prediction* phases of the output $\mathbf{y}$ respectively.

6.4 Experimental setup and results

Experiments are designed to compare the performance of an ABE system that uses features learned from CVAE with systems that use alternative dimensionality reduction techniques. In all cases, performance is assessed with and without mean

and variance normalisation (MVN). The experimental setup, including databases and metrics, is same as that used for the work presented in the previous chapter.

6.4.1 CVAE configuration and training

The CVAE architecture is implemented using the Keras toolkit [264]. Encoders $q_{\phi_{\mathbf{x}}}(\mathbf{z_x}|\mathbf{x})$ and $q_{\phi_{\mathbf{y}}}(\mathbf{z_y}|\mathbf{y})$ consist of two hidden layers with 512 and 256 units, and input layers with 1000 and 10 units respectively. Their outputs are Gaussian-distributed latent variable layers $\mathbf{z_x}$ and $\mathbf{z_y}$ consisting of 10 units for the means $\mu_{\mathbf{z_x}}$, $\mu_{\mathbf{z_y}}$ and log-variances $\log(\sigma^2_{\mathbf{z_x}})$, $\log(\sigma^2_{\mathbf{z_y}})$. The decoders $p_{\theta_{\mathbf{x}}}(\mathbf{x}|\mathbf{z_x})$ and $p_{\theta_{\mathbf{y}}}(\mathbf{y}|\mathbf{z_x}, \mathbf{z_y})$ have 2 hidden layers with 256 and 512 units. Output layers have 1000 and 10 units respectively. All hidden layers have *tanh* activation units whereas Gaussian parameter layers have *linear* activation units. The modelling of log-variances avoids the estimation of negative variances. The CVAE architectures thus have a structure of (512, 256, 10+10, 256, 512) hidden units in accordance with the best performing SSAE architecure (Arch-1C)[4] presented in previous Chapter.

Training is performed jointly in order to minimise the negative conditional log-likelihood in Eq. 6.21 using the Adam stochastic optimisation technique [265] with an initial learning rate of 10^{-3} and hyperparameters $\beta_1 = 0.9, \beta_2 = 0.999$ and $\epsilon = 10^{-8}$. Networks are initialised according to the approach described in [254] so as to improve the rate of convergence. To discourage over-fitting, batch-normalisation [262] is applied before every activation layer. The learning rate is reduced by half when the validation loss increases between 5 consecutive epochs. First, the VAE is trained on input data $\mathbf{x}$ for 50 epochs. The full CVAE is then trained for a further 50 epochs using input $\mathbf{x}$ and output $\mathbf{y}$ data. The model giving the lowest validation loss is used for subsequent processing.

CVAE performance is compared to alternative SAE, SSAE and PCA dimensionality reduction techniques. While the encoder of all architectures (that is used for dimensionality reduction) has a common structure of (1000, 512, 256, 10) units[5], the PCA transformation consists of a weight matrix $\mathbf{W}_{\text{PCA}}$ with size 1000×10.

[4] As shown in Table 5.3, the difference between results obtained using SSAE architectures Arch-1C and Arch-2C with *tanh* activations is insignificant and therefore, Arch-1C is chosen for further experiments as it has relatively lower complexity.

[5] VAE and CVAE architectures have one extra output layer with 10 units that models log-variances.

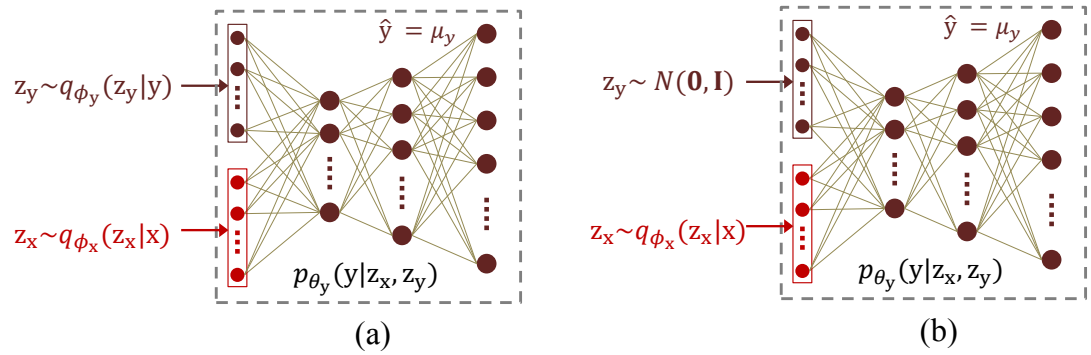


Figure 6.6: A schematic of CVAE-DNN, a DNN formed using a CVAE with stochastic layers $\mathbf{z_x}$ and $\mathbf{z_y}$ during (a) *training* (or *reconstruction*) and (b) *testing* (or *prediction*) phases.

6.4.2 Analysis of weighting factor α

According to the previous discussion, the HB output can be estimated either using the standard GMMR (where the latent representation $\mathbf{z_x}$ is used as an input) or using the CVAE-DNN (see Section 6.3.3). Since better estimation of HB components is crucial to ABE performance, the latent representation $\mathbf{z_x}$ should contain information that is informative of $\mathbf{y}$. We therefore studied the importance of the weighing factor α on the reconstruction error (RE), $\|\mathbf{y} - f(\mathbf{z_y}, \mathbf{z_x}; \theta_\mathbf{y})\|^2$, between the true and estimated HB data using both estimation methods.

For this purpose, VAE and CVAE networks are trained for different values of α_{vae} and α_{cvae}. During training, the network weights with the best validation loss are stored and then used for further processing.

CVAE-DNN

The *trained* CVAE-DNN is used to estimate the D_{KL} and RE values on the validation data, both during *reconstruction* ($\mathbf{z_y} \sim q_{\phi_\mathbf{y}}(\mathbf{z_y}|\mathbf{y})$) and *prediction* ($\mathbf{z_y} \sim p_{\theta_y}(\mathbf{z_y}) = \mathcal{N}(\mathbf{z_y}; \mathbf{0}, \mathbf{I})$) phases[6]. The observations that can be made from the results illustrated in Table 6.1 are as follows:

- Lower values of α ($= 1$) lead to higher values of D_{KL} ($= 2.01$), suggesting that the approximate posterior $q_{\phi_\mathbf{y}}(\mathbf{z_y}|\mathbf{y})$ is far from the prior distribution

[6]Note that the terms *reconstruction* and *prediction* refer to the use of the approximate posterior or prior for the sampling of $\mathbf{z_y}$ (see Section 6.3.3).

Table 6.1: Effect of weighing factor α_{cvae} on D_{KL} and RE during both *training* (or *reconstruction*) and *testing* (or *prediction*) phases for CVAE-DNN. Results shown for the validation dataset.

α_{cvae}	1	2	5	10	20	30
D_{KL} reconstruction	2.01	0.96	0.21	3.3e-4	1.5e-4	9.7e-5
RE reconstruction	3.15	4.73	7.40	8.93	8.97	8.97
RE prediction	12.75	11.40	9.85	8.93	8.97	8.97

$p_{\theta_y}(\mathbf{z_y}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$. This hypothesis is confirmed by the observation of higher REs during *reconstruction* (RE = 12.75) than during *prediction* (RE = 3.15). This is because the decoder $p_{\theta_\mathbf{y}}(\mathbf{y}|\mathbf{z_x}, \mathbf{z_y})$ reconstructs the output $\mathbf{y}$ using latent variables $\mathbf{z_y}$ sampled from the prior during *prediction* phase, but from the approximate posterior during *reconstruction* phase.

- Higher values of α (= 30) give lower values of D_{KL} (= $9.7e{-}5$), suggesting that the posterior distribution is closer to the prior distribution. This hypothesis is confirmed by the observation of similar REs (= 8.97) for *reconstruction* and *prediction* phases. These findings corroborate those of previous work [272].

GMMR

The performance of ABE in terms of RE is also investigated using the standard GMMR technique employed in this thesis. Table 6.2 shows the RE values obtained when the latent representation $\mathbf{z_x}$ obtained from both VAE (via the encoder $q_{\phi_\mathbf{x}}(\mathbf{z_x}|\mathbf{x})$ at stage-I) and CVAE (via the encoder $q_{\phi_\mathbf{x}}(\mathbf{z_x}|\mathbf{x})$ at stage-II) architectures are fed to the GMMR. Note that the RE values obtained in this case are equivalent to those obtained using CVAE-DNN during *prediction* phase (fourth row in Table 6.1).

While the RE values for VAE system (with and without MVN) increase with the increase in $\alpha = \alpha_{\text{vae}}$ value (second row in Table 6.2), the RE performance for CVAE system is consistent (third row in Table 6.2) for all values of $\alpha = \alpha_{\text{cvae}}$. The estimation performance of CVAE system, using a standard regression model, in

Table 6.2: Effect of weighing factor α_{vae} and α_{cvae} on RE in case of estimation using GMMR. Results shown for the validation dataset.

α	1	2	5	10	20	30
VAE	9.10	9.11	9.19	9.39	9.58	9.68
VAE + MVN	9.06	9.10	9.18	9.39	9.58	9.68
CVAE	8.91	8.92	8.93	9.00	9.02	9.02
CVAE + MVN	8.94	8.94	8.95	8.94	8.99	8.98

terms of RE is thus robust to the value of α. This is in contrast to the results obtained using the CVAE-DNN where the RE performance during *prediction* phase is inconsistent (fourth row in Table 6.1) for different values of α.

The aim of the work reported in this chapter is to use the latent representation $\mathbf{z_x}$ learned using a CVAE as a NB feature extraction technique for ABE. Consistent with the work presented in previous chapters ABE performance is thus analysed in the context of dimensionality reduction. All experiments reported in the remainder of this chapter correspond to a value of $\alpha_{\text{vae}} = 10$ and $\alpha = 10^7$.

6.4.3 Objective assessment

Consistent with the previous chapters, objective assessment is performed by evaluation of speech quality using the two distance metrics $d_{\text{RMS-LSD}}$, d_{COSH} and objective estimates of MOS scores MOS-LQO$_{\text{WB}}$ (Section 3.4.2). The lower values of distance metrics and higher values of MOS-LQO$_{\text{WB}}$ indicate better speech quality. Objective results are presented in Table 7.1. The key observations are discussed in the following:

- The performance of PCA ABE system in terms of objective measures $d_{\text{RMS-LSD}}$, d_{COSH}, MOS-LQO$_{\text{WB}}$ is degraded in case of MVN (6.95 $\rightarrow$ 7.35dB, 1.43 $\rightarrow$ 1.45dB, 3.21 $\rightarrow$ 3.14). In contrast, MVN improves performance for SAE (12.45 $\rightarrow$ 7.54dB, 2.96 $\rightarrow$ 1.50dB, 1.95 $\rightarrow$ 3.03) and SSAE (10.50 $\rightarrow$ 6.80dB,

[7]From the results shown in Table 6.2 it is evident that the choice of α_{cvae} does not affect the RE performance of CVAE feature extraction system using GMMR. While VAE system achieves the lowest RE value (9.06 with MVN and 9.10 without MVN) for $\alpha = 1$, it is still inferior to the CVAE performance. Thus, for simplicity, we fix the value α to 10 in our further analysis for both VAE and CVAE systems.

2.11 → 1.34dB, 2.26 → 3.28) techniques significantly. This is because the features learned with conventional auto-encoders may not be orthonormal and uncorrelated.

- ABE performance with PCA dimensionality reduction outperforms that with SAE and VAE techniques, signifying the importance of supervised learning or so-called discriminative fine tuning during feature extraction. PCA ABE system shows better speech quality, in terms of objective measures ($d_{\text{RMS-LSD}}$, d_{COSH}, MOS-LQO$_{\text{WB}}$) by 2.6%, 3.4%, 3.5% compared to SAE + MVN system and by 17.6%, 11.7%, 9% compared to VAE system respectively. The performance of VAE system is consistent with and without MVN. It may provide better results for other values of α (e.g., for $\alpha = 1$).

- SSAE + MVN ABE system outperforms PCA system relatively by %7.5 (7.35 → 6.80dB), %11 (1.45 → 1.34dB), %8 (3.14 → 3.28). This is expected due to the use of semi-supervised feature extraction via SSAE architecture (Section 5.2).

- The CVAE ABE system is the best performing of all and, interestingly, performance is stable with and without MVN. This is likely due to the *probabilistic* learning of latent representations. CVAE ABE system outperforms PCA system relatively by %10.3 (7.35 → 6.59dB), %11 (1.45 → 1.31dB), %8 (3.14 → 3.34). The relative performance improvement of CVAE system w.r.t. SSAE + MVN system is given by %3.1 (6.80 → 6.59dB), %2.2 (1.34 → 1.31dB), %1.8 (3.28 → 3.34).

6.4.4 Subjective assessment

Speech quality performance is also assessed via comparative, subjective CCR listening tests. Tests were performed by 15 listeners who were asked to compare the quality of 12 pairs of speech signals A and B listened to using DT 770 PRO headphones. They were asked to rate the quality of signal A with respect to B according to the following scale: -3 (much worse), -2 (worse), -1 (slightly worse), 0 (about the same), 1 (slightly better), 2 (better), 3 (much better). For further details of CCR listening tests, refer to Section 3.4.1.

The results of subjective tests are illustrated in Table 6.4 in the form of CMOS with corresponding 95% confidence interval (CI$_{95}$). Speech files whose bandwidth is

Table 6.3: Objective assessment results (with mean and standard deviation values). RMS-LSD and d_{COSH} are distance measures (lower values indicate better performance) in dB, whereas MOS-LQO$_{\text{WB}}$ values reflect quality (higher values indicate better performance).

Dimensionality reduction method	$d_{\text{RMS-LSD (dB)}}$	$d_{\text{COSH (dB)}}$	MOS-LQO$_{\text{WB}}$
PCA	6.95 (0.68)	1.43 (0.40)	3.21 (0.35)
PCA + MVN	7.35 (0.70)	1.45 (0.36)	3.14 (0.34)
SAE	12.45 (1.42)	2.96 (0.71)	1.95 (0.20)
SAE + MVN	7.54 (0.74)	1.51 (0.36)	3.03 (0.37)
SSAE	10.50 (1.12)	2.11 (0.48)	2.26 (0.24)
SSAE + MVN	6.80 (0.66)	1.34 (0.32)	3.28 (0.35)
VAE	8.64 (0.79)	1.67 (0.35)	2.75 (0.35)
VAE + MVN	8.60 (0.79)	1.67 (0.35)	2.75 (0.35)
CVAE	**6.59** (0.66)	**1.31** (0.36)	**3.34** (0.35)
CVAE + MVN	**6.69** (0.68)	**1.30** (0.35)	**3.31** (0.35)

Table 6.4: Subjective assessment results for the ABE systems with CVAE, SSAE + MVN and PCA dimensionality reduction techniques in terms of CMOS points with corresponding 95% confidence interval (CI$_{95}$).

Comparison A → B	CMOS [CI$_{95}$]
CVAE → NB	0.90 [0.65; 1.16]
CVAE → PCA	0.13 [0.02; 0.25]
CVAE → SSAE + MVN	0.10 [-0.02; 0.22]
CVAE → WB	-0.96 [-1.16; -0.77]

extended using the proposed CVAE approach were judged to be of superior quality in comparison to original NB signals with a significant CMOS of 0.90. CVAE system produces speech of better quality than PCA and SSAE + MVN systems with CMOS of 0.13 and 0.10. The quality of extended speech signals is still inferior to original WB signals, reflecting from a CMOS of -0.96 points. This shows the scope for further improvements.

6.5 Summary

Conditional variational auto-encoders (CVAE) are directed graphical models that are used for generative modelling. This chapter presents their first application to dimensionality reduction for computationally efficient artificial bandwidth extension (ABE). The key contributions of the work are as follows:

- The strength of CVAEs to model highly complex data distributions is exploited to extract probabilistic, low-dimensional narrowband features from high-dimensional log spectral coefficients with memory.
- The latent representations produced using the proposed approach are shown to improve ABE performance using a standard regression model in comparison to other dimensionality reduction techniques, confirmed by both objective and subjective assessments.
- The probabilistic features are learned in a data-driven manner via generative modelling that does not need any post-processing such as mean and variance normalisation. This is in contrast to bottleneck features learned using conventional stacked auto-encoders.
- Improvements are attributed to the better modelling of high-dimensional spectral data using CVAEs. Crucially, they are achieved without augmenting the complexity of the regression model.

Future work should compare or combine CVAEs with other generative models such as adversarial networks. Better CVAEs training strategies should bring further improvements to ABE performance. A thorough comparison of ABE performance using CVAE with other dimensionality reduction techniques such as robust PCA (RPCA) is also of interest.

Chapter 7

Super-wide bandwidth extension

Many smart devices now support high-quality speech communication services at super-wide bandwidths. Often, however, speech quality is degraded when they are used with networks or devices which lack super-wideband support. Artificial bandwidth extension can then be used to improve speech quality. While approaches to wideband extension have been reported the previous chapters, this chapter proposes an approach to *super*-wide bandwidth extension (SWBE). The algorithm is based upon a classical source filter model in which the spectral envelope and residual error information are extracted from a wideband signal using conventional linear prediction analysis. A form of spectral mirroring is then used to extend the residual error component before an extended super-wideband signal is derived from its combination with the *original wideband envelope*. Experiments confirm improvements to speech quality via both objective and subjective assessments. These show that the quality of super-wideband speech, derived from the bandwidth extension of wideband speech, is comparable to that of speech processed with the standard enhanced voice services (EVS) codec with a bitrate of 13.2kbps. Without the need for statistical estimation of missing super-wideband components, the proposed algorithm is highly efficient.

This chapter is organised as follows. Section 7.1 describes the motivation for the work reported. Section 7.2 presents a brief review of related, past work. Section 7.3 describes the proposed SWBE algorithm. The spectral envelope extension method is explained in detail in Section 7.4. Section 7.5 describes the experimental setup and both subjective and objective assessments. A summary of all the work and contribution is presented in Section 7.6.

7.1 Motivation

The quality of speech offered by modern communications systems and devices has improved enormously in recent times. Whereas many devices were, and continue to be restricted to narrow and wide bandwidths, today's technology such as the enhanced voice services (EVS) codec [51, 52] (Section 1.3.3) developed by the 3rd Generation Partnership Project (3GPP) in 2014, increasingly supports communications at super-wide bandwidths. When used with other devices and networks with compatible support for super-wideband (SWB) services, such technology offers extremely high quality communications.

Often, though, SWB devices are used with other devices and networks which support only narrowband (NB) or wideband (WB) communications. Typically, AMR-NB [31] (Section 1.3.1) and AMR-WB [35] (Section 1.3.2) codecs are used during NB and WB calls respectively. While they usually offer backward compatibility, users of SWB devices will then be restricted to NB or WB communications. A reduction in bandwidth accompanies a reduction in speech quality. Here too, there is potential to improve quality. The extensive body of ABE research in the literature involves the extension of NB to WB speech signals which is the focus of research in earlier chapters of this thesis. In these cases there is substantial potential to improve quality; significant speech components between the NB limit of 3.4 (or 4) kHz and the WB limit of 7 (or 8) kHz can be recovered reliably using ABE. SWB speech signals extend the limit to 16kHz. Super-wide bandwidth extension (SWBE) approaches can then be employed to recover missing high frequency (HF) components between 7 (or 8) kHz and 16kHz from available low frequency (LF) components between 50 Hz to 7 (or 8) kHz.

This chapter presents an efficient approach to SWBE. It is based upon a classical source-filter model in which a WB signal is extended using conventional linear prediction (LP) analysis.

7.2 Past work

Only few approaches to SWBE are reported in the literature (Section 2.8.2). This is perhaps because the potential gain in quality for speech signals from the extension of WB to SWB is much less than the potential when extending from NB to WB. As a result, even modest processing artefacts can no longer be tolerated. Most of the

existing solutions use statistical estimation models which can be computationally demanding for real-time implementations.

In the SWBE approach presented in [210], the spectral content from 6 to 7kHz of a WB speech signal is inserted into the 8 to 12kHz frequency band via a time-scaling operation. This is done via the direct manipulation of DFT coefficients corresponding to the 6-7kHz frequency range without relying on statistical estimation. The proposed approach is shown to outperform the HMM-based SWBE method in [16] while maintaining relatively lower computational complexity.

A generic approach for efficient high-frequency bandwidth extension (EHBE) is proposed in [208] for music and speech signals. The missing HF components are estimated from those in the highest octave of the WB signal. The approach exploits the properties of non-linear operations (e.g., full-wave rectification) to generate harmonics in the HF range. While improvements in quality are reported, the use of non-linear processing typically tends to produce audible intermodulation distortion. However, subjective assessment results for the SWBE approaches reported in [56, 203, 205] – the methods which require the statistical estimation of missing HF components – show that performance is mostly comparable to that of the EHBE algorithm. These works show that SWBE approaches proposed in [208, 210] are thus appealing alternatives to the approaches which are based on statistical estimation techniques, at least for SWBE (WB-to-SWB extension). This is because they have lower computational complexity while maintaining comparable speech quality. Since it performs as well as more recent techniques while not requiring statistical estimation procedure, the EHBE algorithm is used as a baseline approach in this work.

7.3 Super-wide bandwidth extension (SWBE)

A block diagram of the proposed approach to SWBE is presented in Fig. 7.1. There are four key components. First (box 1), the WB input signal $\mathbf{x}_{\text{wb}}$ is windowed for subsequent frame-by-frame processing. Second (box 2), missing HF components are estimated from available LF components. Third (box 3), the original LF components are extracted from the input WB frame. Finally (box 4), an extended SWB output signal $\hat{\mathbf{x}}_{\text{swb}}$ is obtained by combining LF and HF components.

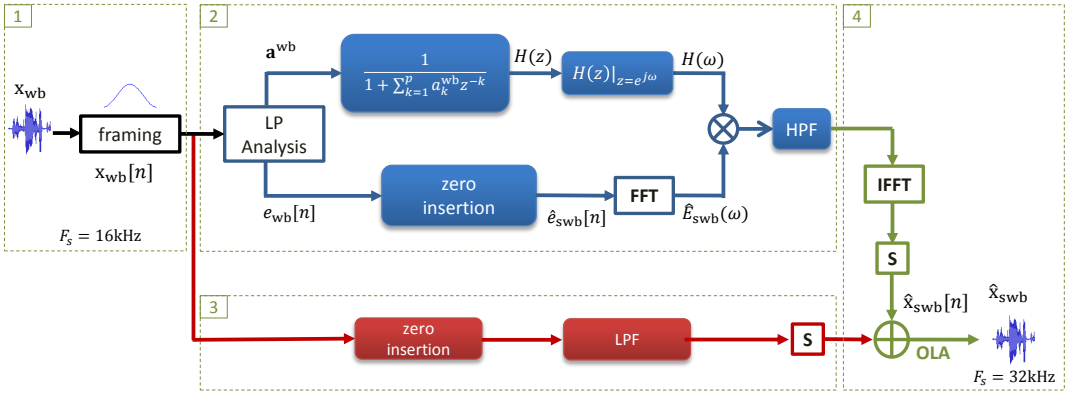


Figure 7.1: A block diagram of the proposed approach to super-wide bandwidth extension (SWBE).

7.3.1 High frequency component estimation

The HF component of the input WB signal sampled at 16 kHz is estimated frame-by-frame via the blue-coloured components illustrated in Fig. 7.1 (box 2). Speech frames $x_{\text{wb}}[n]$ are processed to obtain standard linear prediction (LP) coefficients $\mathbf{a}^{\text{wb}}$ and the residual component $e_{\text{wb}}[n]$ with conventional LP analysis of order $p = 16$. The LP coefficients are used to determine the frequency response $H(\omega)$ from the transfer function $H(z)$, which characterises the spectral envelope of the WB signal.

The residual component $\hat{e}_{\text{swb}}[n]$ is extended by zero insertion in the time domain to obtain $e_{\text{wb}}[n]$. As a form of spectral mirroring, the operation is equivalent to an up-sampling operation without an anti-aliasing filter [131]. The complex frequency domain representation of the excitation signal $\hat{E}_{\text{swb}}(\omega)$ is obtained from the extended residual $\hat{e}_{\text{swb}}[n]$ using the fast Fourier transform (FFT) and then combined by multiplication with the filter/envelope $H(\omega)$. Since the output is a composite of estimated HF components and distorted LF components, the latter are removed via high pass filtering (HPF), thereby preserving HF components only.

7.3.2 Low frequency component upsampling

The LF component of the input signal $\mathbf{x}_{\text{wb}}$ is also extracted frame-by-frame. The processing involved is illustrated by the red-coloured components in Fig. 7.1 (box 3). Each frame $x_{\text{wb}}[n]$ is up-sampled in the time domain using zero insertion. An anti-aliasing low pass filter (LPF) is then applied. The result is an interpolated time

domain signal at a sampling rate of 32kHz comprising only frequency components below 8kHz. Typically, this operation is common to all bandwidth extension algorithms.

7.3.3 Resynthesis

Resynthesis of the extended output $\hat{\mathbf{x}}_{\text{swb}}$ is performed via the green-coloured elements of Fig. 7.1 (box 4). A time domain frame containing only estimated HF components is obtained via the inverse FFT (IFFT). After synchronisation (S) to compensate for delays introduced by the different processes involved in the estimation of LF and HF components, a SWB speech frame with a sampling frequency of 32kHz is obtained from their addition. Synchronisation is also a component of every approach to bandwidth extension. Resynthesis is accomplished using a conventional overlap-add (OLA) [216, Section 12.1.1], [217, Section 5.3.1] technique in order to avoid discontinuities at frame edges.

7.4 Spectral envelope extension

The extension of WB spectral envelope in the proposed SWBE approach is performed by extrapolation. This operation is simply performed by upsampling of residual component via zero insertion while keeping the WB spectral envelope $H(\omega)$ unchanged. This section explains the procedure of spectral envelope estimation for the proposed SWBE algorithm.

7.4.1 Effect of sampling frequency

In order to understand the spectral envelope estimation employed in the proposed SWBE approach, first the effect of the sampling rate on the *effective* frequency response of a filter is explained. A typical system for the processing of continuous-time signals is illustrated in Fig 7.2(a) which comprises a cascade of a continuous-to-discrete-time (C/D) converter, a discrete-time system $h[n]$ and a discrete-to-continuous-time converter. The overall system (shown in the dashed-rectangular box) converts the continuous-time input signal $x_c(t)$ to the continuous-time output signal $y_r(t)$. Its properties are dependent on the choices of the discrete-time system $h[n]$ and the sampling rate $F_s = 1/T_s$ of the C/D and D/C converters [278, Section

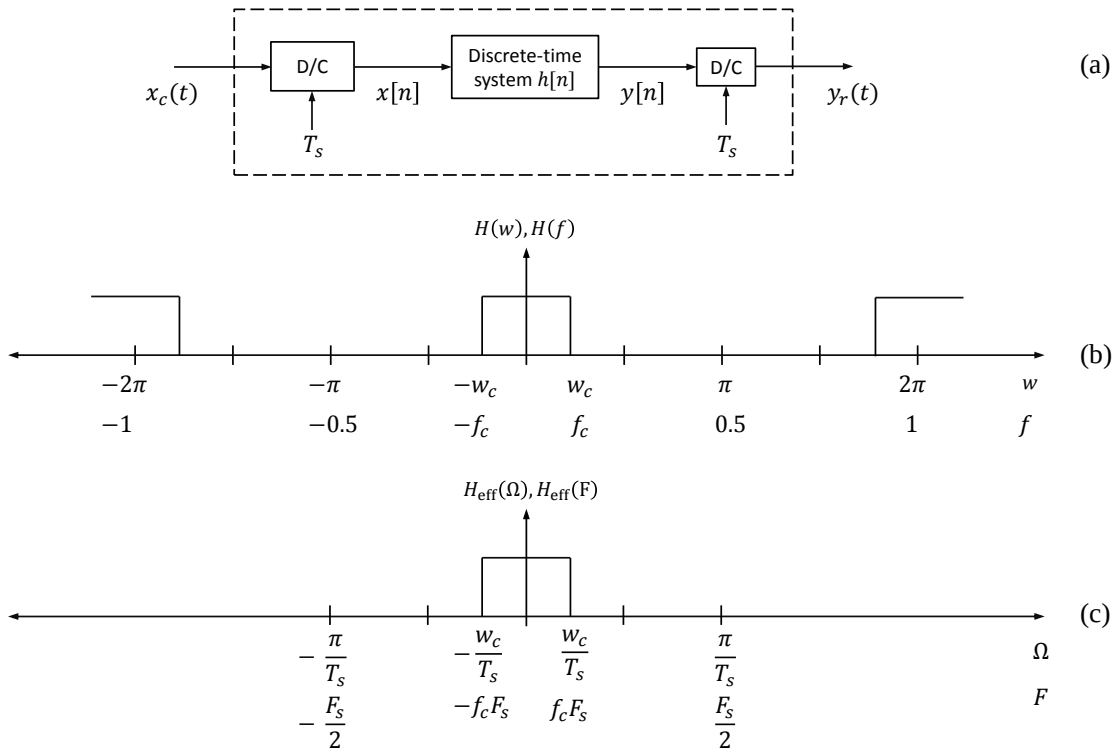


Figure 7.2: (a) Discrete-time processing of continuous-time signals. (b) Frequency response of the discrete-time system $h[n]$. (c) Corresponding *effective* continuous-time frequency response for the bandlimited input $x_c(t)$. Adapted from [278, Section 4.4].

4.4].

The overall system is equivalent to a linear time-invariant (LTI), continuous-time system with effective frequency response (shown in Fig 7.2(b) and (c)) given according to:

$$H_{eff}(\Omega) = \begin{cases} H(w), & |\Omega| < \pi/T_s \\ 0, & |\Omega| \geq \pi/T_s. \end{cases} \tag{7.1}$$

where Ω (radians per second) and ω (radians per sample) are the frequencies[1] for continuous-time and discrete-time signals. They are related according to: $\Omega = \omega/T_s$ [279, Section 1.4.1].

Therefore, the sampling frequency F_s – at which the signal $x_c(t)$ is sampled

[1]Instead of Ω and ω, continuous-time (or analogue) and discrete-time frequencies are also represented by F (cycles per second or hertz (Hz)) and f (cycles per sample) respectively where $\Omega = 2\pi F$, $\omega = 2\pi f$ [279, Section 1.3].

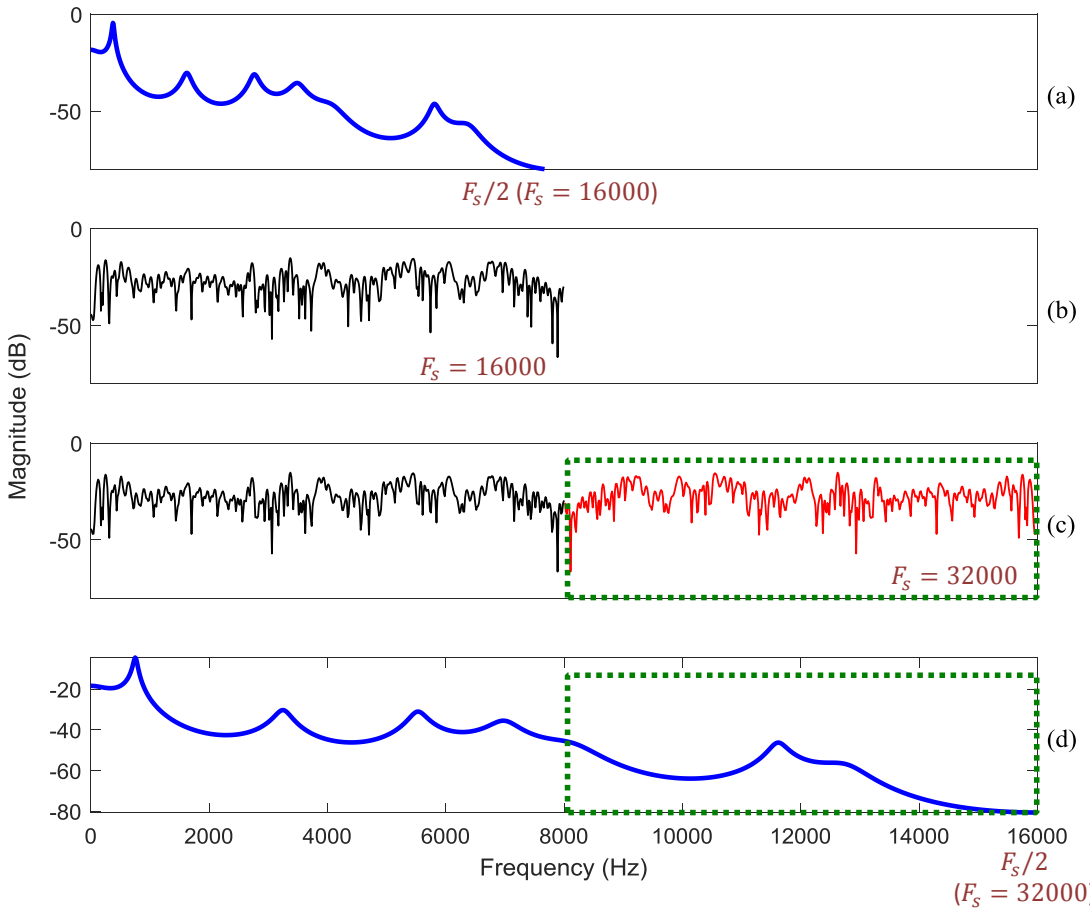


Figure 7.3: Illustration of the envelope extension process for an arbitrary voiced speech frame. (a) wideband (WB) spectral envelope represented by the filter $H(\omega)$, (b) spectrum of residual component $e_{\text{wb}}[n]$, (c) spectrum of the upsampled excitation component $\hat{e}_{\text{swb}}[n]$, (d) the *effective* frequency response of the filter $H(\omega)$ for the input $\hat{e}_{\text{swb}}[n]$.

to obtain the discrete-time signal $x[n]$ – determines the effect of the frequency response $H(w)$ on the output signal $y_r(t)$.

7.4.2 Extension

As discussed above, the effective frequency response $H(\omega)$ of a filter depends on the sampling frequency (F_s) of the input signal. The proposed SWBE approach exploits this property to generate missing HF components.

Illustration of the envelope extension process are shown in Fig 7.3 for an

arbitrary voiced speech frame. During *analysis*, a WB speech frame is represented by a vector of WB LP coefficients $\mathbf{a}^{\text{wb}}$ and residual error component $e_{\text{wb}}[n]$ according to the source-filter model. Fig 7.3(a) illustrates the frequency response of the vocal tract filter defined by the WB LP coefficients according to:

$$H(\omega) = H(z)|_{z=e^{j\omega}}$$

where

$$H(z) = \frac{1}{1 + \sum_{k=1}^{p} a_k^{\text{wb}} z^{-k}}.$$

The LP order (p) is 16. Fig 7.3(b) shows the spectrum of the corresponding WB residual component $e_{\text{wb}}[n]$.

The WB residual component ($F_s = 16$kHz) is then umpsampled to a sampling frequency of 16kHz via zero insertion to obtain $\hat{e}_{\text{swb}}[n]$ whereas the WB LP coefficients are kept unchanged. The spectrum of the upsampled residual component is shown in Fig 7.3(c).

During *synthesis*, the extended excitation $\hat{e}_{\text{swb}}[n]$ is combined with the filter $H(\omega)$. The upsampling operation on the residual component $\hat{e}_{\text{wb}}[n]$ to obtain $\hat{e}_{\text{swb}}[n]$ thus corresponds to stretching of the WB LP spectral envelope. This can be observed from the *effective* frequency response $H(\omega)$ for the input signal $\hat{e}_{\text{swb}}[n]$ which is illustrated in Fig 7.3(d). Only the HF components, contained within the green boxes in Fig. 7.3(c) and (d), bear influence on the resulting SWB signal. They are extracted by high-pass filtering.

7.4.3 Comparison

A comparison of the SWB spectral envelopes extracted from extended (using the proposed SWBE algorithm) and true SWB speech is illustrated in Fig. 7.4 for an arbitrary unvoiced speech frame. Blue and dashed-black profiles in Fig. 7.4(a) show the spectral envelopes of true WB ($p = 16$) and SWB ($p = 32$) speech frames respectively. The red profile shows the stretched copy of the WB spectral envelope; it represents the effective frequency response of the filter $H(\omega)$ to the extended residual component $\hat{e}_{\text{swb}}[n]$. The spectral envelopes (dashed black profiles) extracted from true SWB ($p = 32$) and extended SWB ($p = 32$) speech frames are shown in Fig. 7.4(b) and Fig. 7.4(c) respectively. The corresponding speech spectra are also shown (green profiles). The dashed-blue boxes highlight the HF components

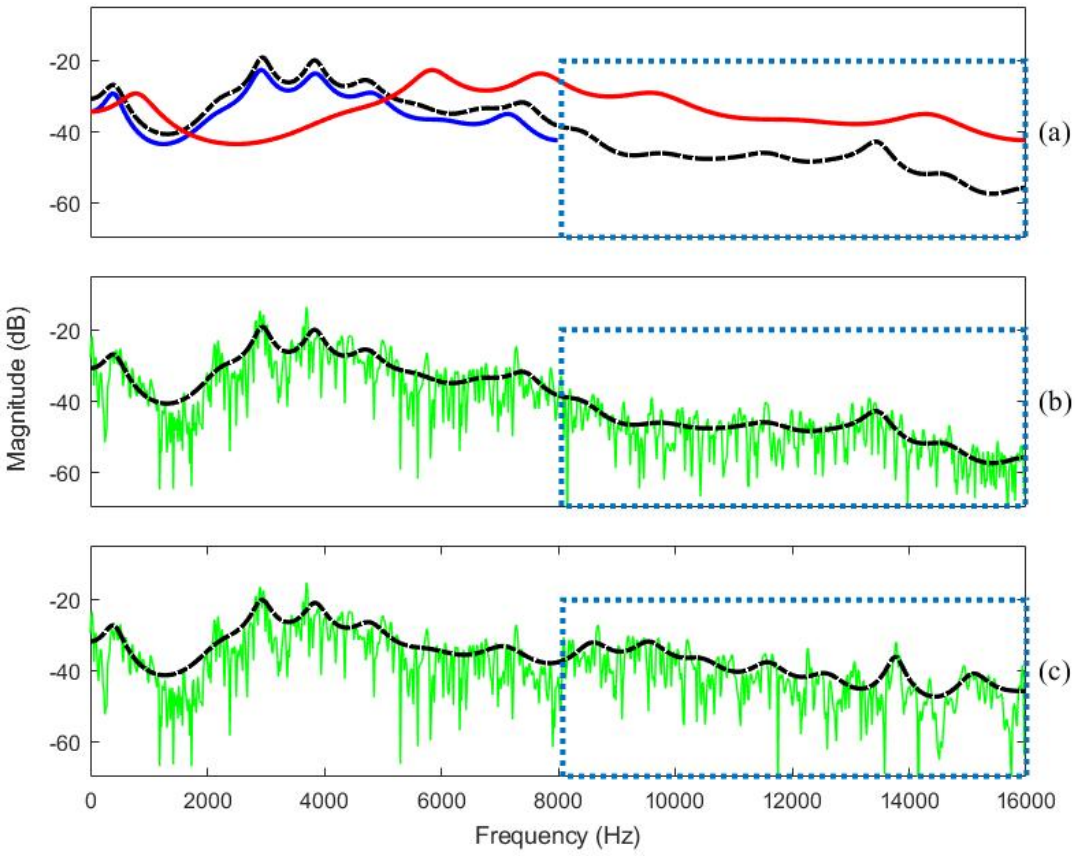


Figure 7.4: A comparison of spectral envelopes for an arbitrary unvoiced speech frame. (a) Spectral envelope profiles are shown for true WB (blue, $p = 16$), true SWB (dashed-black, $p = 32$) speech frames with stretched copy of WB envelope (red). Spectral envelopes (dashed-black profiles, $p = 32$) extracted from (b) true and (c) extended SWB speech frames are shown with spectra of respective speech frames (green profiles).

that are generated within the 8-16kHz frequency range. It is hypothesised that, in this region, profiles of the true SWB (Fig. 7.4(b)) and the extended (Fig. 7.4(c)) speech frames follow spectral shapes which are sufficiently similar to support SWBE. Similar comparison for a voiced speech frame is shown in Fig. 7.5.

7.5 Experimental setup and results

This section reports both objective and subjective assessments of the proposed SWBE algorithm.

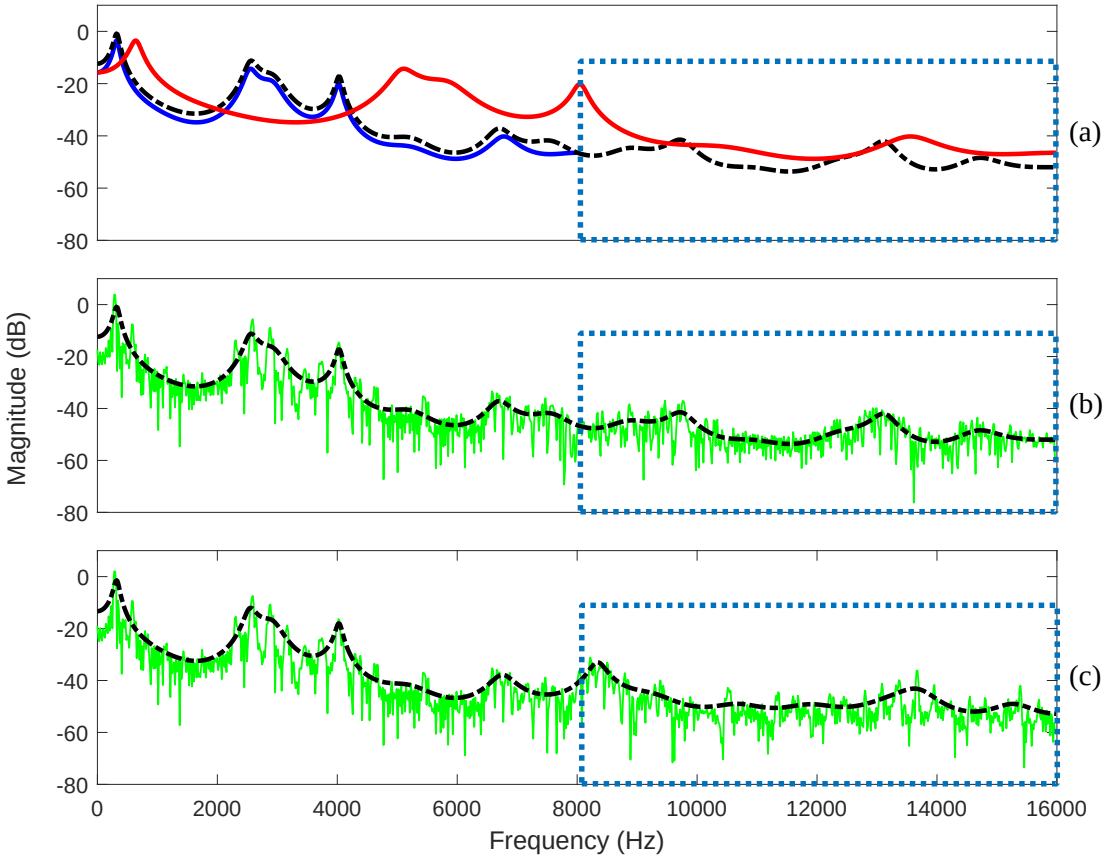


Figure 7.5: A comparison of spectral envelopes similar to that shown in Fig. 7.4 for an arbitrary voiced speech frame.

7.5.1 Databases

All experiments reported here were performed using speech data from the CMU Arctic (Section 3.2.3), TSP speech (Section 3.2.2) and 3GPP (Section 3.2.4) databases. All three databases contain phonetically balanced utterances.

7.5.2 Data pre-processing

Data pre-processing steps are illustrated in Fig. 7.6. All data in the TSP and 3GPP databases were first downsampled to SWB signals so that all three databases then have a common sampling rate of 32kHz. Downsampling was performed using the ResampAudio tool contained in the AFsp package [224]. The active speech level of all utterances in all three databases was then adjusted to -26dBov [227] to

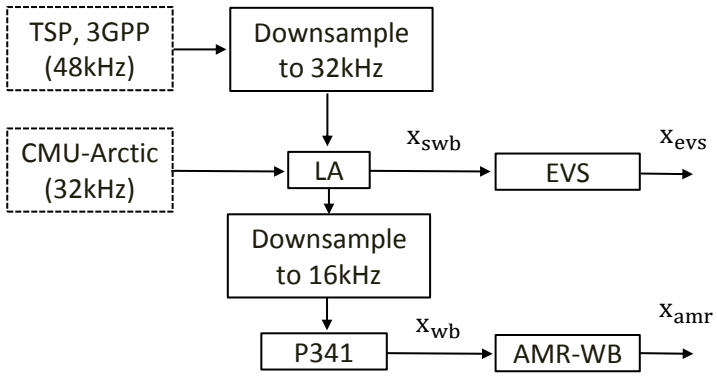


Figure 7.6: Protocol used for data pre-processing. LA = level alignment to -26 dBov.

give SWB data $\mathbf{x}_{\text{swb}}$[2]. They are further encoded and decoded using the EVS [280] codec with active discontinuous transmission in channel aware mode to produce EVS processed signals $\mathbf{x}_{\text{evs}}$. The codec operates at a bitrate of 13.2kbps.

SWB signals $\mathbf{x}_{\text{swb}}$ were then downsampled to 16kHz and passed through a send-side bandpass filter [226] according to ITU-T Rec. P.341 [281], thereby limiting the bandwidth to 50Hz-7kHz, to obtain WB data $\mathbf{x}_{\text{wb}}$. This data was in turn processed with the AMR-WB codec [282] at 12.65kbps in default mode to produce AMR-WB processed signals $\mathbf{x}_{\text{amr}}$. AMR-WB data $\mathbf{x}_{\text{amr}}$ forms the input to the SWBE algorithm ($\mathbf{x}_{\text{wb}}$ in Fig. 7.1 is replaced by $\mathbf{x}_{\text{amr}}$).

7.5.3 Assessment and baseline algorithm

The proposed bandwidth extension algorithm is assessed against AMR-WB and EVS processed speech signals, with the EHBE algorithm [208] being used as a baseline. Since EVS encodes frequencies up to 14kHz when operating at 13.2kbps, bandwidth extended signals produced using either the baseline or the proposed approach are also bandlimited to 14kHz. With a 1024-point FFT, the proposed algorithm was implemented with Hann windows of 25ms duration and 50% overlap, with OLA conditions necessary for perfect reconstruction [216, 217].

[2]Indices $[n]$ (as illustrated in Fig. 7.1) are dropped for convenience.

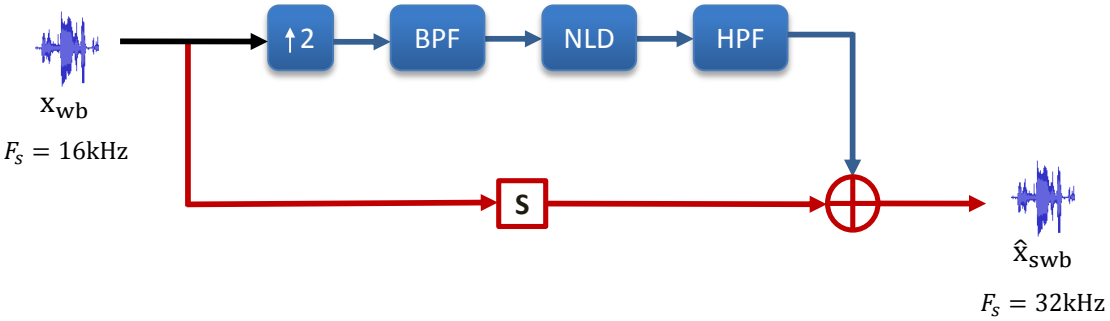


Figure 7.7: An approach to efficient high-frequency bandwidth extension (EHBE) [208]

EHBE baseline algorithm

The baseline algorithm is illustrated in Fig. 7.7. A WB speech signal $\mathbf{x}_{\text{wb}}$ is first upsampled to 32kHz. The highest octave (i.e., 4-8kHz frequency band) present in the input WB signal is extracted using a band-pass filter (BPF). The resulting signal is processed with a non-linear device (NLD), e.g., full-wave rectifier to produce frequency components within 8-16kHz frequency band. The HF components are then extracted using a HPF. The input signal $\mathbf{x}_{\text{wb}}$ (after synchronisation (S) to compensate for delays introduced by filtering operations) is then added to the HF components to obtain extended SWB signal $\hat{\mathbf{x}}_{\text{swb}}$. The EHBE baseline algorithm was implemented in the time domain without framing, as described in [208].

Input WB signals are processed with AMR-WB codec with a bitrate of 12.65kbps because no significant improvement in quality is obtained beyond this bitrate [283]. At 12.65kbps encoding then operates over a frequency range of 0-6.4kHz whereas components up to 8kHz are added during decoding through noise filling [35]. Input signals to both the proposed and baseline algorithms thus extend to 8kHz.

7.5.4 Objective assessment

Objective assessment is performed using the RMS-LSD metric (Section 3.4.2). The average RMS-LSD is determined for estimated HF components only, i.e. in the frequency range 8-14kHz (LF components are not taken into account). It is used to compare EVS-processed and bandwidth-extended speech signals produced using either the proposed algorithm or the EHBE baseline. Comparisons are made with original SWB signals $\mathbf{x}_{\text{swb}}$. All signals were time-aligned before evaluation to

Table 7.1: RMS-LSD results in dB (standard deviation).

	Proposed	EHBE	EVS
CMU Arctic	10.13 (1.68)	11.74 (2.03)	5.00 (0.48)
3GPP	11.06 (1.90)	13.56 (2.30)	4.87 (0.39)
TSP speech	9.29 (0.84)	10.20 (1.04)	4.74 (0.51)
Average	9.92 (1.56)	11.36 (1.96)	4.94 (0.50)

account for any delay introduced by encoding/decoding.

Results presented in Table 7.1 show that the proposed algorithm gives a lower RMS-LSD than the EHBE algorithm. An average RMS-LSD of 9.92dB corresponds to an improvement of 1.44dB over the baseline. As expected, EVS processed signals show lower RMS-LSD values. This is because the EVS codec performs non-blind ABE during decoding exploiting side information related to missing HF information; reconstruction thus results in better RMS-LSD estimates. While results for the proposed algorithm are inferior to those of EVS processed signals, they suggest that it gives a better estimate of the HF spectral shape than the baseline.

7.5.5 Subjective assessment

Subjective assessments were performed using CCR listening tests (Section 3.4.1) in order to compare performance in terms CMOS. Each set of tests involves the pairwise comparison of bandwidth extended signals with (i) AMR-WB signals, (ii) EVS processed signals and (iii) those extended via the EHBE baseline algorithm. Each set of tests was performed by 14 listeners. They were asked to compare the quality of 15 (5 chosen randomly from each of the 3 databases) randomly ordered pairs of speech signals A and B, one of which was treated with the proposed bandwidth extension algorithm. Listeners were asked to rate the quality of signal A with respect to B according to the following scale: -3 (much worse), -2 (slightly worse), -1 (worse), 0 (about the same), 1 (slightly better), 2 (better), 3 (much better). The samples were played using DT 770 PRO headphones.

Subjective assessment results are illustrated in Fig. 7.8. Each group of three bars shows average listener preferences for each of the three comparisons. Blue bars show the percentage of tests in which signals treated with the proposed SWBE algorithm were judged to be of superior quality (scores>0). Green bars show the

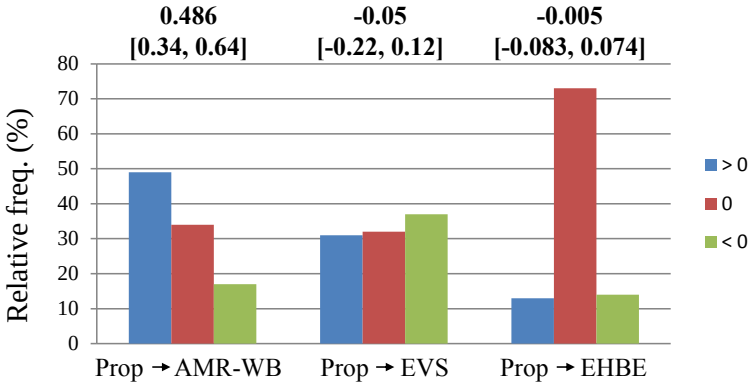


Figure 7.8: Subjective test results in terms of CMOS for bandwidth extended speech generated with the proposed (Prop) algorithm (A) versus either AMR-WB, EVS or EHBE processed speech (B). Each bar indicates the relative frequency that (blue bars) A was preferred to B (score>0), that (green bars) quality was indistinguishable (score=0), or that (red bars) B was preferred to A (score< 0). Scores illustrated to the top are CMOS points with corresponding CI_{95}.

percentage of trials where the same signals were judged to be of inferior quality (scores<0). Red bars show the percentage of tests for which relative quality was indistinguishable (scores=0).

Compared to AMR-WB signals, 49% of speech files treated with the proposed algorithm were judged to be of superior quality. In comparison to EVS processed signals, 32% of trials were found to be of equivalent quality, while 31% were judged to be of superior quality. Quality was found to be inferior for 37% of trials. Up to 73% of comparisons to the EHBE baseline showed no discernible difference. CMOS results (with corresponding 95% confidence interval (CI_{95})) illustrated to the top of Fig. 7.8 also illustrate the improvement in quality compared to AMR-WB signals and equivalence to EVS and EHBE processed signals. The proposed approach outperforms AMR-WB speech quality by significant CMOS points of 0.486. The EVS and EHBE processed speech signals showed slightly higher preference signified by the CMOS points of -0.05 and -0.005 respectively. Overall, these results show that the proposed SWBE algorithm improves consistently on speech quality than AMR-WB signals and to the levels comparable with EVS and EHBE processed speech.

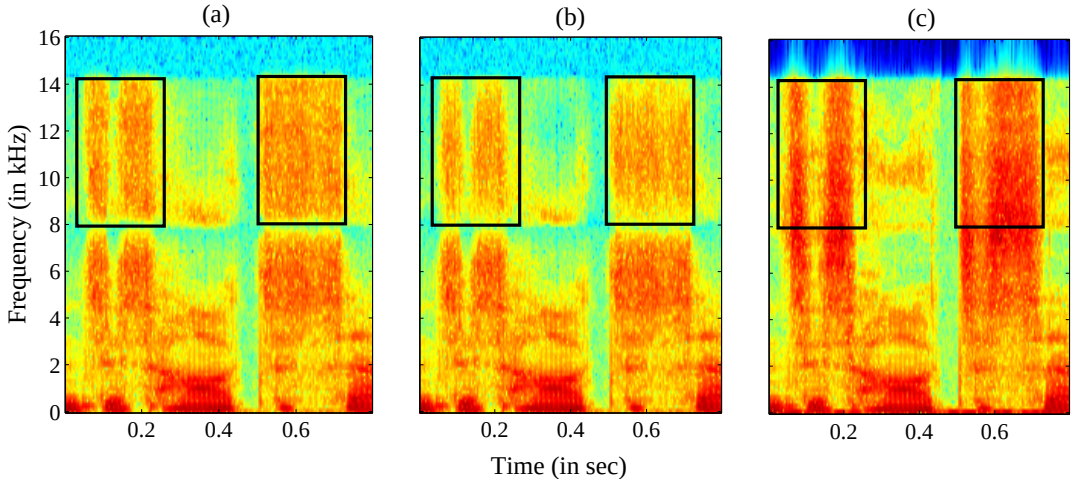


Figure 7.9: Spectrograms of a AMR-WB processed speech segment extended by the proposed algorithm (a) and the EHBE baseline (b) compared to true SWB speech (c). LF components (0-8kHz) in plots (a) and (b) are different than those in plot (c) due to AMR-WB processing.

7.5.6 Discussion

Fig. 7.9 shows a comparison of spectrograms for speech signals after bandwidth extension using (a) the proposed and (b) the baseline algorithms with the true SWB spectrogram illustrated in (c). The spectral gap in both (a) and (b) around 8kHz which arises through AMR-WB processing is generally imperceptible [16]. The comparison of spectrograms in (a) and (b) shows that HF components estimated by the proposed method reflect more reliably the HF components in the true SWB spectrogram (c). This finding confirms the improvements found with objective RMS-LSD assessments. However, subjective assessments show that time domain processing without framing can lead to fewer processing artefacts.

Even though RMS-LSD objective assessment results show that the proposed SWBE algorithm produces speech of lower quality than that produced by the EVS codec, subjective assessment results show only marginal difference. This is because the level discrimination reduces drastically at higher frequencies (especially beyond 8kHz) [284]. As a result resynthesized SWB speech is perceived to be of similar quality.

Lastly, whereas the EHBE algorithm operates on the speech signal directly, the proposed algorithm is based on a classical source filter model. Therefore, when used in combination with a WB codec which employs some form of linear prediction (e.g.

AMR-WB codec), the proposed SWBE algorithm avoids an additional re-synthesis step and therefore introduces lower complexity.

7.6 Summary

This chapter presents an approach to super-wide bandwidth extension that is based on a classical source filter model. The key contributions of this work are as follows:

- An efficient super-wide bandwidth extension algorithm is presented which introduces negligible complexity and is thus well suited to real time implementations. The complexity is reduced by extrapolation of wideband spectral envelope to obtain missing high-frequency spectral envelope information, without requiring statistical estimation.
- Super-wide bandwidth extension is performed for AMR-WB processed speech signals and a comparison of the extended super-wideband speech signals with those processed with the latest EVS-processed is reported. Results of subjective assessments showed that the proposed approach produces speech of notably higher quality than wideband input signals and of comparable quality to speech signals processed with the EVS codec.
- Being codec neutral, the proposed algorithm can be used to improve the speech quality offered by wideband networks and devices and can also be used to preserve quality when super-wideband devices are used alongside wideband services. When used in combination with a wideband codec which operates on some form of linear prediction coefficients, the proposed approach avoids an additional resynthesis step to obtain super-wide bandwidth signals, thereby reducing the complexity.

Chapter 8

Conclusions and future directions

Artificial bandwidth extension (ABE) algorithms estimate missing higher frequency components from available lower frequency components. The research presented in this thesis tackles the problem of using higher dimensional features resulting from *explicit memory inclusion* for efficient ABE. The focus was to improve ABE performance under the constraints of fixed dimensionality, especially when using traditional regression techniques such as Gaussian mixture model regression (GMMR). The work presents novel, efficient solutions to both narrow-to-wideband and wide-to-super-wideband extension. Solutions are based on the classical source-filter model of speech.

8.1 Contributions and conclusions

The contributions and conclusions derived from the research presented in this thesis are explained as follows:

- In Chapter 4, the benefit of using explicit memory that can be captured through static features extracted from neighbouring speech frames is studied via information theoretic analysis. The performance of the *memoryless* baseline ABE system (presented in Chapter 3) is shown to be improved via explicit memory inclusion under the constraint of fixed dimensionality. Improvements in speech quality are obtained with modest increases in complexity and an algorithmic delay (latency) of 20ms. Comparison category rating (CCR) subjective listening tests show that the quality of narrowband speech signals is improved by 0.69 CMOS points. The baseline system employs a GMMR technique to

estimate 10-dimensional highband (HB) linear prediction (LP) coefficients from 10-dimensional narrowband (NB) log-Mel filter energy (logMFE) coefficients. The complexity of the GMMR model resulting from the use of memory is tackled by employing principal component analysis (PCA) as a dimensionality reduction technique. The resulting 10-dimensional, compact NB representation is shown to improve mutual information (MI) by 8.1% relative compared to NB logMFE features. The analysis also shows that memory included from more than two neighbouring speech frames (under the constraint of fixed dimensionality) do not further improve ABE performance. In this case, latency is compatible with the constraints of real-time ABE.

- With a focus only upon the retention of variation in the input NB features, being an unsupervised approach to dimensionality reduction, PCA is not optimised for the estimation of HB information, and is likely a suboptimal solution. The use of semi-supervised stacked auto-encoders (SSAEs) as a better alternative to dimensionality reduction is proposed in Chapter 5. SSAEs are explored to extract low dimensional representations tailored specifically to ABE, with the expectation of better estimation performance. The 10-dimensional higher-level NB features learned via SSAEs when trained on higher-dimensional input NB log power spectrum (LPS) coefficients (with memory inclusion) show significant improvements to speech quality than those obtained from input NB logMFE coefficients. The newly learned NB features learned from LPS and logMFE inputs (with memory) lead to the relative MI improvements of $\approx 23\%$ and 9% over 10-dimensional NB logMFE features respectively. Improvements to speech quality are confirmed by informal listening tests.

- Chapter 6 describes the first application of conditional variational auto-encoders (CVAEs) to supervised dimensionality reduction specifically tailored to ABE. CVAEs, a form of directed, graphical model, are utilised to model the distribution of higher-dimensional log-spectral data (that includes memory) to extract higher-level, compact, latent NB features that are tailored to improve the estimation of missing HB components. This work introduces an approach to the joint optimisation of a CVAE with a probabilistic encoder in the form of an auxiliary neural network which derives its conditioning variable. Speech signals extended using the proposed CVAE feature extraction scheme show improved quality compared to NB signals by 0.90 CMOS points. They are also found to be of superior quality than those extended using PCA and SSAE ABE systems by CMOS points of 0.13 and 0.10 respectively. In contrast to bottleneck features

learned by SSAEs, latent (or probabilistic) features learned by CVAEs also show consistent performance with or without mean-variance normalisation.

- Following standardisation in 2014, the deployment of the enhanced voice services (EVS) codec by telecom operators is increasing rapidly. However, users of super-wideband (SWB) devices will still be restricted to narrowband or wideband communications until the operational bandwidth of devices and networks moves completely to SWB. Chapter 7 proposes a *super*-wide bandwidth extension (SWBE) algorithm which is based on the classical source filter model. The approach is based on linear prediction based analysis-synthesis whereby missing higher frequency components are generated from the combination of the WB residual error signal after zero insertion in the time domain with the original WB spectral envelope. SWB speech produced by the proposed approach shows superior quality to WB speech signal processed with the adaptive multi-rate WB (AMR-WB) codec (at 12.65 kbps). CMOS results show significant improvements by 0.486 points. Results also show that bandwidth-extended signals are comparable in terms of speech quality to signals processed with the EVS codec (at 13.2kbps) with only slightly higher preference observed for the latter. While the proposed approach shows slightly inferior results in subjective listening tests, this is perhaps attributed to the specific implementation. Time domain processing without framing, as used for the baseline system, leads to fewer processing artefacts. The proposed approach, however, avoids an extra resynthesis step when used with a wideband codec that employs some form of linear prediction and thus provides a efficient solution to SWBE with no need for statistical estimation.

8.2 Future directions

The research work carried out in this thesis highlights a number of research directions for future exploration.

- The thesis investigates the importance to ABE of explicit memory. Furthermore, memory captured from NB static features extracted from neighbouring speech frames is exploited without implications for computational complexity. Memory is included via the use of dimensionality reduction and a traditional regression technique using Gaussian mixture models. It should be interesting to investigate and study the relative improvements in ABE performance due

to explicit memory inclusion using hidden Markov models (HMMs) which inherently capture interframe dependancies. The relative improvements in performance of DNN-based ABE algorithms should also be of interest.

- An important aspect of the work presented in this thesis is that deep learning based dimensionality reduction techniques without supervised training perform poorly in comparison to PCA. This motivates the consideration of other dimensionality reduction approaches for ABE.

- Previous works on the assessment of speech quality have illustrated that no ABE algorithm shows consistent results for different languages and codecs. This is perhaps due to the mismatch in training and testing conditions involved during assessment. The training of DNN based models using data that captures such variation may provide better, more robust solutions.

- SWBE approaches which operate without statistical estimation perform well and also reduce computational complexity. Performance can be further improved via statistical estimation of only gain or energy parameters. This is because the estimation of the energy or gain of the HB spectral envelope is especially important to ABE performance. The particular, focused consideration of energy or gain parameter offers potential for better performance.

Bibliography

[1] T. Watson, "How Bell invented the telephone," *American Institute of Electrical Engineers*, vol. 34, no. 8, pp. 1503–1513, 1915. [Cited on page 2].

[2] J. Flood, "Alexander Graham Bell and the invention of the telephone," *Electronics & Power*, vol. 22, no. 3, pp. 159–162, 1976. [Cited on page 2].

[3] N. R. Council, *The evolution of untethered communications*. National Academies Press, 1998. [Cited on pages 2, 3].

[4] A. Inglis, "Transmission features of the new telephone sets," *Bell System Technical Journal*, vol. 17, no. 3, pp. 358–380, 1938. [Cited on page 2].

[5] W. Goodall, "Telephony by pulse code modulation," *The Bell System Technical Journal*, vol. 26, no. 3, pp. 395–409, 1947. [Cited on pages 2, 3, 4].

[6] K. Cattermole, "History of pulse code modulation," in *Proc. Institution of Electrical Engineers*, pp. 889–892, IET, 1979. [Cited on page 2].

[7] "ITU-T Recommendation P.341: Transmission characteristics of national networks," 1998. [Cited on pages 3, 9].

[8] A. Dodd, *The essential guide to telecommunications*. 5th edition, Prentice Hall Professional, 2002. [Cited on pages 4, 5, 11].

[9] Source: `https://www.rfpage.com/evolution-of-wireless-technologies-1g-to-5g-in-mobile-communication/`. [Cited on pages xv, 4].

[10] T. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*. Prentice Hall, 2002. [Cited on pages xv, 5, 6, 7, 78].

[11] D. O'shaughnessy, *Speech communication: Human and Machine*. Universities press, USA, 1987. [Cited on pages 5, 9, 80].

[12] K. Stevens, "Acoustic correlates of some phonetic categories," *The Journal of the Acoustical Society of America*, vol. 68, no. 3, pp. 836–842, 1980. [Cited on page 8].

[13] L. Laaksonen, "Artificial bandwidth extension of narrowband speech-enhanced speech quality and intelligibility in mobile devices," Ph.D. Thesis, Aalto University, Finland, 2013. [Cited on pages 9, 11, 15, 47].

[14] P. Vary and R. Martin, *Digital speech transmission: Enhancement, coding and error concealment*. John Wiley & Sons, 2006. [Cited on pages 9, 11, 12, 13, 14, 16].

[15] A. Nour-Eldin, "Quantifying and exploiting speech memory for the improvement of narrowband speech bandwidth extension," Ph.D. Thesis, McGill University, Canada, 2013. [Cited on pages 10, 32, 34, 38, 56, 58, 72, 73, 74, 75, 77, 81, 94].

[16] P. Jax and P. Vary, "On artificial bandwidth extension of telephone speech," *Signal Processing*, vol. 83, no. 8, pp. 1707–1719, 2003. [Cited on pages 10, 33, 35, 38, 62, 78, 82, 133, 145].

[17] S. Voran, "Listener ratings of speech passbands," in *Proc. IEEE Workshop on Speech Coding For Telecommunications*, pp. 81–82, 1997. [Cited on page 10].

[18] A. Spanias, "Speech coding: A tutorial review," *Proc. of the IEEE*, vol. 82, no. 10, pp. 1541–1582, 1994. [Cited on page 11].

[19] H. Pulakka, "Development and evaluation of artificial bandwidth extension methods for narrowband telephone speech," Ph.D. Thesis, Aalto University, Finland, 2013. [Cited on pages 11, 45].

[20] B. Oliver, J. Pierce, and C. Shannon, "The philosophy of PCM," *Proc. of the IRE*, vol. 36, no. 11, pp. 1324–1331, 1948. [Cited on page 12].

[21] "ITU-T Recommendation G.711: Pulse code modulation (PCM) of voice frequencies," 2001. [Cited on page 12].

[22] "ITU-T Recommendation G.712: Transmission performance characteristics of pulse code modulation channels," 1988. [Cited on page 12].

[23] "ITU-T Recommendation G.726: 40, 32, 24, 16 kbit/s Adaptive Differential Pulse Code Modulation (ADPCM)," 1990. [Cited on page 12].

[24] W. Chu, "Speech coding algorithms," *Foundation and evolution of standardized coders*, 2003. [Cited on pages 12, 14].

[25] "ETSI Recommendation GSM 06.10 : Gsm full rate speech transcoding," 1992. [Cited on page 12].

[26] R. Salami, C. Laflamme, B. Bessette, and J.-P. Adoul, "ITU-T G. 729 Annex A: reduced complexity 8 kb/s CS-ACELP codec for digital simultaneous voice and data," *IEEE Communications Magazine*, vol. 35, no. 9, pp. 56–63, 1997. [Cited on page 13].

[27] R. Salami, C. Laflamme, B. Bessette, and J. Adoul, "Description of ITU-T rec. G. 729 Annex A: Reduced complexity 8 kbit / s CS-ACELP coding," in *Proc. IEEE Int.l Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 1997. [Cited on page 13].

[28] "ETSI Recommendation GSM 06.60 : Digital Cellular Telecommunications System (Phase 2+); Enhanced Full Rate (EFR) Speech Transcoding," 1996. [Cited on page 13].

[29] "ITU-T Recommendation G.729: Coding of speech at 8 kbit/s using Conjugate Structure Algebraic Code-Excited Linear-Prediction (CS-ACELP)," 1996. [Cited on page 13].

[30] "ETSI Recommendation GSM 06.90 : Digital Cellular Telecommunications System (Phase 2+); Adaptive Multi-Rate (AMR) Speech Transcoding," 1998. [Cited on page 13].

[31] "3GPP ts 26.090: Mandatory Speech Codec speech processing functions; Adaptive Multi-Rate (AMR) Speech Codec; Transcoding Functions," 2000. [Online]: `https://www.etsi.org/deliver/etsi_ts/126000_126099/126090/14.00.00_60/`, (ver. 14.0.0 Rel. 14, 2017). [Cited on pages 13, 19, 132].

[32] Global mobile Suppliers Association (GSA), "Mobile HD voice: Global Update report," May, 2016. [Online]: `https://gsacom.com/paper/mobile-hd-voice-global-update-report-2/` (Last accessed : April 2019). [Cited on pages 13, 14, 15].

[33] "ITU-T Recommendation G.722: 7 khz audio-coding within 64 kbit/s," 1988. [Cited on page 14].

[34] "ITU-T Recommendation G.722.1: Low-complexity coding at 24 and 32 kbit/s for hands-free operation in systems with low frame loss," 2005. [Cited on page 14].

[35] "3GPP TS 26.190: Speech Codec Speech Processing Functions; Adaptive Multi-Rate - Wideband (AMR-WB) speech codec; Transcoding functions," 2002. [Online]: `https://www.etsi.org/deliver/etsi_ts/126100_126199/126190/13.00.00_60/ts_126190v130000p.pdf`, ver. 13.0.0 Rel. 13, 2016. [Cited on pages 14, 132, 142].

[36] "ITU-T Recommendation G.722.2: Wideband Coding of Speech at Around 16 kbits/s using Adaptive Multi-Rate Wideband (AMR-WB)," 2002. [Cited on page 14].

[37] P. Ojala, A. Lakaniemi, H. Lepanaho, and M. Jokimies, "The adaptive multirate wideband speech codec: system characteristics, quality advances, and deployment strategies," *IEEE Communications Magazine*, vol. 44, no. 5, pp. 59–65, 2006. [Cited on page 14].

[38] "ITU-T Recommendation G.729.1: G.729 Based embedded variable bit-rate coder: An 8–32 kb/s scalable wideband coder bitstream interoperable with g.729," 2006. [Cited on page 15].

[39] D. Sinder, I. Varga, V. Krishnan, V. Rajendran, and S. Villette, "Recent speech coding technologies and standards," in *Speech and Audio Processing for Coding, Enhancement and Recognition*, pp. 75–109, Springer, 2015. [Cited on pages 15, 17, 18].

[40] "ITU-T Recommendation G.729.1: Annex E: Superwideband scalable extension for G.729.1," 2010. [Cited on page 16].

[41] Y. Lee and S. Choi, "Superwideband bandwidth extension using normalized MDCT coefficients for scalable speech and audio coding," *Advances in Multimedia*, vol. 2013, p. 1, 2013. [Cited on page 16].

[42] "3GPP ts 26.290: Audio Codec Processing Functions; Extended Adaptive Multi-Rate - Wideband AMR-WB+ Codec; Transcoding functions," 2005. [Online]: `https://www.etsi.org/deliver/etsi_ts/126200_126299/126290/15.00.00_60/ts_126290v150000p.pdf`, ver. 15.0.0 Rel. 15, 2018. [Cited on page 16].

[43] J. Makinen, B. Bessette, S. Bruhn, P. Ojala, R. Salami, and A. Taleb, "AMR-WB+: a new audio coding standard for 3rd generation mobile audio services," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, pp. ii–1109, 2005. [Cited on page 16].

[44] "ITU-T Recommendation G.719: Low-complexity full-band audio coding for high-quality conversational applications," 2008. [Cited on page 16].

[45] P. Ekstrand, "Bandwidth extension of audio signals by spectral band replication," in *Proc. IEEE Benelux Workshop on Model Based Processing and Coding of Audio (MPCA)*, pp. 53–58, 2002. [Cited on pages 16, 19].

[46] M. Dietz, L. Liljeryd, K. Kjorling, and O. Kunz, "Spectral Band Replication, a novel approach in audio coding," in *Audio Engineering Society Convention*, Audio Engineering Society, 2002. [Cited on pages 16, 19].

[47] J. Herre and M. Dietz, "MPEG-4 high-efficiency AAC coding [standards in a nutshell]," *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 137–142, 2008. [Cited on page 16].

[48] "Full HD Voice," *Huawei white paper*, October, 2014. [Online]: `https://www.huawei.com/ilink/en/download/HW_377700` (Last accessed : April 2019). [Cited on page 17].

[49] "3GPP TR 22.813, Study of use cases and requirements for enhanced voice codecs for the Evolved Packet System (EPS)," 2010. [Cited on page 17].

[50] A. Rämö and H. Toukomaa, "Subjective quality evaluation of the 3GPP EVS codec," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5157–5161, 2015. [Cited on page 18].

[51] "3GPP TS 26.445: Codec for Enhanced Voice Services; Detailed algorithmic description," 2014. [Online]: `https://www.etsi.org/deliver/etsi_ts/126400_126499/126445/13.04.01_60/ts_126445v130401p.pdf, ver. 13.4.0 Rel. 13, 2016` (Last accessed : April 2019). [Cited on pages 18, 132].

[52] "3GPP TS 26.441: Codec for Enhanced Voice Services; General overview," 2014. [Online]: `https://www.etsi.org/deliver/etsi_ts/126400_126499/126441/13.00.00_60/ts_126441v130000p.pdf, ver. 13.0.0 Rel. 13, 2016` (Last accessed : April 2019). [Cited on pages 18, 132].

[53] S. Bruhn, H. Pobloth, M. Schnell, B. Grill, J. Gibbs, L. Miao, K. Järvinen, L. Laaksonen, N. Harada, N. Naka, *et al.*, "Standardization of the new 3gpp evs codec," in *Proc. IEEE Int.l Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5703–5707, 2015. [Cited on page 18].

[54] M. Dietz, M. Multrus, V. Eksler, V. Malenovsky, E. Norvell, H. Pobloth, L. Miao, Z. Wang, L. Laaksonen, A. Vasilache, *et al.*, "Overview of the EVS codec architecture," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5698–5702, 2015. [Cited on page 18].

[55] Global mobile Suppliers Association (GSA), "Enhanced Voice Services (EVS): Market Update," September, 2018. [Online]: `https://gsacom.com/paper/evs-enhanced-voice-services-market-update-september-2018/` (Last accessed : April 2019). [Cited on page 18].

[56] X. Liu and C. Bao, "Blind bandwidth extension of audio signals based on non-linear prediction and hidden Markov model," *APSIPA Transactions on Signal and Information Processing*, vol. 3, p. e8, 2014. [Cited on pages 18, 49, 50, 133].

[57] S. Villette, S. Li, P. Ramadas, and D. J. Sinder, "eAMR: Wideband speech over legacy narrowband networks," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5110–5114, 2017. [Cited on page 19].

[58] C. Beaugeant, M. Schonle, and I. Varga, "Challenges of 16 khz in acoustic pre-and post-processing for terminals," *IEEE Communications Magazine*, vol. 44, no. 5, pp. 98–104, 2006. [Cited on page 20].

[59] S. Villette, S. Li, P. Ramadas, and D. J. Sinder, "An objective evaluation methodology for blind bandwidth extension.," in *Proc. INTERSPEECH*, pp. 2548–2552, 2016. [Cited on page 20].

[60] L. Laaksonen, H. Pulakka, V. Myllyla, and P. Alku, "Development, evaluation and implementation of an artificial bandwidth extension method of telephone speech in mobile terminal," *IEEE Transactions on Consumer Electronics*, vol. 55, no. 2, pp. 780–787, 2009. [Cited on pages xv, 20, 21].

[61] B. Geiser, "High-definition telephony over heterogeneous networks," Ph.D. Thesis. Rheinisch-Westfälischen Technische Hochschule Aachen, Germany, 2012. [Cited on page 20].

[62] S. Moller, M. Waltermann, B. Lewcio, N. Kirschnick, and P. Vidales, "Speech quality while roaming in next generation networks," in *Proc. IEEE Int.l Conf. on Communications*, pp. 1–5, 2009. [Cited on pages 20, 22].

[63] S. Voran, "Subjective ratings of instantaneous and gradual transitions from narrowband to wideband active speech," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4674–4677, 2010. [Cited on pages 20, 22].

[64] "3GPP TS 26.976: Performance characterization of the Adaptive Multi-Rate Wideband (AMR-WB) speech codec," 2003. [Online]: `https://www.etsi.org/deliver/etsi_tr/126900_126999/126976/15.00.00_60/tr_126976v150000p.pdf`, ver. 15.0.0 Rel. 15, 2018. [Cited on page 22].

[65] P. Nidadavolu, V. Iglesias, J. Villalba, and N. Dehak, "Investigation on neural bandwidth extension of telephone speech for improved speaker recognition," in *Proc. IEEE Int.l Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6111–6115, 2019. [Cited on page 22].

[66] K. Li, Z. Huang, Y. Xu, and C.-H. Lee, "DNN-based speech bandwidth expansion and its application to adding high-frequency missing features for automatic speech recognition of narrowband speech," in *Annual Conf. of the Int. Speech Comm. Association*, 2015. [Cited on page 22].

[67] D. Haws and X. Cui, "Cyclegan bandwidth extension acoustic modeling for automatic speech recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech*

and Signal Processing (ICASSP), pp. 6780–6784, 2019. [Cited on pages 22, 41].

[68] R. Masumura, T. Tanaka, T. Moriya, Y. Shinohara, T. Oba, and Y. Aono, "Large context end-to-end automatic speech recognition via extension of hierarchical recurrent encoder-decoder models," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5661–5665, 2019. [Cited on page 22].

[69] M. Faúndez-Zanuy, M. Nilsson, and W. B. Kleijn, "On the relevance of bandwidth extension for speaker identification," in *Proc. IEEE European Signal Processing Conference (EUSIPCO)*, pp. 1–4, 2002. [Cited on page 22].

[70] R. Kaminishi, H. Miyamoto, S. Shiota, and H. Kiya, "Investigation on blind bandwidth extension with a non-linear function and its evaluation of x-vector-based speaker verification," pp. 4055–4059, 2019. [Cited on page 22].

[71] C. Liu, Q.-J. Fu, and S. S. Narayanan, "Effect of bandwidth extension to telephone speech recognition in cochlear implant users," *The Journal of the Acoustical Society of America*, vol. 125, no. 2, pp. EL77–EL83, 2009. [Cited on page 23].

[72] W. Nogueira, J. Abel, and T. Fingscheidt, "Artificial speech bandwidth extension improves telephone speech intelligibility and quality in cochlear implant users," *The Journal of the Acoustical Society of America*, vol. 145, no. 3, pp. 1640–1649, 2019. [Cited on page 23].

[73] M. Croll, *Sound-quality improvement of broadcast telephone calls*. Research Department, Engineering Division, BBC, 1972. [Cited on pages 31, 38].

[74] P. Patrick, "Enhancement of band-limited speech signals," Ph.D. Thesis, Loughborough University of Techology, UK, 1983. [Cited on page 32].

[75] H. Yasukawa, "Signal restoration of broad band speech using nonlinear processing," in *Proc. IEEE European Signal Processing Conference*, pp. 1–4, 1996. [Cited on page 32].

[76] M. Dietrich, "Performance and implementation of a robust ADPCM algorithm for wideband speech coding with 64 kbit/s," in *Proc. Int. Zürich Seminar on Digital Communications*, 1984. [Cited on page 32].

[77] U. Lindgren and H. Gustafsson, "Speech bandwidth extension," 2002, US Patent no. 2002/0128839 A1. [Cited on page 32].

[78] P. Jax, "Enhancement of bandlimited speech signals: Algorithms and theoretical bounds," Ph.D. Thesis, Aachen University (RWTH), Germany, 2002. [Cited on pages xv, 32, 33, 35, 37, 38, 42, 53, 54, 55, 69, 82].

[79] B. Iser and G. Schmidt, "Bandwidth extension of telephony speech," in *Speech and Audio Processing in Adverse Environments*, pp. 135–184, Springer, 2008. [Cited on pages 32, 38].

[80] J. Flanagan, *Speech analysis synthesis and perception.* 2nd edition, Springer, 1972. [Cited on page 32].

[81] K.-Y. Park and H. Kim, "Narrowband to wideband conversion of speech using GMM based transformation," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 1843–1846, 2000. [Cited on pages 33, 35, 78].

[82] Y. Cheng, D. O'Shaughnessy, and P. Mermelstein, "Statistical recovery of wideband speech from narrowband speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 544–548, 1994. [Cited on page 33].

[83] Y. Qian and P. Kabal, "Dual-mode wideband speech recovery from narrowband speech.," in *Proc. INTERSPEECH*, 2003. [Cited on pages 33, 35, 78].

[84] G. Chen and V. Parsa, "HMM-based frequency bandwidth extension for speech enhancement using line spectral frequencies," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, pp. I–709, 2004. [Cited on page 33].

[85] A. Nour-Eldin and P. Kabal, "Mel-frequency cepstral coefficient-based bandwidth extension of narrowband speech„" in *Proc. INTERSPEECH*, pp. 53–56, 2008. [Cited on pages 33, 35, 44, 72, 76, 78, 86].

[86] J. Abel, M. Strake, and T. Fingscheidt, "Artificial bandwidth extension using deep neural networks for spectral envelope estimation," in *Proc. of Int. Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 1–5, 2016. [Cited on pages 33, 36, 44, 103].

[87] P. Bauer, "Artificial bandwidth extension of telephone speech signals using phonetic: A priori knowledge," Ph.D. Thesis, Technische Universität CaroloWilhelmina zu Braunschweig, Germany, 2017. [Cited on pages 33, 36].

[88] H. Carl and U. Heute, "Bandwidth enhancement of narrow-band speech signals," in *Proc. IEEE European signal processing conference (EUSIPCO)*, pp. 1178–1181, 1994. [Cited on pages 33, 38].

[89] V. Iyengar, R. Rabipour, P. Mermelstein, and B. Shelton, "Speech bandwidth extension method and apparatus," 1995, US Patent no. 5,455,888. [Cited on page 33].

[90] Y. Yoshida and M. Abe, "An algorithm to reconstruct wideband speech from narrowband speech based on codebook mapping," in *Proc. Int. Conf. on Spoken Language Processing*, 1994. [Cited on page 33].

[91] H. YASUKAWA, "Spectrum broadening of telephone band signals using multirate processing for speech quality enhancement," *IEICE Transations on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 78, no. 8, pp. 996–998, 1995. [Cited on page 33].

[92] J. Epps, "Wideband extension of narrowband speech for enhancement and coding," Ph.D. Thesis, University of New South Wales, Australia, 2000. [Cited on page 33].

[93] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Transactions on communications*, vol. 28, no. 1, pp. 84–95, 1980. [Cited on page 33].

[94] J. Epps and W. H. Holmes, "A new technique for wideband enhancement of coded narrowband speech," in *IEEE Workshop on Speech Coding Proceedings. Model, Coders, and Error Criteria (Cat. No. 99EX351)*, pp. 174–176, 1999. [Cited on pages 33, 34].

[95] E. Larsen and R. Aarts, *Audio bandwidth extension: application of psychoacoustics, signal processing and loudspeaker design.* John Wiley & Sons, 2005. [Cited on pages 33, 50].

[96] G. Miet, A. Gerrits, and J.-C. Valiere, "Low-band extension of telephone-band speech," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pp. 1851–1854, 2000. [Cited on page 33].

[97] C. Avendano, H. Hermansky, and E. A. Wan, "Beyond Nyquist: Towards the recovery of broad-bandwidth speech from narrow-bandwidth speech," in *Proc. Fourth European Conference on Speech Communication and Technology*, 1995. [Cited on page 33].

[98] Y. Nakatoh, M. Tsushima, and T. Norimatsu, "Generation of broadband speech from narrowband speech using piecewise linear mapping.," in *EUROSPEECH*, 1997. [Cited on pages 34, 36].

[99] D. Reynolds and R. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Transactions on speech and audio processing*, vol. 3, no. 1, pp. 72–83, 1995. [Cited on page 34].

[100] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 285–288, 1998. [Cited on page 35].

[101] Y. Stylianou, O. Cappe, and E. Moulines, "Statistical methods for voice quality transformation," in *Fourth European Conference on Speech Communication and Technology*, 1995. [Cited on page 35].

[102] Y. Qian and P. Kabal, "Combining equalization and estimation for bandwidth extension of narrowband speech," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, pp. I–713, 2004. [Cited on pages 35, 38].

[103] A. Nour-Eldin, T. Shabestary, and P. Kabal, "The effect of memory inclusion on mutual information between speech frequency bands," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 3, pp. III–III, 2006. [Cited on pages 35, 44, 72, 78].

[104] A. Nour-Eldin and P. Kabal, "Objective analysis of the effect of memory inclusion on bandwidth extension of narrowband speech," in *Proc. INTERSPEECH*, pp. 2489–2492, 2007. [Cited on pages 35, 44, 72].

[105] A. H. Nour-Eldin and P. Kabal, "Combining frontend-based memory with MFCC features for bandwidth extension of narrowband speech," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4001–4004, 2009. [Cited on page 35].

[106] H. Pulakka, U. Remes, K. Palomäki, M. Kurimo, and P. Alku, "Speech bandwidth extension using Gaussian mixture model-based estimation of the highband mel spectrum," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5100–5103, 2011. [Cited on page 35].

[107] P. Jax and P. Vary, "Wideband extension of telephone speech using a hidden Markov model," in *IEEE Workshop on Speech Coding Proceedings. Meeting the Challenges of the New Millennium*, pp. 133–135, 2000. [Cited on page 35].

[108] P. Jax and P. Vary, "Artificial bandwidth extension of speech signals using MMSE estimation based on a hidden Markov model," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. I–I, 2003. [Cited on page 35].

[109] P. Bauer and T. Fingscheidt, "A statistical framework for artificial bandwidth extension exploiting speech waveform and phonetic transcription," in *IEEE European Signal Processing Conference (EUSIPCO)*, pp. 1839–1843, 2009. [Cited on pages 35, 36, 82].

[110] P. Bauer, M.-A. Jung, J. Qi, and T. Fingscheidt, "On improving speech intelligibility in automotive hands-free systems," in *Proc. IEEE Int. Symposium on Consumer Electronics (ISCE 2010)*, pp. 1–5, 2010. [Cited on page 36].

[111] P. Bauer, R.-L. Fischer, M. Bellanova, H. Puder, and T. Fingscheidt, "On improving telephone speech intelligibility for hearing impaired persons," in *Speech Communication; 10. ITG Symposium*, pp. 1–4, VDE, 2012. [Cited on page 36].

[112] P. Bauer, J. Jones, and T. Fingscheidt, "Impact of hearing impairment on fricative intelligibility for artificially bandwidth-extended telephone speech in noise," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 7039–7043, IEEE, 2013. [Cited on page 36].

[113] P. Bauer, J. Abel, and T. Fingscheidt, "HMM-based artificial bandwidth extension supported by neural networks," in *Proc. IEEE Int. Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 1–5, 2014. [Cited on pages 36, 82].

[114] I. Katsir, D. Malah, and I. Cohen, "Evaluation of a speech bandwidth extension algorithm based on vocal tract shape estimation," in *Proc. IEEE*

Int. Workshop on Acoustic Signal Enhancement (IWAENC), pp. 1–4, VDE, 2012. [Cited on page 36].

[115] M. Hosoki, T. Nagai, and A. Kurematsu, "Speech signal band width extension and noise removal using subband HMN," in *Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. I–245, IEEE, 2002. [Cited on page 36].

[116] M. Sanna and M. Murroni, "A codebook design method for fricative enhancement in artificial bandwidth extension," in *Proc. Int. Mobile Multimedia Communications Conf.*, p. 39, ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications, 2009. [Cited on page 36].

[117] K. Kalgaonkar and M. Clements, "Vocal tract area based artificial bandwidth extension," in *IEEE Workshop on Machine Learning for Signal Processing*, pp. 480–485, 2008. [Cited on pages 36, 43, 82].

[118] K. Kalgaonkar and M. Clements, "Sparse probabilistic state mapping and its application to speech bandwidth expansion," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4005–4008, 2009. [Cited on page 36].

[119] C. YağLı, M. Turan, and E. Erzin, "Artificial bandwidth extension of spectral envelope along a viterbi path," *Speech Communication*, vol. 55, no. 1, pp. 111–118, 2013. [Cited on page 36].

[120] C. Yağli and E. Erzin, "Artificial bandwidth extension of spectral envelope with temporal clustering," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 5096–5099, 2011. [Cited on page 36].

[121] J. Han, G. Mysore, and B. Pardo, "Language informed bandwidth expansion," in *Proc. IEEE Int. Workshop on Machine Learning for Signal Processing*, pp. 1–6, 2012. [Cited on page 36].

[122] B. Iser and G. Schmidt, "Neural networks versus codebooks in an application for bandwidth extension of speech signals," in *Eighth European Conference on Speech Communication and Technology*, 2003. [Cited on page 36].

[123] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, B. Kingsbury, *et al.*, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal processing magazine*, vol. 29, 2012. [Cited on page 36].

[124] J. Abel and T. Fingscheidt, "Artificial speech bandwidth extension using deep neural networks for wideband spectral envelope estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 71–83, 2018. [Cited on pages 36, 103, 104].

[125] Y. Li and S. Kang, "Artificial bandwidth extension using deep neural network-based spectral envelope estimation and enhanced excitation estimation," *IET Signal Processing*, vol. 10, no. 4, pp. 422–427, 2016. [Cited on pages 37, 38].

[126] Y. Tachioka and J. Ishii, "Long short-term memory recurrent-neural-network-based bandwidth extension for automatic speech recognition," *Acoustical Science and Technology*, vol. 37, no. 6, pp. 319–321, 2016. [Cited on page 37].

[127] Y. Wang, S. Zhao, J. Li, J. Kuang, and Q. Zhu, "Recurrent neural network for spectral mapping in speech bandwidth extension," in *Proc. IEEE Global Conf. on Signal and Information Processing (GlobalSIP)*, pp. 242–246, 2016. [Cited on page 37].

[128] K. Schmidt and B. Edler, "Blind bandwidth extension based on convolutional and recurrent deep neural networks," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5444–5448, 2018. [Cited on page 37].

[129] Y. Wang, S. Zhao, D. Qu, and J. Kuang, "Using conditional restricted boltzmann machines for spectral envelope modeling in speech bandwidth extension," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pp. 5930–5934, 2016. [Cited on page 37].

[130] Y. Wang, S. Zhao, J. Li, and J. Kuang, "Speech bandwidth extension using recurrent temporal restricted boltzmann machines," *IEEE Signal Processing Letters*, vol. 23, no. 12, pp. 1877–1881, 2016. [Cited on page 37].

[131] J. Makhoul and M. Berouti, "High-frequency regeneration in speech coding systems," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, pp. 428–431, 1979. [Cited on pages 38, 134].

[132] J. Fuemmeler, R. Hardie, and W. Gardner, "Techniques for the regeneration of wideband speech from narrowband speech," *EURASIP Journal on Applied Signal Processing*, vol. 2001, no. 1, pp. 266–274, 2001. [Cited on page 38].

[133] P. Jax and P. Vary, "Enhancement of band-limited speech signals," in *Proc. Aachen Symposium on Signal Theory*, pp. 331–336, 2001. [Cited on page 38].

[134] U. Kornagel, "Spectral widening of the excitation signal for telephone-band speech enhancement," in *Proc. Int. Workshop on Acoustic Echo and Noise Control*, pp. 215–218, 2001. [Cited on page 38].

[135] M. Turan and E. Erzin, "Synchronous overlap and add of spectra for enhancement of excitation in artificial bandwidth extension of speech.," in *Proc. INTERSPEECH*, pp. 2588–2592, 2015. [Cited on page 38].

[136] B. Iser, G. Schmidt, and W. Minker, *Bandwidth extension of speech signals*. Springer, 2008. [Cited on page 38].

[137] H. Pulakka, P. Alku, L. Laaksonen, and P. Valve, "The effect of highband harmonic structure in the artificial bandwidth expansion of telephone speech," in *Proc. Annual Conf. of the Int. Speech Communication Association*, 2007. [Cited on page 38].

[138] T. Van Pham, F. Schaefer, and G. Kubin, "A novel implementation of the spectral shaping approach for artificial bandwidth extension," in *Int. Conf. on Communications and Electronics*, pp. 262–267, 2010. [Cited on page 39].

[139] J. Kontio, L. Laaksonen, and P. Alku, "Neural network-based artificial bandwidth expansion of speech," *IEEE Transactions on Audio, Speech, and language processing*, vol. 15, no. 3, pp. 873–881, 2007. [Cited on page 39].

[140] H. Pulakka and P. Alku, "Bandwidth extension of telephone speech using a neural network and a filter bank implementation for highband mel spectrum," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2170–2183, 2011. [Cited on page 39].

[141] A. Uncini, F. Gobbi, and F. Piazza, "Frequency recovery of narrow-band speech using adaptive spline neural networks," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pp. 997–1000, 1999. [Cited on page 39].

[142] L. Laaksonen, J. Kontio, and P. Alku, "Artificial bandwidth expansion method to improve intelligibility and quality of AMR-coded narrowband speech," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. I–809, 2005. [Cited on page 39].

[143] H. Pulakka, V. Myllyla, L. Laaksonen, and P. Alku, "Bandwidth extension of telephone speech using a filter bank implementation for highband mel spectrum," in *18th European Signal Processing Conference*, pp. 979–983, 2010. [Cited on page 39].

[144] K. Li and C.-H. Lee, "A deep neural network approach to speech bandwidth expansion," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4395–4399, 2015. [Cited on pages 39, 40, 44].

[145] B. Liu, J. Tao, Z. Wen, Y. Li, and D. Bukhari, "A novel method of artificial bandwidth extension using deep architecture," in *Proc Annual Conf. of Int. Speech Communication Association*, 2015. [Cited on pages 39, 44].

[146] B. Liu, J. Tao, *et al.*, "A novel research to artificial bandwidth extension based on deep BLSTM recurrent neural networks and exemplar-based sparse representation," 2016. [Cited on page 39].

[147] Y. Gu, Z.-H. Ling, and L.-R. Dai, "Speech bandwidth extension using bottleneck features and deep recurrent neural networks.," in *Proc. INTERSPEECH*, pp. 297–301, 2016. [Cited on pages 39, 103].

[148] D. Yu and M. Seltzer, "Improved bottleneck features using pretrained deep neural networks," in *Proc. Annual Conf. of the Int. Speech Communication Association*, 2011. [Cited on pages 39, 95].

[149] R. Peharz, G. Kapeller, P. Mowlaee, and F. Pernkopf, "Modeling speech with sum-product networks: Application to bandwidth extension," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 3699–3703, 2014. [Cited on page 40].

[150] M. Zöhrer, R. Peharz, and F. Pernkopf, "On representation learning for artificial bandwidth extension," in *Proc. Annual Conf. of the Int. Speech Communication Association*, 2015. [Cited on page 40].

[151] J. Sadasivan, S. Mukherjee, and C. S. Seelamantula, "Joint dictionary training for bandwidth extension of speech signals," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 5925–5929, 2016. [Cited on page 40].

[152] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984. [Cited on page 40].

[153] A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio.," *SSW*, vol. 125, 2016. [Cited on page 40].

[154] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015. [Cited on page 40].

[155] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *arXiv preprint arXiv:1412.7062*, 2014. [Cited on page 40].

[156] Y. Gu and Z.-H. Ling, "Waveform modeling using stacked dilated convolutional neural networks for speech bandwidth extension.," in *Proc. INTERSPEECH*, pp. 1123–1127, 2017. [Cited on page 40].

[157] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. Courville, and Y. Bengio, "Samplernn: An unconditional end-to-end neural audio generation model," *arXiv preprint arXiv:1612.07837*, 2016. [Cited on page 40].

[158] Z.-H. Ling, Y. Ai, Y. Gu, and L.-R. Dai, "Waveform modeling and generation using hierarchical recurrent neural networks for speech bandwidth extension," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 5, pp. 883–894, 2018. [Cited on page 40].

[159] V. Kuleshov, S. Enam, and S. Ermon, "Audio super resolution using neural networks," *arXiv preprint arXiv:1708.00853*, 2017. [Cited on page 41].

[160] T. Lim, R. Yeh, Y. Xu, M. Do, and M. Hasegawa-Johnson, "Time-frequency networks for audio super-resolution," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 646–650, 2018. [Cited on page 41].

[161] Z. Jin, A. Finkelstein, G. Mysore, and J. Lu, "FFTNet: A real-time speaker-dependent neural vocoder," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2251–2255, 2018. [Cited on page 41].

[162] B. Feng, Z. Jin, J. Su, and A. Finkelstein, "Learning bandwidth expansion using perceptually-motivated loss," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 606–610, 2019. [Cited on page 41].

[163] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Advanced Neural Information Processing Systems (NIPS)*, pp. 2672–2680, 2014. [Cited on page 41].

[164] J. Sautter, F. Faubel, M. Buck, and G. Schmidt, "Artificial bandwidth extension using a conditional generative adversarial network with discriminative training," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7005–7009, 2019. [Cited on pages 41, 42].

[165] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. on computer vision and pattern recognition*, pp. 1125–1134, 2017. [Cited on page 41].

[166] S. E. Eskimez and K. Koishida, "Speech super resolution generative adversarial network," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3717–3721, 2019. [Cited on page 41].

[167] K. Roth, A. Lucchi, S. Nowozin, and T. Hofmann, "Stabilizing training of generative adversarial networks through regularization," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 2018–2028, 2017. [Cited on page 41].

[168] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. on computer vision*, pp. 2223–2232, 2017. [Cited on page 41].

[169] M. Nilsson and W. B. Kleijn, "Avoiding over-estimation in bandwidth extension of telephony speech," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 869–872, 2001. [Cited on page 42].

[170] J. Sautter, F. Faubel, M. Buck, and G. Schmidt, "Discriminative training of deep regression networks for artificial bandwidth extension," in *IEEE Int. Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 540–544, 2018. [Cited on page 42].

[171] M. Nilsson, H. Gustaftson, S. Andersen, and W. Kleijn, "Gaussian mixture model based mutual information estimation between frequency bands in speech," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, pp. I–525, 2002. [Cited on pages 42, 70, 74, 78].

[172] P. Jax and P. Vary, "Feature selection for improved bandwidth extension of speech signals," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. I–697, 2004. [Cited on pages 42, 44, 70, 78, 82, 84, 94].

[173] R. Kohavi and G. John, "Wrappers for feature subset selection," *Artificial intelligence*, vol. 97, no. 1-2, pp. 273–324, 1997. [Cited on page 43].

[174] J. Sautter, F. Faubel, and G. Schmidt, "Feature selection for dnn-based bandwidth extension," in *Proc. Jahrestagung für Akustik (DAGA)*, 2018. [Cited on page 43].

[175] A. Nour-Eldin and P. Kabal, "Memory-based approximation of the gaussian mixture model framework for bandwidth extension of narrowband speech," in *Annual Conf. of the Int. Speech Communication Association*, 2011. [Cited on page 43].

[176] S. Möller, *Assessment and prediction of speech quality in telecommunications*. Springer Science & Business Media, 2000. [Cited on pages 44, 46].

[177] P. Loizou, "Speech quality assessment," in *Multimedia analysis, processing and communications*, pp. 623–654, Springer, 2011. [Cited on page 44].

[178] T. Bäckström, *Speech Coding: with Code-Excited Linear Prediction*. Springer, 2017. [Cited on page 45].

[179] "ITU-T Recommendation P. 800: Methods for subjective determination of transmission quality," 1996. [Cited on pages 45, 46, 47].

[180] "3GPP Recommendation P.805 : Subjective evaluation of conversational quality," 2007. [Cited on page 45].

[181] H. Pulakka, L. Laaksonen, S. Yrttiaho, V. Myllylä, and P. Alku, "Conversational quality evaluation of artificial bandwidth extension of telephone speech," *The Journal of the Acoustical Society of America*, vol. 132, no. 2, pp. 848–861, 2012. [Cited on pages 45, 46].

[182] "ITU-T Recommendation P. 800.1: Mean opinion score (MOS) terminology," 2016. [Cited on page 46].

[183] N. Côté, *Integral and diagnostic intrusive prediction of speech quality*. Springer Science & Business Media, 2011. [Cited on page 46].

[184] M. Guéguin, R. Le Bouquin-Jeannès, V. Gautier-Turbin, G. Faucon, and V. Barriac, "On the evaluation of the conversational speech quality in telecommunications," *EURASIP Journal on Advances in Signal Processing*, vol. 2008, p. 93, 2008. [Cited on page 46].

[185] A. Raake, *Speech quality of VoIP: assessment and prediction*. John Wiley & Sons, 2007. [Cited on page 46].

[186] E. Hänsler and G. Schmidt, *Speech and audio processing in adverse environments*. Springer Science & Business Media, 2008. [Cited on page 46].

[187] T. Barnwell, "Correlation analysis of subjective and objective measures for speech quality," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 5, pp. 706–709, 1980. [Cited on page 47].

[188] "ITU-T Recommendation P.862: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," 2001. [Cited on pages 47, 68, 69].

[189] "ITU-T Recommendation P.862.2: Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs," 2005. [Cited on pages 47, 69].

[190] "ITU-T Recommendation P.863: Perceptual objective listening quality assessment," 2011. [Cited on page 47].

[191] J. Abel, M. Kaniewska, C. Guillaume, W. Tirry, T. Fingscheidt, J. Abel, M. Kaniewska, C. Guillaume, W. Tirry, and T. Fingscheidt, "An instrumental quality measure for artificially bandwidth-extended speech signals," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 25, no. 2, pp. 384–396, 2017. [Cited on pages 47, 48].

[192] H. Pulakka, L. Laaksonen, and P. Alku, "Quality improvement of telephone speech by artificial bandwidth expansion-listening tests in three languages," in *Ninth Int. Conf. on Spoken Language Processing*, 2006. [Cited on page 47].

[193] H. Pulakka, L. Laaksonen, M. Vainio, J. Pohjalainen, and P. Alku, "Evaluation of an artificial speech bandwidth extension method in three languages," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 6, pp. 1124–1137, 2008. [Cited on page 47].

[194] S. Möller, E. Kelaidi, F. Köster, N. Côté, P. Bauer, T. Fingscheidt, T. Schlien, H. Pulakka, and P. Alku, "Speech quality prediction for artificial bandwidth extension algorithms.," in *Proc. INTERSPEECH*, pp. 3439–3443, 2013. [Cited on pages 47, 69].

[195] P. Bauer, C. Guillaumé, W. Tirry, and T. Fingscheidt, "On speech quality assessment of artificial bandwidth extension," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6082–6086, 2014. [Cited on pages 48, 69].

[196] H. Pulakka, V. Myllylä, A. Rämö, and P. Alku, "Speech quality evaluation of artificial bandwidth extension: Comparing subjective judgments and instrumental predictions," in *Proc. Annual Conf. of the Int. Speech Communication Association*, 2015. [Cited on page 48].

[197] J. Abel, M. Kaniewska, C. Guillaume, W. Tirry, H. Pulakka, V. Myllylä, J. Sjöberg, P. Alku, I. Katsir, D. Malah, *et al.*, "A subjective listening test of six different artificial bandwidth extension approaches in English, Chinese, German, and Korean," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5915–5919, 2016. [Cited on page 48].

[198] Y.-t. Sha, C. Bao, M.-s. Jia, and X. Liu, "High frequency reconstruction of audio signal based on chaotic prediction theory," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 381–384, 2010. [Cited on page 49].

[199] H. Liu, C. Bao, X. Liu, X. Zhang, and L. Zhang, "Audio bandwidth extension based on RBF neural network," in *IEEE Int. Symposium on Signal Processing and Information Technology (ISSPIT)*, pp. 150–154, 2011. [Cited on page 49].

[200] H.-j. Liu, C. Bao, and X. Liu, "Spectral envelope estimation used for audio bandwidth extension based on RBF neural network," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 543–547, 2013. [Cited on page 49].

[201] X. Zhang, C. Bao, X. Liu, and L. Zhang, "A blind bandwidth extension method of audio signals based on Volterra series," in *Proc. Asia Pacific Signal and Information Processing Association (APSIPA) Annual Summit and Conference*, pp. 1–4, 2012. [Cited on page 49].

[202] H. Kantz and T. Schreiber, *Nonlinear time series analysis*. Cambridge university press, 2004. [Cited on page 49].

[203] C. Bao, X. Liu, Y.-T. Sha, and X.-T. Zhang, "A blind bandwidth extension method for audio signals based on phase space reconstruction," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2014, no. 1, pp. 1–9, 2014. [Cited on pages 49, 133].

[204] X. Liu, C. Bao, M.-s. Jia, and Y.-t. Sha, "Nonlinear bandwidth extension based on nearest-neighbor matching," in *Proc. Asia Pacific Signal and Information Processing Association (APSIPA) Annual Summit and Conference*, pp. 169–172, 2010. [Cited on pages 49, 50].

[205] X. Liu, C. Bao, L. Zhang, X. Zhang, F. Bao, and B. Bu, "Nonlinear bandwidth extension of audio signals based on hidden Markov model," in *IEEE Int. Symposium on Signal Processing and Information Technology (ISSPIT)*, pp. 144–149, 2011. [Cited on pages 50, 133].

[206] X. Liu and C. Bao, "Audio bandwidth extension based on ensemble echo state networks with temporal evolution," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 24, no. 3, pp. 594–607, 2016. [Cited on page 50].

[207] X. Liu and C. Bao, "Audio bandwidth extension based on temporal smoothing cepstral coefficients," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2014, no. 1, pp. 1–16, 2014. [Cited on page 50].

[208] E. Larsen, R. Aarts, and M. Danessis, "Efficient high-frequency bandwidth extension of music and speech," in *Audio Engineering Society Convention*, Audio Engineering Society, 2002. [Cited on pages xviii, 50, 133, 141, 142].

[209] J. Abel, E. Seidel, and T. Fingscheidt, "Enhancing the EVS Codec in Wideband Mode by Blind Artificial Bandwidth Extension to Superwideband," in *Proc. IEEE Int. Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 281–285, 2018. [Cited on page 50].

[210] B. Geiser and P. Vary, "Artificial bandwidth extension of wideband speech by pitch-scaling of higher frequencies," *INFORMATIK 2013–Informatik angepasst an Mensch, Organisation und Umwelt*, 2013. [Cited on pages 50, 133].

[211] Y. Wang, S. Zhao, K. Mohammed, S. Bukhari, and J. Kuang, "Superwideband extension for AMR-WB using conditional codebooks," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3695–3698, 2014. [Cited on page 50].

[212] J. Makhoul, "Spectral linear prediction: Properties and applications," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 23, no. 3, pp. 283–296, 1975. [Cited on pages 55, 79].

[213] J. Markel and A. Gray, *Linear prediction of speech*. Springer Science & Business Media, 2013. [Cited on page 55].

[214] C. Bishop, *Pattern recognition and machine learning*. Springer, 2006. [Cited on pages 55, 57, 84, 94, 111].

[215] S. Kay, *Fundamentals of statistical signal processing*. Prentice Hall PTR, 1993. [Cited on page 57].

[216] M. Goodwin, "The STFT, Sinusoidal Models, and Speech Modification," in *Springer handbook of speech processing*, pp. 229–258, Springer, 2007. [Cited on pages 62, 91, 135, 141].

[217] T. Dutoit and F. Marques, *Applied Signal Processing: A MATLAB-Based Proof of Concept*. Springer, 2010. [Cited on pages 63, 135, 141].

[218] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, and D. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon technical report N*, vol. 93, 1993. [Cited on page 63].

[219] P. Kabal, "TSP Speech Database," *McGill University, Database Version : 1.0*, pp. 02–10, 2002. [Online]: `http://mmsp.ece.mcgill.ca/Documents/Data/`. [Cited on page 64].

[220] E. Rothauser, "IEEE recommended practice for speech quality measurements," *IEEE Transactions on Audio and Electroacoustics*, vol. 17, pp. 225–246, 1969. [Cited on page 64].

[221] J. Kominek and A. Black, "CMU ARCTIC databases for speech synthesis," 2003. [Online] : `http://festvox.org/cmu_arctic/index.html`. [Cited on page 64].

[222] A. Black and K. Tokuda, "The Blizzard challenge 2005: Evaluating corpus-based speech synthesis on common databases," in *Proc. INTERSPEECH*, 2005. [Cited on page 64].

[223] "ITU-T Recommendation P. 501, Test signals for use in telephonometry," 2012. [Online]: `https://www.itu.int/rec/T-REC-P.501-201201-I/en`. [Cited on page 65].

[224] P. Kabal, "The AFsp package," 2002. [Online]: `http://www-mmsp.ece.mcgill.ca/Documents/Downloads/AFsp/`. [Cited on pages 65, 140].

[225] "ITU-T Recommendation G. 191: Software tools for speech and audio coding standardization," 2005. [Cited on page 65].

[226] "ITU-T Recommendation G. 191, Software Tool Library 2009 User's Manual," 2009. [Cited on pages 65, 141].

[227] "ITU-T Recommendation P. 56, Objective measurement of active speech level," 2011. [Cited on pages 65, 140].

[228] L. Rabiner, B.-H. Juang, and J. Rutledge, *Fundamentals of speech recognition*. PTR Prentice Hall Englewood Cliffs, 1993. [Cited on pages 67, 68].

[229] A. Gray and J. Markel, "Distance measures for speech processing," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 5, pp. 380–391, 1976. [Cited on page 68].

[230] "ITU-T Recommendation P.862: Mapping function for transforming P.862 raw result scores to MOS-LQO," 2003. [Cited on page 69].

[231] T. Cover and J. Thomas, *Elements of information theory*. John Wiley & Sons, 2012. [Cited on page 70].

[232] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech communication*, vol. 52, no. 1, pp. 12–40, 2010. [Cited on page 82].

[233] I. Jolliffe, *Principal component analysis*. Springer (Second Ed.), 2002. [Cited on page 94].

[234] J. Gehring, Y. Miao, F. Metze, and A. Waibel, "Extracting deep bottleneck features using stacked auto-encoders," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3377–3381, 2013. [Cited on page 95].

[235] T. Sainath, B. Kingsbury, and B. Ramabhadran, "Auto-encoder bottleneck features using deep belief networks," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4153–4156, 2012. [Cited on page 95].

[236] S. Takaki and J. Yamagishi, "A deep auto-encoder based low-dimensional feature extraction from fft spectral envelopes for statistical parametric speech synthesis," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5535–5539, 2016. [Cited on page 95].

[237] M. H. Soni, T. B. Patel, and H. A. Patil, "Novel subband autoencoder features for detection of spoofed speech.," in *Interspeech*, pp. 1820–1824, 2016. [Cited on page 95].

[238] L. Deng, M. L. Seltzer, D. Yu, A. Acero, A.-r. Mohamed, and G. Hinton, "Binary coding of speech spectrograms using a deep auto-encoder," in *Proc. Annual Conf. of Int. Speech Communication Association*, 2010. [Cited on page 95].

[239] S. H. Mohammadi and A. Kain, "Semi-supervised training of a voice conversion mapping function using a joint-autoencoder," in *Proc. Annual Conf. of Int. Speech Communication Association*, 2015. [Cited on page 95].

[240] S. H. Mohammadi and A. Kain, "A voice conversion mapping function based on a stacked joint-autoencoder.," in *Proc. INTERSPEECH*, pp. 1647–1651, 2016. [Cited on page 95].

[241] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`. [Cited on pages 96, 99].

[242] P. Baldi and K. Hornik, "Neural networks and principal component analysis: Learning from examples without local minima," *Neural networks*, vol. 2, no. 1, pp. 53–58, 1989. [Cited on page 96].

[243] H. Bourlard and Y. Kamp, "Auto-association by multilayer perceptrons and Singular Value Decomposition," *Biological cybernetics*, vol. 59, no. 4-5, pp. 291–294, 1988. [Cited on page 96].

[244] N. Japkowicz, S. J. Hanson, and M. Gluck, "Nonlinear autoassociation is not equivalent to PCA," *Neural computation*, vol. 12, no. 3, pp. 531–545, 2000. [Cited on page 96].

[245] Y. Bengio, Y. LeCun, *et al.*, "Scaling learning algorithms towards ai," *Large-scale kernel machines*, vol. 34, no. 5, pp. 1–41, 2007. [Cited on page 97].

[246] Y. Bengio *et al.*, "Learning deep architectures for ai," *Foundations and trends® in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009. [Cited on page 97].

[247] G. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006. [Cited on page 97].

[248] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *Journal of Machine Learning Research*, vol. 11, no. Dec, pp. 3371–3408, 2010. [Cited on pages 97, 100].

[249] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 153–160, 2007. [Cited on pages 97, 100].

[250] D. Erhan, P.-A. Manzagol, Y. Bengio, S. Bengio, and P. Vincent, "The difficulty of training deep architectures and the effect of unsupervised pre-training," in *Artificial Intelligence and Statistics*, pp. 153–160, 2009. [Cited on page 97].

[251] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio, "Why does unsupervised pre-training help deep learning?," *Journal of Machine Learning Research*, vol. 11, no. Feb, pp. 625–660, 2010. [Cited on page 97].

[252] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. Int. Conf. on Artificial Intelligence and Statistics*, pp. 249–256, 2010. [Cited on pages 97, 98].

[253] J. Bergstra, G. Desjardins, P. Lamblin, and Y. Bengio, "Quadratic polynomials learn better image features," *Technical report, 1337*, 2009. [Cited on page 98].

[254] K. He *et al.*, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. on computer vision*, pp. 1026–1034, 2015. [Cited on pages 98, 102, 123].

[255] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. Int. Conf. on artificial intelligence and statistics*, pp. 315–323, 2011. [Cited on page 98].

[256] A. Maas, A. Hannun, and A. Ng, "Rectify nonlinearities improve neural network acoustic model," in *ICML Workshop on Deep Learning for Audio, Speech, and Language Processing*, 2013. [Cited on page 98].

[257] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," *arXiv preprint arXiv:1511.07289*, 2015. [Cited on page 98].

[258] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 1097–1105, 2012. [Cited on page 98].

[259] M. Zeiler, M. Ranzato, R. Monga, M. Mao, K. Yang, Q. V. Le, P. Nguyen, A. Senior, V. Vanhoucke, J. Dean, *et al.*, "On rectified linear units for speech processing," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3517–3521, 2013. [Cited on page 98].

[260] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012. [Cited on page 98].

[261] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014. [Cited on pages 99, 103].

[262] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. conf. on machine learning (ICML)*, pp. 448–456, 2015. [Cited on pages 99, 103, 123].

[263] J. Guo, U. A. Nookala, and A. Alwan, "CNN-based joint mapping of short and long utterance i-vectors for speaker verification using short utterances," *Proc. INTERSPEECH*, pp. 3712–3716, 2017. [Cited on page 101].

[264] F. Chollet *et al.*, "Keras." `https://github.com/keras-team/keras`, 2015. [Cited on pages 102, 123].

[265] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014. [Cited on pages 102, 123].

[266] D. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013. [Cited on pages xvii, 110, 115].

[267] C. Doersch, "Tutorial on variational autoencoders," *arXiv preprint arXiv:1606.05908*, 2016. [Cited on page 115].

[268] X. Yan, J. Yang, K. Sohn, and H. Lee, "Attribute2image: Conditional image generation from visual attributes," in *Proc. European Conference on Computer Vision*, pp. 776–791, 2016. [Cited on pages 115, 116, 118].

[269] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 3483–3491, 2015. [Cited on pages 115, 116, 118].

[270] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling, "Semi-supervised learning with deep generative models," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 3581–3589, 2014. [Cited on page 118].

[271] W.-N. Hsu, Y. Zhang, and J. Glass, "Learning latent representations for speech generation and transformation," pp. 1273–1277, 2017. [Cited on page 118].

[272] M. Blaauw and J. Bonada, "Modeling and transforming speech using variational autoencoders.," in *Proc. INTERSPEECH*, pp. 1770–1774, 2016. [Cited on pages 118, 125].

[273] C.-C. Hsu *et al.*, "Voice conversion from non-parallel corpora using variational auto-encoder," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA) Annual Summit and Conference*, pp. 1–6, 2016. [Cited on page 118].

[274] K. Akuzawa, Y. Iwasawa, and Y. Matsuo, "Expressive speech synthesis via modeling expressions with variational autoencoder," pp. 3067–3071, 2018. [Cited on page 118].

[275] Y. Jung, Y. Kim, Y. Choi, and H. Kim, "Joint learning using denoising variational autoencoders for voice activity detection," pp. 1210–1214, 2018. [Cited on page 118].

[276] S. Latif, R. Rana, J. Qadir, and J. Epps, "Variational autoencoders for learning latent representations of speech emotion," pp. 3107–3111, 2018. [Cited on page 118].

[277] L. Pandey, A. Kumar, and V. Namboodiri, "Monoaural audio source separation using variational autoencoders," pp. 3489–3493, 2018. [Cited on page 118].

[278] R. Schafer and A. Oppenheim, *Discrete-time signal processing*. Prentice Hall Englewood Cliffs, NJ, 1989. [Cited on pages xviii, 135, 136].

[279] J. Proakis, *Digital signal processing: principles algorithms and applications*. Pearson Education India, 2001. [Cited on page 136].

[280] "3GPP TS 26.442: Codec for Enhanced Voice Services; ANSI C Code (fixed point)," 2015. [Online]: `https://www.etsi.org/deliver/etsi_ts/126400_126499/126442/13.03.00_60, ver. 13.0.0 Rel. 13, 2016` (Last accessed : April 2019). [Cited on page 141].

[281] "ITU-T Recommendation P.341: Transmission Characteristics for Wideband Digital Loudspeaking and Hands-free Telephony Terminals," 2011. [Cited on page 141].

[282] "3GPP TS 26.173: ANSI-C Code for the Adaptive Multi-Rate - Wideband (AMR-WB) speech codec," 2002. [Online]: `https://www.etsi.org/deliver/etsi_ts/126100_126199/126173/13.01.00_60`, ver. 13.1.0 Rel. 13, 2016. [Cited on page 141].

[283] A. Rämö, "Voice quality evaluation of various codecs," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4662–4665, 2010. [Cited on page 142].

[284] M. Florentine, S. Buus, and C. Mason, "Level discrimination as a function of level for tones from 0.25 to 16 khz," *The Journal of the Acoustical Society of America*, vol. 81, no. 5, pp. 1528–1541, 1987. [Cited on page 145].