FACULTY OF ENGINEERING
AND ARCHITECTURE

Single-Channel Speech Enhancement with Prior Knowledge

Yanjue Song

Doctoral dissertation submitted to obtain the academic degree of
Doctor of Engineering

Supervisors
Prof. Nilesh Madhu, PhD - Prof. Kris Demuynck, PhD
Department of Electronics and Information Systems
Faculty of Engineering and Architecture, Ghent University

March 2025

GHENT
UNIVERSITY

ISBN 978-94-6355-963-8
NUR 959, 984
Wettelijk depot: D/2025/10.500/23

Members of the Examination Board

Chair

Prof. Hennie De Schepper, PhD, Ghent University

Other members entitled to vote

Prof. Tony Belpaeme, PhD, Ghent University
Prof. Paul Devos, PhD, Ghent University
Prof. Hong-Goo Kang, PhD, Yonsei University, South Korea
Prof. Gerhard Schmidt, PhD, Christian-Albrechts-Universität zu Kiel, Germany

Supervisors

Prof. Nilesh Madhu, PhD, Ghent University
Prof. Kris Demuynck, PhD, Ghent University

Acknowledgments

Completing this dissertation has been a journey shaped by the support of many incredible people. To everyone who walked this path with me, thank you.

My deepest gratitude goes to my promoter, Prof. Nilesh Madhu: None of this would have been possible without your faith in my potential. I still remember the thrill of receiving your call offering me this position—a moment that launched an adventure far richer than I had ever imagined. Our discussions in meeting rooms, punctuated by scribbles on the whiteboard, transformed my thinking in ways I will carry forward. Thank you also for the unwavering encouragement during moments of self-doubt and for showing me how to turn setbacks into stepping stones.

To my co-supervisor, Prof. Kris Demuynck: Thank you for your unique insights and your knack for reframing scientific challenges from fresh perspectives. My Dutch summary would have been much more robotic without your revision.

To the ASPIRE group: You turned our office into a second home. I enjoyed our lively academic discussions as much as the debate over vegetables during countless lunches (where *many* sandwiches were consumed) or our sporadic after-work gatherings with various themes. Thank you for being both colleagues and friends. My days would have been far dimmer without your companionship.

A special thank-you to Prof. Hong-Goo Kang and your DSP&AI group at Yonsei University: I am deeply grateful for the warm and considerate hospitality extended to me during my two-month stay in Seoul. Even though I did not know any Korean, you made me feel at home from day one. Your thoughtful good-bye gift, the mouse pad, sits faithfully by my keyboard as I write these words. Thank you for showing me the short-cut through the labyrinth of speech synthesis. Seoul will stay in my memory with crisp autumn skies and the aroma of warm coffee around every corner.

Beyond academia, this journey was brightened by the people who grounded me.

To Chunhui: Just as your name implies, you are my sunshine on rainy days in Belgium. Thank you for your endless support—the warm meals that eased my anxiety when data and algorithms refused to cooperate, your heartfelt celebrations of every small success, and the laughter you brought into my toughest days. I am truly grateful for the special moments we have shared.

To Yang: Thank you for being my emotional safe haven. Your non-judgmental ear, whether over a dinner table or via video call, has been a constant source of strength. Our reunions across cities and time zones are treasures I hold dear.

To Yingyi and Jiajia: Our friendship has spanned seven time zones, two con-

tinents, and survived five relocations. Despite the distance, your messages have always made the world feel a little smaller, reminding me that a shred of my past continues to shine in my present.

To my family: Your love has been my anchor throughout this journey. Thank you for supporting my less conventional career and life choices—even when you might not have fully agreed with them. You taught me to embrace uncertainty long before I began this PhD, and I am forever grateful for your trust and encouragement for me to carve my own path.

To all those who have been part of this journey, whether mentioned here or not: thank you for making the journey as meaningful as the destination.

Ghent, March 2025
Yanjue Song

Table of Contents

Summary **ix**

Samenvatting **xiii**

1 Introduction **1**

2 Background **3**
2.1 Signal Model 4
2.1.1 Source-Filter Model 4
2.2 Statistical Methods 5
2.2.1 Framework 5
2.2.2 Gain Function 7
2.2.3 Improved *A Priori* SNR Estimation 8
2.2.4 Contributions 9
2.3 Deep Learning Methods 11
2.3.1 Training Targets 11
2.3.2 Loss Functions 12
2.3.3 Architectures 14
2.3.4 Contributions 15
2.4 Phase Reconstruction 16
2.4.1 Phase Retrieval 16
2.4.2 Contributions 17
2.5 Self-Supervised Learning (SSL) 18
2.5.1 SSL Models 18
2.5.2 Neural Vocoder 19
2.5.3 Contributions 21
2.6 Publications 22
2.6.1 Publications in international journals (listed in the Science Citation Index) 22
2.6.2 Publications in International Conferences 22
References 24

3 Improved CEM for Speech Harmonic Enhancement **31**
3.1 Introduction 32
3.2 Cepstral excitation manipulation (CEM) baseline 35

3.2.1 Overview of CEM-based Speech Enhancement Framework 35
3.2.2 CEM_{ID} in detail: F_0 Detection 36
3.2.3 CEM_{ID} in detail: Excitation Manipulation 38
3.2.4 Speech Estimation 38
3.3 Improved Excitation Manipulation 39
3.3.1 Analysis of the drawbacks of CEM 39
3.3.2 Residual Amplitude Estimation (RAE) 41
3.3.2.1 Adaptive Amplifying Factor τ 42
3.3.2.2 Amplitude Decay $\omega_l(m)$ 43
3.3.3 Cepstrum Smoothing (CC) 44
3.4 Evaluation 46
3.4.1 Experimental Setup 47
3.4.2 Noise estimation 48
3.4.3 Quality Measures 48
3.4.4 Experimental Results and Discussion 50
3.4.4.1 Overall Average of Instrumental Metrics 52
3.4.4.2 ΔPESQ and $\Delta\text{PESQ}_{\text{act}}$ 53
3.4.4.3 POLQA 54
3.4.4.4 NA and SSDR 55
3.4.4.5 STOI 57
3.5 Conclusions 58
References 60

4 Speech Harmonic Recovery in Deep Neural Network via Cepstral Excitation Manipulation Loss 63
4.1 Introduction 64
4.2 Signal model and CRUSE architecture 66
4.2.1 Signal model 66
4.2.2 CRUSE architecture 67
4.3 Exploitation of excitation information 69
4.3.1 F0 detection 69
4.3.2 CEM loss function component 69
4.3.3 Harmonic indicator 70
4.3.4 Summary 70
4.4 Experimental evaluation 70
4.5 Results and discussion 71
4.6 Conclusions 73
References 75

5 Gradient Based Phase Reconstruction 77
5.1 Introduction 78
5.2 STFT domain speech enhancement 80
5.2.1 DNN baselines: CRUSE and Complex CRUSE (C-CRUSE) 81
5.2.2 Phase derivatives 81
5.3 Phase estimation 82

5.3.1 Estimating $\Delta_f \Phi(l,m)$ and $\Delta_t \Psi(l,m)$ 82
5.3.2 Phase retrieval from clean speech amplitudes 82
5.3.3 Phase reconstruction for speech enhancement 84
5.4 Experimental evaluation and discussion 84
5.4.1 Results & Discussion 86
5.5 Conclusions 88
References 90

6 Disentanglement and Robustness of Self-Supervised Speech Representations 93
6.1 Introduction 94
6.2 Experiments 96
6.2.1 Methodology 96
6.2.2 Metrics 98
6.3 Results 99
6.3.1 Preserved Information 99
6.3.2 Distortion 100
6.4 Conclusions 102
References 103

7 Spectrum-Aware Neural Vocoder Based on Self-Supervised Learning for Speech Enhancement 105
7.1 Introduction 106
7.2 Methodology 108
7.2.1 Generative Model Framework 108
7.2.2 Feature Extraction 108
7.2.3 Neural Vocoder: HiFi-GAN 109
7.2.4 Fusion Methods 109
7.2.5 Research Questions 111
7.3 Experiment Setup 111
7.3.1 Metrics 111
7.3.2 Ablation Studies 112
7.4 Results and Discussions 114
7.5 Conclusions 115
References 117

8 Conclusions and Future Work 121
8.1 Conclusions 121
8.2 Future Research 123

A Optimal Estimation of Speech Envelopes for the Two-Stage Speech Enhancement 125
A.1 Introduction 127
A.2 Speech Enhancement Framework 131
A.3 Cepstral Envelope Estimation 133

A.3.1 Feature Extraction 134
A.3.1.1 LPCC 134
A.3.1.2 Cepstral Coefficients (CCoef) 134
A.3.2 Codebook 135
A.3.3 Envelope Estimator 137
A.3.3.1 Feedforward DNN Classifier 137
A.3.3.2 Recurrent Neural Network-Based Classifier . . 137
A.3.3.3 CRNN-Based Envelope Estimation by Regression 138
A.4 Evaluations 138
A.4.1 Experimental Setup 138
A.4.2 Quality Measures 140
A.4.3 Oracle Test Results 141
A.4.3.1 Codebook Optimisation: Findings from Oracle Tests 142
A.4.3.2 Feature Representation 143
A.4.3.3 Quantisation Error 143
A.4.4 Practical System Evaluation Results 144
A.4.4.1 Impact of Separate Codebook 144
A.4.4.2 Comparative Benchmark against DNN Baseline 145
A.5 Conclusions 148
References 150

B List of Acronyms **155**

Summary

Speech is a primary mode of human communication and has become increasingly important in our digital lives. Its role in online communication and human-machine interaction is growing as more and more devices are equipped with at least one microphone. However, real-world speech signals often suffer from environmental interference such as background noise and reverberation, making them difficult to comprehend or process. These distortions not only cause listening fatigue for human listeners but also degrade the performance of speech-based systems. Therefore, it is crucial to improve the quality and intelligibility of recorded speech in these challenging environments. This forms the key research topic of this dissertation: to enhance the speech signal captured by a single microphone in adverse conditions.

In recent years, speech enhancement methods have evolved from traditional statistical methods towards deep neural network (DNN)-based approaches. This trend has also changed the focus of research. Explicit signal models are required to develop and improve statistical approaches, where speech properties are thoroughly examined to advance the field. These methods, based on simplified signal models, have been proven effective when background noise changes slowly. With DNNs emerging as a more powerful alternative for modelling signal statistics, research has shifted towards optimising network architectures and loss functions to fully leverage the benefits of the general end-to-end training scheme, while signal models usually play a lesser role in DNN-based approaches. Although DNNs can model complex distributions, prior knowledge about speech remains valuable, particularly when training data or computational resources are limited.

This dissertation explores different means of incorporating prior knowledge into speech enhancement systems. We start with investigating how the speech estimate can be refined by using the source-filter model in statistical enhancement methods. To that end, Chapter 3 explores a simple way of introducing domain knowledge to complement the classical Gaussian or super Gaussian assumptions on the distribution of speech spectral coefficients. By closely examining the naïve cepstral excitation manipulation (CEM) method, which manipulates excitation signals of intermediate speech estimates in the cepstrum, we propose to optimise its speech preservation performance by two modifications: a) an adaptive harmonic emphasising factor related to the dynamic range of the excitation signal of the current frame, which guarantees sufficient enhancement for voiced frames without introducing artefacts in unvoiced frames; b) cepstral smoothing of excitation signals, which reduces musical noise while retaining essential fine structures.

Next, the enhancement of the speech spectral envelope, the other component of the source-filter decomposition, is investigated with a codebook-based approach. The spectral envelope of the initial speech estimate is matched to and then replaced by the closest template from the pre-defined codebook extracted from clean speech corpora. Our analysis of the codebook entry distribution indicates that a manual division of speech active/inactive frames, though proposed in existing research, is not necessary and even degrades the subsequent template matching accuracy. Furthermore, the investigation into envelope representations shows that codebooks based on cepstral coefficients outperform those using linear prediction coding. Building on this, we introduce a temporal modelling component through a recurrent neural network (RNN) to improve the envelope classification accuracy. Additionally, we study a regression-based approach for envelope enhancement to mitigate potential quantisation errors in the codebook-based approach. However, despite all the optimisation on the two-stage framework, the improvement in overall performance is limited, as presented in Appendix A.

This bottleneck may stem from inaccuracies in another key component of statistical speech enhancement systems, the noise estimator. However, accurately modelling noise distributions using statistical methods is highly challenging—if not impossible—due to their vast variety. In contrast, speech exhibits a more structured nature, making it well-suited for modelling with DNNs to achieve better performance.

Therefore, in Chapter 4, we examine the performance of an efficient DNN-based speech enhancement method, the convolutional recurrent U-net architecture for speech enhancement (CRUSE). It was observed in audio enhanced by CRUSE that weak speech harmonics are prone to be suppressed along with noise, which is similar to the problem addressed in Chapter 3 in conventional methods. Inspired by our work on CEM, we introduce the source-filter model into DNNs to mitigate this issue. Specifically, in addition to the globally averaged mean-square error (MSE), the importance of speech harmonics to speech quality is emphasised by an additional cepstral loss term to measure prediction errors in speech harmonics. To further reinforce harmonic preservation, an amplifying factor is employed to boost harmonics in clean references. Experiments demonstrate that the proposed loss function term effectively preserves speech harmonics without increasing computational cost during inference.

Traditionally, single-channel speech enhancement focusses on spectral magnitude estimation and leaves the noisy phase untouched. Although a distorted phase degrades speech quality less than a distorted magnitude, improving phase can still yield benefits—which is studied in Chapter 5. While absolute phase is tricky to estimate, phase derivatives are linked to log magnitudes, which allows the possibility to train DNNs to estimate these derivatives from magnitudes. This method, however, requires access to clean magnitudes, which are unavailable in the speech enhancement task. As an approximation thereof, we employ enhanced speech magnitudes (from the previous chapter) to estimate phase derivatives. Since fully synthetic phase reconstruction sounds unnatural, the final phase estimate is obtained by balancing initial estimates (noisy phases can be regarded as initial es-

timates as well) with synthetic ones. Experiments show that the proposed method further improves the quality of speech signals enhanced by CRUSE or its variants, providing the deep learning methods with complementary domain knowledge.

In the previous chapters, we explored explicit signal models for encoding domain knowledge regarding magnitude and phase of speech in the short-time Fourier transform (STFT) domain. These models were derived from human observation and understanding of speech signals. In the following chapters, we shift our focus to a deep-learning-based method, self-supervised learning (SSL) models, to encode prior knowledge in a data-driven manner.

SSL models learn to extract multiple levels of information from large datasets without requiring human annotation. Various experiments have demonstrated their ability to encode various speech-related information in the feature maps (embeddings) extracted from speech signals, enabling the reconstruction of natural speech. For speech enhancement, we adopt an SSL-based re-synthesis framework that synthesises clean speech signals from distorted embeddings extracted from captured mixtures of speech and noise.

To optimise such a system, Chapter 6 examines the information encoded in SSL embeddings and their robustness to interference. Our quantitative analysis finds that embeddings extracted by transformer encoder representations from alteration (TERA) are least affected by noise and reverberation among all tested models, while retaining useful acoustic, phonetic, and semantic details.

In Chapter 7, we propose the spectrum-aware SSL-based neural vocoder for speech enhancement, with TERA as the pre-trained SSL model. Existing research on speech synthesis has shown that SSL-based neural vocoders can generate more natural-sounding speech when provided with acoustic details missing in the embeddings. We extend this idea to speech enhancement by fusing SSL embeddings with noisy spectrograms at the vocoder, enabling it to extract relevant acoustic features directly from noisy spectrograms. Ablation studies on fusion methods confirm that the introduction of noisy spectrograms improves the quality of enhanced signals, especially in their naturalness.

Samenvatting
– Summary in Dutch –

Spraak is de primaire vorm van communicatie tussen mensen. Maar ook in ons digitaal leven wordt spraak steeds belangrijker. De rol van spraak in online communicatie en mens-machine-interactie groeit naarmate meer en meer apparaten worden uitgerust met ten minste èèn microfoon. Spraaksignalen in de echte wereld hebben echter vaak last van omgevingsinterferentie zoals achtergrondruis en galm, waardoor ze moeilijk te begrijpen of te verwerken zijn. Ruis en galm veroorzaken niet alleen luistermoeheid bij menselijke luisteraars, maar verminderen ook de prestaties van de computeralgoritmes die de spraak verwerken. Daarom is het cruciaal om de kwaliteit en verstaanbaarheid van opgenomen spraak in dergelijke uitdagende omgevingen te verbeteren. Dit vormt het onderwerp van dit proefschrift: het verbeteren van het spraaksignaal opgenomen door een enkele microfoon in ongunstige omstandigheden.

In de afgelopen jaren zijn spraakverbeteringsalgoritmes geëvolueerd van traditionele statistische methoden naar benaderingen op basis van diepe neurale netwerken (deep neural network, DNN). Deze trend heeft de focus van het onderzoek gewijzigd. Om statistische benaderingen te ontwikkelen en te verbeteren, zijn expliciete signaalmodellen vereist waarbij spraakeigenschappen grondig worden onderzocht. De resulterende algoritmes, gebaseerd op vereenvoudigde signaalmodellen, zijn effectief gebleken wanneer achtergrondgeluid langzaam wijzigt. DNN's bieden een krachtiger alternatief voor het modelleren van signaalstatistieken, wat maakt dat het onderzoek verschuift naar het optimaliseren van netwerkarchitecturen en kostfuncties om de voordelen van end-to-end trainingsschema's volledig te benutten. De signaalmodellen spelen daarbij maar een ondergeschikte rol. Hoewel DNN's complexe distributies kunnen modelleren, blijft voorkennis over spraak waardevol, vooral wanneer trainingsdata of rekenkracht beperkt zijn.

Dit proefschrift onderzoekt verschillende manieren om voorkennis in spraakverbeteringsalgoritmes te integreren. Ons onderzoek in hoofdstuk 3 begint met het verfijnen van de schatting van storingsvrije spraak in statistische verbeteringsalgoritmes, dit op basis van het bron-filter model. Het bron-filter model vormt een eenvoudige manier om domeinkennis te introduceren ter aanvulling van de aanname van een klassieke Gaussiaanse of super Gaussiaanse verdeling van de spectrale coëfficiënten van spraak. Na nauwkeurig onderzoek van de nogal naïve manipulatie van de tussentijdse spraakschattingen in het cepstrale domein door middel van CEM (cepstral excitation method), stellen we voor om CEM via twee aanpas-

singen zo aan te passen dat de spraakeigenschappen beter behouden blijven. De eerste aanpassing introduceert een adaptieve harmonische benadrukkingsfactor gerelateerd aan het dynamische bereik van het excitatiesignaal van het huidige frame, wat voldoende verbetering garandeert voor stemhebbende frames zonder artefacten in stemloze frames te introduceren. De tweede aanpassing betreft een cepstrale uitvlakking van de excitatiesignalen, wat muzikale ruis vermindert terwijl essentiële fijne structuur behouden blijft.

Voor de verbetering van de spectrale omhullende van de spraak, de filtercomponent in de bron-filter decompositie, wordt vervolgens gekeken naar een codeboekgebaseerde benadering. De spectrale omhullende van de initiële spraakschatting wordt gematcht en vervolgens vervangen door het dichtstbijzijnde sjabloon uit het vooraf gedefinieerde codeboek, een codeboek dat is geleerd uit ruisvrije spraakdata. Onze analyse van de codeboekdistributie geeft aan dat een rigide opdeling in spraakactieve/inactieve frames, hoewel voorgesteld in de literatuur, niet noodzakelijk is en zelfs de nauwkeurigheid van de daaropvolgende sjabloonmatching vermindert. Bovendien toont het onderzoek naar enveloprepresentaties aan dat codeboeken gebaseerd op cepstrale coëfficiënten beter presteren dan codeboeken gebaseerd op lineaire predictie. Voortbouwend hierop introduceren we een temporele modelleringscomponent door middel van een RNN (recurrent neural network) om zo de nauwkeurigheid van de classificatie van de spectrale omhullende te verbeteren. Daarnaast bestuderen we een regressie-gebaseerde benadering voor de verbetering van de omhullende om potentiële kwantisatiefouten in de codeboekgebaseerde benadering te verminderen. Ondanks alle optimalisaties, blijft de verbetering over het algemeen beperkt (zie appendix A).

De beperkte verbetering kan voortkomen uit onnauwkeurigheden in een andere sleutelcomponent van statistische spraakverbeteringssystemen, namelijk de ruisschatter. Het nauwkeurig modelleren van de ruisdistributie met behulp van statistische methoden is zeer uitdagend—zo niet onmogelijk—vanwege de grote variëteit in types ruis. Spraak daarentegen vertoont een meer gestructureerde aard, wat maakt dat spraak wel goed geschikt is voor modellering met DNN's om zo betere prestaties te bereiken.

Daarom onderzoeken we in Hoofdstuk 4 de prestaties van een efficiënte DNN-gebaseerde spraakverbeteringsmethode, met name CRUSE (convolutional recurrent U-net architecture for speech enhancement). We namen wel waar dat in audio verbeterd door CRUSE de zwakke spraakharmonischen geneigd zijn om samen met ruis te worden onderdrukt, wat vergelijkbaar is met het probleem dat in Hoofdstuk 3 via conventionele methoden werd aangepakt. Geïnspireerd door ons werk aan CEM, introduceren we het bron-filter model in de DNN's om dit probleem te mitigeren. Specifiek, naast de globaal gemiddelde *mean-square error (MSE)*, wordt het belang van spraakharmonischen voor de spraakkwaliteit benadrukt door een extra cepstrale verliesterm in de DNN-training om zo de voorspellingsfouten in spraakharmonischen te meten. Om het behoud van de harmonischen verder te ondersteunen, wordt een versterkingsfactor toegepast op de harmonischen van de ruisvrije spraakreferenties. Experimenten tonen aan dat de voorgestelde verliesterm effectief spraakharmonischen behoudt zonder de rekenkosten tijdens inferen-

tie te verhogen.

Traditioneel richt enkelkanaalsspraakverbetering zich op spectrale amplitudeschatting en laat de ruisfase onaangeroerd. Hoewel een vervormde spectrale fase de spraakkwaliteit minder degradeert dan een vervormde amplitude, kan faseverbetering nog steeds voordelen opleveren. Dit wordt bestudeerd in Hoofdstuk 5. De absolute fase is lastig te schatten. Fase-afgeleiden daarentegen zijn gerelateerd aan log-amplitudes, wat de mogelijkheid biedt om DNN's te trainen om deze afgeleiden te schatten op basis van amplitudes. Deze methode vereist echter toegang tot ruisvrije amplitudes, en deze zijn niet beschikbaar in de spraakverbeteringstaak. Als een benadering gebruiken we daarom verbeterde spraakamplitudes (uit het vorige hoofdstuk) om de fase-afgeleiden te schatten. Aangezien de volledig synthetische fasereconstructie onnatuurlijk klinkt, wordt de uiteindelijke faseschatting verkregen door initiële schattingen (de fase van het ruizige signaal kan ook als initiële schatting worden beschouwd) te balanceren met synthetische. Experimenten tonen aan dat de voorgestelde methode de kwaliteit van spraaksignalen verder verbetert bovenop de verbeteringen bekomen door CRUSE of varianten daarop.

In de vorige hoofdstukken onderzochten we expliciete signaalmodellen voor het coderen van domeinkennis met betrekking tot amplitudes en fases van spraak bekomen door kortetermijn Fourier-transformaties. Deze modellen zijn afgeleid van menselijke observatie en begrip van spraaksignalen. In de volgende hoofdstukken verschuiven we onze focus naar een op deep learning-gebaseerde methode, *self-supervised learning (SSL)*, om voorkennis op een data-gedreven manier te introduceren.

Verschillende experimenten hebben aangetoond dat ze in staat zijn om de diverse aspecten van spraak effectief te coderen in de *feature maps (embeddings)* die zijn geëxtraheerd uit spraaksignalen, in die mate dat zelfs de reconstructie van natuurlijke spraak uit deze embeddings mogelijk is. Voor de spraakverbetering passen we een op SSL-gebaseerd resynthese framewerk aan om ruisvrije spraaksignalen te synthetiseren uit vervormde embeddings berekend uit de opgenomen mix van ruis en spraak met galm.

Om zo'n systeem te optimaliseren, onderzoekt Hoofdstuk 6 de informatie die is gecodeerd in SSL-embeddings en hun robuustheid tegen interferentie. Onze kwantitatieve analyse toont aan dat embeddings geëxtraheerd door TERA *(transformer encoder representations from alteration)* het minst worden beïnvloed door ruis en galm van alle geteste modellen, terwijl de nuttige akoestische, fonetische en semantische details behouden blijven.

In Hoofdstuk 7 stellen we vervolgens de spectrum-bewuste SSL-gebaseerde neurale vocoder voor spraakverbetering voor, met TERA als het vooraf getrainde SSL-model. Bestaand onderzoek naar spraaksynthese heeft aangetoond dat SSL-gebaseerde neurale vocoders nog beter klinkende spraak kunnen genereren wanneer ze worden gevoed met extra akoestische details die ontbreken in de embeddings. We passen dit idee op een vernieuwende manier toe door spraakverbetering via SSL-embeddings te fuseren met ruisspectrogrammen bij de vocoder, waardoor de vocoder de relevante akoestische kenmerken direct uit de ruisspectrogrammen kan extraheren. Ablatiestudies over fusiemethoden bevestigen dat de introduc-

tie van de ruisspectrogrammen de kwaliteit van de verbeterde signalen verbetert, vooral wat betreft hun natuurlijkheid.

1 Introduction

Speech is one of the most natural and fundamental forms of human communication. In the digital era, with the rise of the internet and communication technologies, speech is no longer confined to face-to-face interactions. Clear speech capture has become essential for both long-distance online communication and for the development of efficient human-machine interfaces. However, achieving high-quality speech capture in real-world environments remains challenging due to environmental interference, such as background noise and reverberation. These distortions affect both the quality and intelligibility of speech, posing challenges to human listeners and automated systems. To enhance communication quality and ensure robust performance of downstream applications such as virtual assistants, speech enhancement, which aims to recover the clean speech from distorted recordings, is one fundamental task.

Research on improving the quality and intelligibility of recorded speech signals has been an active research area for several decades. Traditionally, when only one microphone is available for recording, speech enhancement has relied on statistical methods with mathematically tractable assumptions on the distributions of speech and noise signals. While effective to some extent, these assumptions often oversimplify the complex nature of real-world signals, leading to degraded performance, particularly in challenging conditions like in low signal-to-noise ratio (SNR) environments or with highly non-stationary noise.

In recent years, deep learning-based methods have emerged as a powerful alternative. Their ability to model complex distributions significantly improves speech enhancement performance in challenging scenarios. Deep learning methods learn

the mapping from distorted signals to clean speech signals in a data-driven manner. However, despite the success of deep learning methods compared to the traditional statistical predecessors, their general end-to-end training scheme pays little attention to incorporating prior knowledge, which could potentially benefit the overall speech enhancement performance. There are various sources of prior knowledge: the domain knowledge summarised by humans and expressed as explicit mathematical models, such as the structure of magnitude or phase; or speech-related information extracted from observed mixtures by a fully data-driven method—knowledge discovered by representation learning approaches such as self-supervised learning (SSL).

This dissertation seeks to address this gap by exploring how to make better use of various types of prior knowledge in speech enhancement systems. Specifically, the dissertation investigates the integration of the following speech models:

- The source-filter model based on a simplified speech production mechanism;
- The gradient-based phase retrieval method;
- SSL models which learn to extract speech-related features from large amounts of data without manual annotation.

Different approaches are proposed to combine these models with the speech enhancement systems. We start with the optimisation of the existing two-stage statistical speech enhancement framework based on the source-filter model. By individually idealising the two components of speech signals decomposed by this model, the initial estimate is further refined. Our optimisation shows the benefit of more realistic modelling during refinement. However, an in-depth analysis reveals the limitations of shallow neural networks combined with statistical noise estimators, prompting a shift towards more advanced deep learning methods. Since clean speech magnitude and phase, the two components of speech in the time-frequency domain, show diverse properties according to the domain knowledge about speech, we propose to separately enhance their estimates by incorporating individual signal models. In the final part of this dissertation, we investigate the potential of emerging SSL models as prior knowledge encoders, aiming to generate high-quality, natural speech under adverse conditions.

By integrating these varied approaches, this dissertation contributes to advancing the research on obtaining high-quality speech from signals recorded across diverse and challenging environments with different prior information about speech.

2
Background

This chapter provides an overview of the existing single-channel speech enhancement methods and positions this dissertation within the broader research landscape. By reviewing the key advancements and identifying the limitations of current approaches, it lays the groundwork for the contributions of this work.

The chapter begins by introducing the fundamental signal models in Section 2.1. Speech enhancement is traditionally performed in the time-frequency domain, where estimating the clean magnitude has been the primary focus. Section 2.2 reviews the statistical single-channel speech enhancement methods, presenting major milestones along with their limitations. With the rise of deep learning, machine learning-based methods have become increasingly prominent, which are briefly summarised in Section 2.3. In contrast to the extensive research on magnitude estimation, the phase spectrogram remains relatively under-explored. Section 2.4 reviews phase estimation in single-channel speech enhancement and phase retrieval approaches as a foundation for the proposed phase reconstruction method. Finally, Section 2.5 introduces self-supervised learning (SSL) models, an alternative way to implicitly encode prior knowledge derived from a large amount of unlabelled data. Its potential and usage in speech enhancement tasks are overviewed.

⋆ ⋆ ⋆

2.1 Signal Model

When a speech signal is captured by a single microphone remotely, it typically faces two types of interference: a) background noise, which comes from other sound sources and adds to the speech signal, and b) reverberation caused by sound reflecting off surrounding surfaces. This is modelled by a room impulse response (RIR) h_{RIR} to be convolved with the speech signal. As a result, the digital signal y captured by the microphone at time instance k can therefore be expressed as:

$$y(k) = h_{\mathrm{RIR}} * s(k) + v(k)\,, \tag{2.1}$$

where $*$ denotes convolution, $s(k)$ is the speech signal, and $v(k)$ the background noise. The goal of speech enhancement is to recover the clean speech s from the noisy and/or reverberant mixture y.

In practice, speech signals are usually first segmented, which is crucial to both real-time processing and accurate analysis of non-stationary characteristics. For this purpose, window functions that typically range from 20 to 40 milliseconds are employed to segment signals into short, overlapping frames. This segmentation allows for further analysis and processing on a frame-by-frame basis.

The short-time Fourier transform (STFT) is widely employed in speech signal processing given the sparsity of speech signals in the time-frequency domain. With the analysis window function $\mathrm{win}(i)$, a frame length of M and a frame shift of D, the STFT coefficients of signal y at frame l and frequency bin m is obtained by:

$$Y(l,m) = \sum_{i=0}^{M-1} \left(\mathrm{win}(i)\, y(lD+i)\, \exp^{-\jmath 2\pi \frac{im}{M}}\right). \tag{2.2}$$

In the STFT domain, accordingly, the signal model of Eq. (2.1) is approximated as:

$$Y(l,m) = H_{\mathrm{RIR}}(m)S(l,m) + V(l,m)\,, \tag{2.3}$$

where Y, S, V denote the STFT coefficients of the observed signal, clean speech, and background noise, respectively. H_{RIR} is the frequency response of the RIR. Consequently, the speech enhancement task translates into estimating S from Y.

Leveraging statistical properties of speech signals, STFT-domain methods are mainstream in speech enhancement, especially in conventional statistical approaches. They also constitute the majority of frameworks studied and optimised in this dissertation. After being processed, the estimated clean speech coefficients are converted back to the time domain by the inverse short-time Fourier transform (iSTFT) with overlap-add (OLA) to reconstruct enhanced signals.

2.1.1 Source-Filter Model

The speech signal s can be approximated using an auto-regressive (AR) model [1], where the current speech sample is predicted by a linear combination of previous

samples:

$$\widehat{s}(k) = \sum_{i=1}^{N} b_i s(k-i)\,. \tag{2.4}$$

This technique is called linear prediction coding (LPC), and the coefficients of this filter: $\mathbf{b}(l) = [b_1(l), b_2(l), \ldots, b_N(l)]$ are known as the LPC coefficients. By minimising the mean-square error (MSE) between the actual speech signal $s(k)$ and the predicted signal $\widehat{s}(k)$ for each frame, the LPC filter $\mathbf{b}(l)$ can be efficiently estimated from the autocorrelation of the speech signal using the Levinson-Durbin recursion [1].

Once $\mathbf{b}(l)$ is obtained, the speech signal can be expressed as:

$$\begin{aligned} s(k) &= \widehat{s}(k) + r(k) \\ &= \sum_{i=1}^{N} b_i s(k-i) + r(k)\,, \end{aligned} \tag{2.5}$$

where $r(k)$ is the prediction residual. In the z-domain, speech signal is then:

$$\begin{aligned} S(z) &= R(z) \cdot \frac{1}{1 - \sum_{i=1}^{N} b_i z^{-i}}, \\ &= R(z) \cdot \frac{1}{1 - B(z)}, \\ &= R(z) \cdot H(z)\,, \end{aligned} \tag{2.6}$$

Speech signals are stationary for short durations typically between 20 to 400 ms. Therefore, the filter $H(z)$ needs to be estimated frame by frame. With the STFT coefficients, Eq. (2.6) is approximated as:

$$S(l,m) = H(l,m) \cdot R(l,m)\,. \tag{2.7}$$

This simplified speech production model is termed the 'source-filter model'. The decomposition in Eq. (2.7) corresponds to the two components: the decomposition residual R mainly represents the excitation signal generated by the vibration of the vocal cords, whereas the envelope H approximates the influence of the vocal tract. For voiced sounds, the excitation signal is assumed periodic and has a fundamental frequency (F0) physically decided by the vocal tract; for unvoiced sounds, white noise is assumed. Thereby, the speech signal is viewed as the result of filtering the excitation signal R through the vocal tract filter H [1].

2.2 Statistical Methods

2.2.1 Framework

Statistical speech enhancement methods typically operate in the time-frequency domain. It should be noted that only additive noise is considered within this frame-

work, i.e., $y(k) = s(k) + v(k)$.

In the polar coordinate system, the STFT representation of one frame of the speech signal in the complex domain is expressed in two parts, the magnitude $A(l, m)$ and the phase $\Phi(l, m)$:

$$S(l, m) = A(l, m) \exp\{\jmath\, \Phi(l, m)\}\,. \tag{2.8}$$

As the magnitude spectrogram $A(l, m)$ of clean speech signals exhibits a well-defined structure, it is possible to estimate this component from noisy observations $Y(l, m)$. However, the phase $\Phi(l, m)$ lacks a discernible pattern, making it challenging to recover the clean phase. On the other hand, statistical analysis indicates that under the Gaussian assumption, the noisy phase is the optimal estimate of the clean phase in the minimum mean square error (MMSE) sense [2]. Experimental evidence further verifies that a noisy phase has much less impact on the quality of enhanced signals compared to a noisy magnitude [3]. Consequently, most statistical approaches focus on estimating the clean magnitude and subsequently combine the processed magnitude spectrogram with the observed noisy phase as the final estimate.

Figure 2.1 depicts the steps of a typical statistical speech enhancement approach. The captured mixture signal is first segmented and converted into the STFT domain with the pre-defined window function. The speech enhancement operation, estimating the clean speech magnitude spectrogram from the noisy mixture, can be formulated as applying a gain function $G(l, m)$ to $|Y(l, m)|$. Therefore, the gain function is a real-valued mask on a frame-by-frame basis. The estimated magnitude spectrogram is then combined with the noisy phase $\Phi_Y(l, m)$ to get the final STFT speech estimate $\widehat{S}(l, m)$, i.e.:

$$\begin{aligned} \widehat{S}(l, m) &= G(l, m) \cdot |Y(l, m)| \cdot \exp\{\jmath\, \Phi_Y(l, m)\} \\ &= G(l, m) \cdot Y(l, m)\,. \end{aligned} \tag{2.9}$$

Subsequently, the STFT-domain estimate is converted back to the time-domain signal using the synthesis window and OLA by iSTFT.

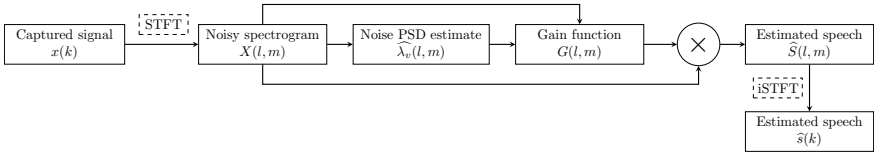


Figure 2.1: A general framework of a statistical STFT domain speech enhancement method. Solid boxes represent data contained, while dashed boxes indicate operations.

2.2.2 Gain Function

The gain function $G(l,m)$, a crucial parameter in this framework, controls the balance between noise suppression and speech preservation. Various methods exist to determine this parameter based on different assumptions about speech and noise distributions. Most approaches rely on two key parameters: the noise power spectral density (PSD) $\lambda_v(l,m)$, and the *a priori* signal-to-noise ratio (SNR) $\xi(l,m)$. Since the true values of these parameters are unknown in practice, they must be estimated before calculating the gain function.

When estimating noise PSD, it is usually assumed that the background noise varies *more slowly* than the foreground speech. This assumption is maintained in established noise PSD estimators such as the minimum statistics [4] and minimum mean square error-speech presence probability (MMSE-SPP) based approach [5]. The basic assumption of the slowly changing noise limits the aforementioned noise PSD estimation methods to primarily track the *stationary or quasi-stationary* components of the background noise. As a result, their performance deteriorates significantly in the presence of highly non-stationary noise, which consequently degrades the estimate of $G(l,m)$. Specifically, the underestimated noise leads to noise residual that distorts the enhanced signals, while noise overestimation leads to speech attenuation.

The *a priori* SNR is defined as:

$$\xi(l,m) = \frac{\lambda_s(l,m)}{\lambda_v(l,m)}\,, \tag{2.10}$$

where λ_s denotes the speech PSD. With the noise floor $\widehat{\lambda_v}(l,m)$ estimated, the naïve estimate of ξ with a lower boundary of $\xi_{\min}$ can be obtained in a maximum likelihood manner as [2]:

$$\widehat{\xi}(l,m) = \min\left\{\frac{|Y(l,m)|^2}{\widehat{\lambda_v}(l,m)} - 1, \xi_{\min}\right\}. \tag{2.11}$$

Once the noise floor and *a priori* SNR are estimated, the gain function can be calculated. For example, the spectral subtraction (SS) method with a minimal gain of $G_{\min}$ can be then written as:

$$G^{\mathrm{SS}}(l,m) = \left(\max\left\{1 - \frac{\lambda_v(l,m)}{|Y(l,m)|^2}, G_{\min}\right\}\right)^{\beta^{\mathrm{SS}}}\,, \tag{2.12}$$

which is an equivalent of the Wiener filter when the compression factor $\beta^{\mathrm{SS}} = 1/2$.

Assuming that speech spectral coefficients follow a Gaussian distribution, the gain function is further optimised by minimising the averaged MSE between the estimated magnitude spectra and the clean speech references, yielding the classical

minimum mean-square error short-time spectral amplitude (MMSE-STSA) estimator [2]. Building on this, the minimum mean-square error log-spectral amplitude (MMSE-LSA) speech estimator is derived by minimising a log-scale MSE cost function under the same Gaussian assumption to mirror the logarithmic perception of the human auditory system on loudness [6]:

$$G^{\mathrm{LSA}}(l,m) = \frac{\xi(l,m)}{1+\xi(l,m)} \exp\left(\frac{1}{2}\int_{\mu(l,m)}^{\infty} \frac{e^{-t}}{t}\, dt\right), \tag{2.13}$$

where $\gamma(l,m) = \frac{|Y(l,m)|^2}{\lambda_v(l,m)}$ is the *a posteriori* SNR, and $\mu(l,m) = \frac{\xi(l,m)}{1+\xi(l,m)}\gamma(l,m)$.

Further analysis of the speech spectral amplitude distribution suggests that super-Gaussian distributions, such as the Laplacian [7] or the Gamma [7] distribution, provide a better fit to the actual distribution compared to the commonly considered Gaussian model. These advanced distributions lead to alternative gain functions [8, 9]. In this dissertation, however, we employ the widely used MMSE-LSA gain function as the baseline, as it is well-studied and allows for direct and fair comparisons.

Estimation errors in the noise floor and *a priori* SNR are inevitable, often resulting in unpleasant musical noise artefacts in the enhanced speech. To address this issue, various methods have been proposed. One widely used approach is the decision-directed (DD) method [2], which applies recursive smoothing to the *a priori* SNR. By reducing abrupt temporal fluctuations in $\widehat{\xi}(l,m)$, this technique ensures that the enhanced speech signal sounds more pleasant:

$$\widehat{\xi}(l,m) = \max\left\{\alpha\frac{|\widehat{S}(l-1,m)|^2}{\widehat{\lambda_v}(l-1,m)} + (1-\alpha)\left(\frac{|Y(l,m)|^2}{\widehat{\lambda_v}(l,m)} - 1\right), \xi_{\min}\right\}, \tag{2.14}$$

where α is the weighting factor (typically set close to 1) to ensure a gradual adaptation to new observations.

2.2.3 Improved *A Priori* SNR Estimation

Nevertheless, there are still problems with the DD approach. Firstly, one frame delay is introduced by the recursive smoothing, which can be solved as in [10] by the two-step noise reduction (TSNR), where ξ is iteratively updated for each frame to eliminate the one-frame delay.

Another problem is that only temporal correlation is considered in the DD approach. For voiced frames, the clean speech spectral magnitude is structured with energy concentrated at F0 and integer multiples of it (speech harmonics). This relationship is not considered in the aforementioned method, and provides the room to further improve the speech estimate under the TSNR framework.

This spectral dependency can be highlighted by the excitation signal component of speech signals according to the source-filter model presented in Section 2.1.1. In [11], it is proposed to highlight the speech harmonics in the cepstrum of the excitation signal. Cepstrum is defined as applying the discrete cosine transform (DCT) to the logarithmic amplitude spectrum, making periodic structures more prominent. Thereby, F0 is identified as the peak within the typical range for human voices for each frame. To emphasise the harmonic structure of the voiced frames and suppress musical noise, the cepstral bin corresponding to F0 is amplified by a *constant* factor and the cepstral coefficients of the higher quefrencies are set to *zero*. This modified intermediate speech estimate is used to update the *a priori* SNR estimate, leading to an improved speech preservation.

The other component of the LPC decomposition, the speech envelope signal, is related to the uttered phoneme, and therefore to the intelligibility of speech [12]. Since the total number of unique phonemes is limited in one language [13], a spectral envelope codebook obtained from a clean speech corpus can be used to improve the speech intelligibility. The initial speech estimates are refined by replacing their spectral envelopes with the closest matches from the codebook, allowing for a more accurate *a priori* SNR estimation [14]. Distortions are removed if the correct template is found and substituted. Thereby, the envelope estimation problem is converted into a classification problem.

2.2.4 Contributions

Previous attempts such as [11, 14, 15] have shown that incorporating more domain knowledge into statistical speech enhancement frameworks benefits overall performance. In this direction, we have identified further opportunities for optimisation. Two avenues are explored here for a better integration of domain knowledge embodied by the source-filter model.

We start with the optimisation of the manipulation of the excitation signal cepstrum proposed in [11]. Two potential sources of performance degradation are identified in this baseline method. Acoustic details are discarded together with residual noise when directly nulling all cepstral coefficients unrelated to the F0 or the envelope. An important observation in [11] is that nulling cepstral coefficients outperforms replacing them by the best-matched excitation templates. This implies the importance of preserving acoustic details in an adaptive manner. In addition, it is suboptimal to use one pre-fixed constant amplifying factor for harmonic emphasis across all frames. Since the dynamic range of excitation signals varies widely from frame to frame, an input-dependent amplifying factor would further strengthen the speech harmonic boosting ability of this method.

These issues are addressed in Chapter 3. Primarily, two improvements are proposed:

- Cepstral smoothing (cepstral convolution): For adaptive acoustic detail preservation, the cepstrum is smoothed along the quefrency axis, which provides a balance between full nulling for musical noise reduction and the preservation of signal fine structures. This leads to more natural-sounding audio without the need for pre-trained templates.

- Dynamic amplifying factor (residual amplitude estimation): The harmonic amplification factor is determined frame by frame based on the range of the excitation signal, allowing for harmonic emphasis optimised for each frame.

Experimental results demonstrate that the two proposed modifications are complementary and collectively contribute to sharpening the speech harmonics in the enhanced audio, leading to higher speech quality and stronger noise suppression.

Next, this dissertation investigates the enhancement of spectral envelope. Despite the established relationship between speech envelope and intelligibility, it remains unclear why previous attempts such as [14, 15] failed to improve speech intelligibility (measured by short-time objective intelligibility (STOI)) with the codebook-based envelope enhancement. To answer this question, we analyse and optimise each component of the envelope enhancement system in the following aspects:

- Optimising codebook construction: By analysing the oracle distribution of codebook entries and experimenting with different coefficient choices, we optimise the codebook construction method.

- Introducing recurrent neural network (RNN): We explore the potential of shallow RNNs as an alternative to traditional methods such as Gaussian mixture model (GMM)-hidden Markov model (HMM) for codebook entry classification.

- Exploring the regression network performance: As an ablation study, we redefine the envelope enhancement as a regression task to assess the benefit of the codebook method and investigate the possible distortion caused by quantisation error. For this purpose, we replace the codebook-based classification system by the regression network.

Together, these optimisations yield measurable improvements in instrumental speech quality metrics; however, the overall performance gains remain limited, regardless of the choice for envelope estimation networks. The details of this research are presented in Appendix A.

These results highlight the limitation of this iterative statistical approach. The gain function relies on both the *a priori* SNR and the noise floor. Despite modifications aimed at improving *a priori* SNR estimation, the method remains unable to

recover the missing or highly distorted segments where the initial speech estimation fails. However, enhancing noise floor estimation is particularly challenging, as noise exhibits far greater variability than speech, while the modelling capacity of statistical methods remains limited. This inherent shortcoming of statistical approaches underscores the need to explore deep neural networks for better speech enhancement performance.

2.3 Deep Learning Methods

In the past decade, deep neural networks (DNNs) have proven to be a powerful tool in speech enhancement. Their capacity to model the non-linear relationship between distorted and clean speech, without requiring explicitly defined density functions, makes them well-suited for learning complicated distributions of clean speech signals—which is difficult to model using statistical approaches and therefore becomes a performance bottleneck.

During the training stage, a supervised DNN learns the mapping between the inputs and targets from paired data. This is achieved through an iterative update of network parameters to minimise the discrepancy between its predictions and actual *targets* measured by *loss functions*. At the inference stage, parameter updating stops and the trained DNN utilises the learnt mapping to predict the desired output for a given input. Leveraging this end-to-end training scheme, if appropriate training data are provided, where the input is a noisy, reverberant signal and the target is the clean, dry speech, noise suppression and reverberation removal can be accomplished by a single model.

To process non-stationary sequential data such as audio, the (overlapped) segmentation of signals is widely adopted in deep-learning-based speech enhancement methods as pre-processing. The exploration of deep learning for speech enhancement started with an analogy to statistical methods discussed in Section 2.2, which operates in the STFT domain with a focus on the clean magnitude estimation. Recently, however, studies have demonstrated that the strong learning ability of DNNs allows the possibility of complex target prediction [16–21] or a direct speech enhancement in the time domain [22–24] as well. Since speech is more structured in the STFT domain—allowing for the possibility of adding additional, interpretable information for fine-tuning the models, this overview still focuses mainly on the *STFT domain* DNN methods.

2.3.1 Training Targets

In the STFT domain, the two most common training targets for deep-learning-based speech enhancement are: a) the clean spectrogram, and b) a mask to be applied to the noisy mixture to obtain the clean spectrogram. Accordingly, these

two approaches are sometimes referred to as 'mapping' (directly predicting the clean speech spectrogram from the noisy input) and 'masking' (predicting a mask to remove interferences in the noisy spectrogram), respectively.

In the early stages of DNN-based speech enhancement, phase was often considered intractable and unimportant to the task, a view inherited from the established statistical methods [2, 6]. Therefore, most research focussed on estimating the clean magnitude spectrogram to be combined with the distorted phase of the observed mixture. As a result, training targets are limited to the real domain, i.e., the clean magnitude spectrogram [25] or the real-valued mask [26]. When comparing the two approaches, *masking methods* are generally reported to outperform mapping counterparts, especially in improving speech intelligibility. However, mapping methods exhibit a distinct advantage in low SNR scenarios [19, 25]. This is because masking approaches are more susceptible to numerical instability when mask values become excessively small, whereas the outputs of direct regression remain within a more confined range. Given the distinct strengths of each approach, there were also attempts to combine the two approaches to leverage their complementary benefits in different scenarios such as [27].

As the importance of phase in speech enhancement has been gradually recognised [28, 29], attention shifted towards the challenge of modelling complex-valued spectrograms. Predicting complex spectra presents unique challenges because phase values wrap around 2π with no clear pattern, and are sensitive to time shifts in the signal. These unique properties make defining an appropriate training target with phase more challenging than for magnitude alone. Studies have reported the advantages of using complex-valued targets to train DNNs for speech enhancement, confirmed by the improved scores of instrumental metrics and subjective listening tests. The complex mapping targets include magnitude accompanied with phase [20, 30], or real and imaginary parts [18, 21]. The complex masking targets include the phase sensitive mask (PSM) [16], the complex ideal ratio mask [17], and the real and imaginary component of the complex mask [31]. Given the diverse evidence presented, there is no definite conclusion on the optimal choice of training target for speech enhancement DNNs.

It should be noted that even trained with a complex-domain target, the networks sometimes struggle to learn and instead converge at a real-valued local minimum, effectively learning targets in the real domain [20]. A similar behaviour was also observed in our experiments with neural networks using complex-domain targets.

2.3.2 Loss Functions

The choice of loss function plays a critical role in the performance of DNNs as it guides the network training process. One simple way to quantitatively evaluate the performance of a network is to measure the distance (for example, by L1-distance

or MSE) between network outputs and training target ground truth. Given the wide dynamic range of speech spectra and the non-linearity of human perception of loudness, logarithm compression [32] or power-law compression [33] is often employed in loss functions to highlight low-energy or perceptually important regions.

Signal approximation loss measures performance in a slightly different manner. Since the ultimate goal of a speech enhancement system is to recover the underlying clean speech signal, the difference between the clean reference and the enhanced output serves as a reliable indicator of network performance, regardless of the chosen training target (mapping or masking). Following this perspective, loss functions remain unchanged for mapping networks, whereas a different loss function is defined for masking networks: instead of measuring the distance between the predicted and ideal masks, the signal approximation loss applies the output mask to the observed mixtures and compares the resulting STFT coefficients with the clean ones.

With training targets extended to the complex domain, errors are computed similarly to magnitude-based loss functions, leading to losses on the complex spectra. A systematic study in [33] indicates that a linear combination of the compressed magnitude loss and the compressed complex loss provides benefits in all tested scenarios:

$$\mathcal{L} = (1-\beta)\mathcal{L}_{\mathrm{mag}} + \beta\mathcal{L}_{\mathrm{cx}}\,, \tag{2.15}$$

where $\mathcal{L}_{\mathrm{mag}}$ denotes the magnitude loss and $\mathcal{L}_{\mathrm{cx}}$ the complex loss, and β is the hyper-parameter to balance the two terms. Grid search shows that for the compressed MSE loss, $\beta = 0.3$ provides the best combination between the real and complex terms.

While distance-based measurements are straightforward to implement, they do not directly correlate with perceptual quality, unlike instrumental evaluation metrics designed to mimic human perception. Therefore, researchers have explored the usage of perceptual speech quality metrics as loss functions, aiming to directly link loss values with perceived quality of enhanced signals. Common quality metrics repurposed as speech enhancement loss functions include perceptual evaluation of speech quality (PESQ) [34–39] and STOI [23, 40, 41]. However, the original PESQ implementation [42] is not fully differentiable, preventing its direct use as a loss function. To address this, various techniques have been proposed, such as differentiable implementations [34, 37] or neural network approximations [35, 36, 38, 39]. Using STOI as a loss function, on the other hand, has been proven to be equal to employing spectral MSE loss in theory and by experiments [41].

It is important to note that these metric-based loss functions are typically combined with distance-based terms, as there is a risk of overfitting when optimising solely for one single instrumental quality measure. For example, the ablation study in [37] shows that using only a faithful implementation of PESQ as the loss function can lead to a band-pass effect, dramatically improving PESQ scores but sig-

nificantly degrading speech quality. Using a neural network approximation cannot avoid such an overfitting issue either [39].

Although many attempts have been made to find better loss functions for speech enhancement, distance-based losses remain widely used and serve as the foundation for many other sophisticated losses due to their simplicity and effectiveness. In this dissertation, we follow this common practice optimised in [33] as the baseline to benchmark the proposed methods.

2.3.3 Architectures

DNNs are composed of different types of basic components, and each architectural component offers distinct capabilities in the context of speech enhancement. For example, convolutional layers focus on capturing local features, whereas the RNN is efficient in learning long-term dependency. Given their complementary strengths, hybrid architectures that combine these different layers often outperform models that rely solely on one type of architecture. Experimental results demonstrate that architectures incorporating temporal modelling perform better on the speech enhancement task, such as those using RNNs [16, 18, 20, 43, 44] and transformers [45, 46].

One of the most widely used architectures that has been proven successful for speech enhancement is the encoder-decoder framework (also known as UNet), which is frequently combined with recurrent layers at its bottleneck [18, 27, 31, 36, 43, 44], as illustrated in Figure 2.2. In this design, the encoder consists of convolutional layers, which extract and compress local features from the noisy input, while the decoder recovers the clean speech by reconstructing the target from the encoded features using transposed convolutional layers. Skip connections are commonly used between encoder and decoder layers to restore fine structures that may be lost during encoding. In addition, the recurrent layers such as long short-term memories (LSTMs) or gated recurrent units (GRUs) at the bottleneck capture the temporal dependencies across frames, enabling the model to track speech features over time. This combination allows the system to efficiently learn both the short-term and long-term properties of speech, which are crucial for robust performance in noisy environments.

One comprehensive optimisation of the efficiency of such an architecture can be found in [44]. Based on the convolutional recurrent neural network (CRNN) proposed in [43], several modifications have been made to further reduce its computational complexity. LSTMs at the bottleneck are replaced by more efficient grouped GRUs. Additionally, the traditional skip connection that concatenates two different feature maps is replaced by the add-skip connection. This change reduces the dimensionality of feature maps in the decoder, improving computational efficiency and lowering memory usage with minimal performance degradation. Furthermore,

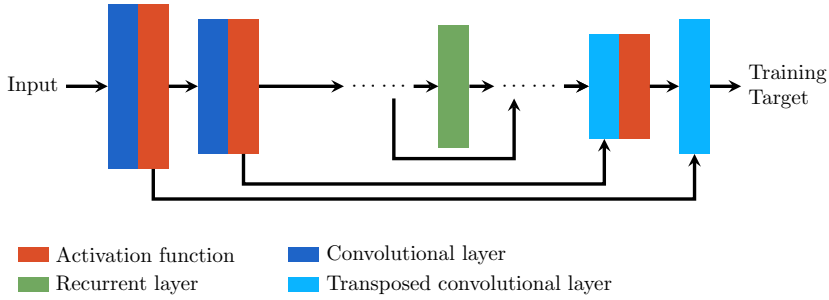


Figure 2.2: A typical UNet with skip connections and one recurrent layer for speech enhancement.

the influence of network depth has been investigated via a grid search to determine the optimal number of layers, balancing model complexity with performance. We employ this optimised network as the baseline system for future investigations and optimisations.

2.3.4 Contributions

When training a DNN, there is no need for explicit assumptions about the distributions of speech and noise spectral coefficients, as the network learns them during its end-to-end training. However, the distance-based loss functions discussed earlier tend to focus on minimising the global average error, without considering the structured nature of speech spectra. As noted in [47], this often leads to the attenuation of weak harmonics in the enhanced signal, especially in low SNR regions. While this may be a globally optimised result according to the loss function, it can lead to undesirable artefacts and distortions in the enhanced speech.

The clean speech magnitude spectrogram, as compactly described by the source-filter model, has a well-defined structure, but this domain knowledge is usually ignored in typical deep-learning-based speech enhancement methods. To incorporate the two, Chapter 4 proposes a novel loss function in the cepstral domain, termed cepstral excitation manipulation (CEM) loss. Building upon our previous exploration for better harmonic preservation based on the source-filter model in Chapter 3, this function is designed to measure the speech estimation errors at harmonics. To further address the common challenge of harmonic over-suppression in extreme conditions, we additionally introduce an amplifying factor to highlight the harmonics of the clean references.

Experiments show that this extra loss function enables the network to more effectively capture structured segments of speech from noisy mixtures, outperforming the naïve baseline by better speech harmonic preservation and stronger intra-harmonic attenuation, resulting in overall speech quality improvement. This

suggests that incorporating source-filter model principles as additional loss function terms effectively helps networks leverage domain knowledge of speech magnitude structure without increasing computational complexity during inference. Compared to other approaches that emphasises speech harmonic via loss functions such as [47], our CEM loss is not limited to the combination with the MSE loss function and can be integrated into a broad variety of frameworks.

2.4 Phase Reconstruction

Although the importance of phase is gradually recognised, its estimation remains challenging with the assistance of DNNs. In traditional statistical methods, the noisy phase is usually left intact. In deep learning methods, the introduction of complex training targets and complex loss functions has made it possible to estimate the phase together with the magnitude [16–21, 30]. However, it has been observed that, even with complex training targets, networks often make few changes to the noisy phase spectrogram [20, 48].

Further analysis indicates that predicting both the magnitude and the phase in a coupled manner potentially leads to a less accurate magnitude estimation [48]. This is because when the phase is intractable, instead of predicting a matched phase, a lower loss in the complex domain can be achieved by predicting the projection of the magnitude spectrogram on the noisy phase, which distorts the actual magnitude without improving phase estimation. In other words, when optimising a complex loss function, the magnitude estimate may compensate for phase estimate errors [48], which defeats the purpose of the complex-domain targets.

2.4.1 Phase Retrieval

Although clean speech phase does not exhibit a clear, predictable pattern, its relationship with magnitude has been widely explored, particularly in phase retrieval problems, which recover phase from a known clean magnitude.

One such relationship arises from the temporal redundancy in the STFT coefficients. Caused by the overlapped analysis window, this redundancy forms the basis of a classical iterative solution to the phase retrieval problem, the Griffin-Lim algorithm (GLA) [49], and inspires its DNN version, the deep Griffin-Lim algorithm [50]. However, its application in speech enhancement is limited for two reasons: a) GLA offers little improvement when the magnitude is distorted [49], and b) it is acausal because of the iterative STFT and iSTFT operation, which is problematic for real-time applications.

Another phase retrieval approach utilises phase gradients to find a matched phase for the clean magnitude spectrogram. A determined relationship between the *phase derivatives* and the log-amplitude can be derived when adopting a known

window (e.g., a Gaussian window) as the analysis window [51]. Non-iterative phase retrieval solutions such as phase gradient heap integration (PGHI) [51] are proposed based on this relationship. These methods first predict phase derivatives for the given magnitude, and then selectively accumulate the derivatives to obtain the final phase estimate. Since the phase gradients are related to the log-magnitude, they exhibit certain structure and are invariant to time-shifts, making them learnable by DNNs. This leads to the DNN-based phase retrieval methods using derivatives [52–54], where DNNs are trained to predict the phase derivatives from the magnitude spectra.

2.4.2 Contributions

While several efforts have been made on DNN-based phase estimation for speech enhancement, to the best of our knowledge, the relationship between log magnitudes and phase derivative has not been fully explored for this task yet. Speech enhancement surely benefits from a better phase estimation, but the phase retrieval solution cannot be directly used in speech enhancement, because the clean speech magnitude is not available.

To bridge this gap, Chapter 5 proposes a two-stage cascaded approach that separates magnitude enhancement from phase reconstruction. In this method, the clean magnitude spectrogram is first estimated using a typical speech enhancement network. Then, this estimated magnitude is treated as an approximation of the clean spectrogram to obtain the phase derivatives. The final estimate is constructed by fusing multiple results. By integrating phase reconstruction with common deep learning methods, our approach incorporates the phase consistency constraint with the general DNN-based speech enhancement process.

Compared to the conventional phase retrieval problem, phase reconstruction for speech enhancement presents two key challenges. First, the magnitude spectrogram estimated by a speech enhancement network is often imperfect, unlike the direct access to clean magnitude spectrogram in phase retrieval. These imperfections introduce distortions in phase derivative estimation, which then accumulate in the final reconstruction. Second, the initial phase estimate from speech enhancement typically sounds more natural than a fully synthetic phase, even in a real-valued system where the noisy phase serves as the initial estimate. Therefore, it is crucial to evaluate the impact of using an imperfect magnitude in gradient-based phase retrieval and to explore how leveraging the initial phase estimate can improve the final reconstruction.

Chapter 5 investigates how structured phase derivative can be incorporated into a speech enhancement system for phase reconstruction. A better phase is collectively obtained by combining the derivative-based retrieval results based on estimated magnitudes with the initial phase estimates. To study the potential ad-

verse effect of speech magnitude distortion, two versions of gradient estimation networks are trained and benchmarked: one operates with the ideal assumption that the magnitude estimate closely resembles the clean magnitude, i.e., this network learns the mapping from clean magnitudes to clean phase derivatives; while another accounts for magnitude distortion by a matched condition training, i.e., this network learns the mapping from estimated speech magnitudes to clean phase derivatives.

Experimental results show that the proposed phase reconstruction method is capable of improving the audio quality of network outputs trained for both real-valued and complex-valued targets, especially in the presence of stationary or semi-stationary noise. These results, particularly the improvement on signals enhanced by networks with complex-domain targets, suggest that the proposed phase reconstruction approach effectively introduces phase derivative information that has not been fully captured by the standard DNN training scheme into the speech enhancement system.

2.5 Self-Supervised Learning (SSL)

2.5.1 SSL Models

So far, we have focused on incorporating the human discovered domain knowledge into the speech enhancement system. Alternatively, self-supervised learning (SSL) models provide a fully DNN-based approach to encode prior knowledge through representation learning. These models aim to extract and separate multiple levels of speech-related information from audio and summarise it into representations in the latent space. One key advantage of the SSL method is its ability to leverage large amounts of data to learn underlying patterns without human annotation, effectively functioning as a deep-learning-based knowledge discovery system.

To train such an SSL model, it is required to design a pre-defined pseudo task whose ground truth references can be easily obtained without manual annotation. Common pseudo-tasks include contrastive learning [55], clustering-based label classification [56], or masked input recovery [57]. In terms of the architecture design, SSL models are typically divided into an encoder and a decoder. Once the network is fully trained for the pseudo task, the decoder is discarded, and the feature maps of the encoder are treated as the desired representations. These representations are assumed to capture essential information from the input signal, which can then be used as features for various downstream tasks.

These learned speech representations, often referred to as embeddings, have been shown to capture a wide range of information, including phonetic, semantic, and paralinguistic features. Numerous studies have demonstrated that embeddings extracted by SSL models generalise well across various tasks, such as keyword

spotting, emotion recognition, speaker identification, or automatic speech recognition (ASR). Notably, only a small amount of labelled data is required to train a shallow back-end model for a certain task based on SSL embeddings [58].

Given the richness of information extracted by SSL models from speech signals, their potential for speech enhancement is worth exploring. SSL models pretrained on large corpora can capture clean speech properties and analyse speaker and content information. By incorporating them into the speech enhancement system, we could introduce a broader range of prior information that extends beyond the domain knowledge explicitly considered in the previous chapters, making SSL models a valuable addition.

One approach to use SSL models for speech enhancement is to leverage them as feature extractors of loss functions [59, 60]. For example, the correlation analysis in [60] demonstrates that the distance between the embedding extracted from the enhanced signal and the one from the clean reference correlates well with perceptual speech intelligibility scores.

Alternatively, since SSL embeddings capture essential acoustic information, speech (or enhancement masks) can be directly synthesised from the extracted embeddings [61–63]. For this purpose, the ideal SSL model should be robust against background noise and reverberation.

A crucial design choice in SSL-based speech enhancement is selecting a pretrained SSL model from the many publicly available options. Current research often relies on empirical selection such as exhaustive search [62]. There lacks a quantitative analysis of embedding robustness across different SSL models, presenting an area for further investigation.

2.5.2 Neural Vocoder

To syntheses speech audio from SSL embeddings, a neural vocoder is required. Based on the generative adversarial network (GAN) training scheme, HiFi-GAN [64], a widely used architecture for audio generation, is commonly employed for this purpose.

HiFi-GAN learns to synthesise speech in an adversarial manner. The GAN consists of two parts: the generator ($\mathbb{G}$) which synthesises speech signal, and the discriminator ($\mathbb{D}$) which classifies its input signal as real (from real data) or fake (synthesised by $\mathbb{G}$). The adversarial loss of GAN is defined as follows:

$$\mathcal{L}_{\text{GAN}} = -E\left\{(\mathbb{D}(\mathbb{G}(y)))\right\} - E\left\{(1 - \mathbb{D}(s))\right\} . \tag{2.16}$$

The discriminator aims to *maximise* this loss to identify the generated data and thereby capture the real data distribution, while the generator attempts to *minimise* the loss to generate an authentic output that can 'fool' the discriminator. Through this min-max game during training, the generator learns the real data distribution,

and only $\mathbb{G}$ is needed at the inference stage. In speech enhancement tasks, where noisy observations are available, conditional GANs (cGANs) are typically more appropriate, as they can condition the generation process on input signals to obtain the target underlying clean speech.

HiFi-GAN [64] was originally proposed for the text-to-speech (TTS) task, mapping mel-spectrograms to authentic waveforms. Its architecture includes a generator and two discriminators: the multi-period discriminator (MPD) and the multi-scale discriminator (MSD). The generator consists of residual blocks at various scales to convert input features, while the two discriminators independently aggregate the convolutional neural network (CNN) outputs to make their final decisions. Accordingly, each component is optimised by its own loss function. Following the standard binary classifier training scheme, the discriminators are updated using the binary cross-entropy (BCE) loss on their classification results of reference and generated signals, whereas the generator is updated with a weighted summation of three loss terms: a) the log-Mel distance between the generated (enhanced) signal and the target, b) the classification results of the discriminators, and c) the feature matching loss, which measures the MSE between the discriminator feature maps when taking generator outputs and these when taking reference clean signals. This architecture is later adapted as the general neural vocoder solution to map speech representations in other domains (such as SSL embeddings) to the time-domain speech signal [62, 65].

As identified in speech synthesis tasks [65], one challenge for SSL-embedding-based neural vocoders is that the deeper layers of SSL models may lack sufficient acoustic details, and thus yield less natural output. It is proposed in [65] to address the problem by empirically supplementing the missing information with corresponding feature extraction modules such as the F0 tracker and the speaker embedding extractor. Nevertheless, this approach is inefficient as a general solution to benefit other neural vocoders since it involves a trial-and-error process. Moreover, the manual identification and extraction of missing information may not be easy. In contrast, a different strategy is adapted in [62]: feature maps (also termed 'hidden states') of each layer in the SSL model are combined together with a learnt weight as the input to the HiFi-GAN-based neural vocoder.

A close examination of the learnt weights in [62] shows that for all investigated SSL models, the first hidden state where least compression is performed always contributes most, suggesting that acoustic details are essential to the speech enhancement task. The canonical correlation analysis (CCA) of various SSL models verifies that the first layer is closely related to local features, whereas phonetic or linguistic information is more prominent in the deeper layers [66]. This analysis also indicates that relying heavily on the first hidden state might underutilise the full potential of SSL models, which are designed to capture both local and high-level speech information. If SSL models are utilised in speech enhancement

in a more efficient way, the overall performance might be improved by the full exploration of the automatically encoded prior knowledge.

2.5.3 Contributions

We identify several questions to be answered for efficient applications of SSL models with more and more of them available. While SSL models offer a promising approach to leveraging vast amounts of data, a critical issue often overlooked in research is the discrepancy between clean training audio and the highly distorted recordings encountered in real-world scenarios. This mismatch poses a potential challenge to the effective application of SSL models in downstream tasks such as speech enhancement. Furthermore, although SSL embeddings undoubtedly encapsulate essential information, it remains unclear how easily the task-relevant information can be retrieved from these embeddings. Gaining insights into these questions could benefit early-stage model selection, enhancing overall development efficiency.

To quantitatively answer these questions and facilitate the use of SSL models for speech enhancement, Chapter 6 proposes a framework for evaluating embedding robustness and analysing preserved information in the extracted embeddings without the need to train task-specific networks. This analysis helps us understand how SSL models perform under real-world conditions and provides a foundation for their application in speech enhancement. Our analysis on HuBERT, WavLM, TERA, and Wav2Vec2.0 shows that TERA demonstrates the highest robustness against interference among the evaluated models, making it a strong candidate for the speech re-synthesis enhancement framework.

Existing SSL-based neural vocoders for speech enhancement [62, 63] demonstrate the potential of the pre-trained large models on the task; however, determining the optimal supplementary information to improve the neural vocoder performance as proposed in [65] remains a challenge. In Chapter 7, we tackle it by the integration of noisy spectrograms alongside SSL embeddings in neural vocoders for speech synthesis. As it is costly and challenging to manually identify and extract the missing acoustic information accordingly, we propose to provide neural vocoders with direct access to noisy spectrograms, enabling them to learn extracting acoustic details missing in the embedding from noisy spectrograms. Besides, this direct access could potentially help the neural vocoder to focus on the high-level information summarised by SSL models in the embeddings when the local features are more easily accessible via the spectrograms. Since two input modalities are provided to the neural vocoder, we explore different fusion strategies to further optimise the enhancement performance. Experimental results show that compared to the baseline denoising vocoder which takes only SSL embeddings [62], the introduction of the noisy spectrogram improves the audio quality of the speech

signals synthesised by the SSL-based speech enhancement system, particularly in terms of the naturalness.

2.6 Publications

2.6.1 Publications in international journals (listed in the Science Citation Index [1])

1. **Song, Y.**, and Madhu, N. (2022). Improved CEM for Speech Harmonic Enhancement in Single Channel Noise Suppression. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30*, 2492-2503.

2. **Song, Y.**, Van Hoecke, E., and Madhu, N. (2022). Portable and Non-Intrusive Fill-State Detection for Liquid-Freight Containers Based on Vibration Signals. *Sensors, 22*(20), 7901.

3. **Song, Y.**, and Madhu, N. (2023). Investigations on the Optimal Estimation of Speech Envelopes for the Two-Stage Speech Enhancement. *Sensors, 23*(14), 6438.

4. Song, S., **Song, Y.**, and Madhu, N. (2024). Robust Detection of Background Acoustic Scene in the Presence of Foreground Speech. *Applied Sciences, 14*(2), 609.

5. Baert, M., Moons, B., Pittevils, J., **Song, Y.**, Madhu, N., and Hoebeke, J. (2024). Evaluation of BLE-Based Audio Broadcasting Under Probabilistic Interference. *Computer Communications, 222*, 130-140.

2.6.2 Publications in International Conferences

1. **Song, Y.**, Kindt, S., and Madhu, N. (2022). Drone Ego-Noise Cancellation for Improved Speech Capture Using Deep Convolutional Autoencoder Assisted Multistage Beamforming. In *2022 25th International Conference on Information Fusion (FUSION)* (pp. 1-8). IEEE.

2. **Song, Y.**, and Madhu, N. (2023). Aiding Speech Harmonic Recovery in DNN-Based Single Channel Noise Reduction Using Cepstral Excitation Manipulation (CEM) Components. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1-5). IEEE.

[1] The publications listed are recognised as 'A1 publications', according to the following definition used by Ghent University: A1 publications are articles listed in the Science Citation Index Expanded, the Social Science Citation Index or the Arts and Humanities Citation Index of the ISI Web of Science, restricted to contributions listed as article, review, letter, note or proceedings paper.

3. **Song, Y.**, Kim, D., Madhu, N., and Kang, H.-G. (2024). On the Disentanglement and Robustness of Self-Supervised Speech Representations. In *2024 International Conference on Electronics, Information, and Communication (ICEIC)* (pp. 1-4). IEEE.

4. **Song, Y.**, and Madhu, N. (2024). Phase Reconstruction in Single Channel Speech Enhancement Based on Phase Gradients and Estimated Clean-Speech Amplitudes. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1461-1465). IEEE.

5. **Song, Y.**, Kim, D., Kang, H.-G., and Madhu, N. (2024). Spectrum-Aware Neural Vocoder Based on Self-Supervised Learning for Speech Enhancement. In *2024 32nd European Signal Processing Conference (EUSIPCO)* (pp. 16-20). IEEE.

6. Kim, D., **Song, Y.**, Madhu, N., and Kang, H.-G. (2024). Enhancing Neural Speech Embeddings for Generative Speech Models. In *2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (pp. 1-6). IEEE.

7. Grzywalski, T., Botteldooren, D., **Song, Y.**, and Madhu, N. (2024). Salient Sound Extraction Using Deep Neural Networks Predicting Complex Masks. In *2024 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)* (pp. 166-171). IEEE.

References

[1] P. Vary and R. Martin. *Digital speech transmission: Enhancement, coding and error concealment*. John Wiley & Sons, 2006.

[2] Y. Ephraim and D. Malah. *Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator*. IEEE Transactions on acoustics, speech, and signal processing, 32(6):1109–1121, 1984.

[3] D. Wang and J. Lim. *The unimportance of phase in speech enhancement*. IEEE Transactions on Acoustics, Speech, and Signal Processing, 30(4):679–681, 1982.

[4] R. Martin. *Noise power spectral density estimation based on optimal smoothing and minimum statistics*. IEEE Transactions on speech and audio processing, 9(5):504–512, 2001.

[5] T. Gerkmann and R. C. Hendriks. *Unbiased MMSE-based noise power estimation with low complexity and low tracking delay*. IEEE Transactions on Audio, Speech, and Language Processing, 20(4):1383–1393, 2011.

[6] Y. Ephraim and D. Malah. *Speech enhancement using a minimum mean-square error log-spectral amplitude estimator*. IEEE transactions on acoustics, speech, and signal processing, 33(2):443–445, 1985.

[7] R. Martin and C. Breithaupt. *Speech enhancement in the DFT domain using Laplacian speech priors*. In Proc. IWAENC, volume 3, pages 87–90. Citeseer, 2003.

[8] R. Martin. *Speech enhancement based on minimum mean-square error estimation and supergaussian priors*. IEEE transactions on speech and audio processing, 13(5):845–856, 2005.

[9] C. Breithaupt, M. Krawczyk, and R. Martin. *Parameterized MMSE spectral magnitude estimation for the enhancement of noisy speech*. In 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, pages 4037–4040. IEEE, 2008.

[10] C. Plapous, C. Marro, and P. Scalart. *Improved signal-to-noise ratio estimation for speech enhancement*. IEEE transactions on audio, speech, and language processing, 14(6):2098–2108, 2006.

[11] S. Elshamy, N. Madhu, W. Tirry, and T. Fingscheidt. *Instantaneous a priori SNR estimation by cepstral excitation manipulation*. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(8):1592–1605, 2017.

[12] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen. *An algorithm for intelligibility prediction of time–frequency weighted noisy speech*. IEEE Transactions on audio, speech, and language processing, 19(7):2125–2136, 2011.

[13] W. F. Twaddell. *On defining the phoneme*. Language, 11(1):5–62, 1935.

[14] S. Elshamy, N. Madhu, W. Tirry, and T. Fingscheidt. *DNN-supported speech enhancement with cepstral estimation of both excitation and envelope*. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(12):2460–2474, 2018.

[15] S. Eishamy, N. Madhu, W. Tirry, and T. Fingscheidt. *A priori SNR computation for speech enhancement based on cepstral envelope estimation*. In 2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC), pages 351–355. IEEE, 2018.

[16] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux. *Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks*. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 708–712. IEEE, 2015.

[17] D. S. Williamson, Y. Wang, and D. Wang. *Complex ratio masking for monaural speech separation*. IEEE/ACM transactions on audio, speech, and language processing, 24(3):483–492, 2015.

[18] K. Tan and D. Wang. *Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement*. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6865–6869. IEEE, 2019.

[19] Z.-Q. Wang, P. Wang, and D. Wang. *Complex spectral mapping for single- and multi-channel speech enhancement and robust ASR*. IEEE/ACM transactions on audio, speech, and language processing, 28:1778–1787, 2020.

[20] D. Yin, C. Luo, Z. Xiong, and W. Zeng. *Phasen: A phase-and-harmonics-aware speech enhancement network*. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 9458–9465, 2020.

[21] A. Li, C. Zheng, G. Yu, J. Cai, and X. Li. *Filtering and refining: A collaborative-style framework for single-channel speech enhancement*. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30:2156–2172, 2022.

[22] A. Pandey and D. Wang. *TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain*. In ICASSP 2019-2019

IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6875–6879. IEEE, 2019.

[23] M. Kolbæk, Z.-H. Tan, S. H. Jensen, and J. Jensen. *On loss functions for supervised monaural time-domain speech enhancement*. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28:825–838, 2020.

[24] K. Wang, B. He, and W.-P. Zhu. *TSTNN: Two-stage transformer based neural network for speech enhancement in the time domain*. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 7098–7102. IEEE, 2021.

[25] S. A. Nossier, J. Wall, M. Moniri, C. Glackin, and N. Cannings. *Mapping and masking targets comparison using different deep learning based speech enhancement architectures*. In 2020 international joint conference on neural networks (IJCNN), pages 1–8. IEEE, 2020.

[26] Y. Wang, A. Narayanan, and D. Wang. *On training targets for supervised speech separation*. IEEE/ACM transactions on audio, speech, and language processing, 22(12):1849–1858, 2014.

[27] M. Strake, B. Defraene, K. Fluyt, W. Tirry, and T. Fingscheidt. *Separated noise suppression and speech restoration: LSTM-based speech enhancement in two stages*. In 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pages 239–243. IEEE, 2019.

[28] K. Paliwal, K. Wójcicki, and B. Shannon. *The importance of phase in speech enhancement*. speech communication, 53(4):465–494, 2011.

[29] T. Gerkmann, M. Krawczyk, and R. Rehr. *Phase estimation in speech enhancement—unimportant, important, or impossible?* In 2012 IEEE 27th Convention of Electrical and Electronics Engineers in Israel, pages 1–5. IEEE, 2012.

[30] T. Peer and T. Gerkmann. *Phase-aware deep speech enhancement: It's all about the frame length*. JASA Express Letters, 2(10), 2022.

[31] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie. *DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement*. arXiv preprint arXiv:2008.00264, 2020.

[32] J. Du, Y. Tu, L.-R. Dai, and C.-H. Lee. *A regression approach to single-channel speech separation via high-resolution deep neural networks*. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24(8):1424–1437, 2016.

[33] S. Braun and I. Tashev. *A consolidated view of loss functions for supervised deep learning-based speech enhancement*. In 2021 44th International Conference on Telecommunications and Signal Processing (TSP), pages 72–76. IEEE, 2021.

[34] J. Kim, M. El-Khamy, and J. Lee. *End-to-end multi-task denoising for joint SDR and PESQ optimization*. arXiv preprint arXiv:1901.09146, 2019.

[35] S.-W. Fu, C.-F. Liao, and Y. Tsao. *Learning with learned loss function: Speech enhancement with quality-net to improve perceptual evaluation of speech quality*. IEEE Signal Processing Letters, 27:26–30, 2019.

[36] Z. Xu, M. Strake, and T. Fingscheidt. *Deep noise suppression maximizing non-differentiable PESQ mediated by a non-intrusive PESQNet*. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30:1572–1585, 2022.

[37] L. Thieling, L. Nippert, and P. Jax. *Using Perceptual Evaluation of Speech Quality (PESQ) Loss for DNN-Based Speech Enhancement*. In Speech Communication; 15th ITG Conference, pages 61–65. VDE, 2023.

[38] Z. Xu, M. Sach, J. Pirklbauer, and T. Fingscheidt. *Employing Real Training Data for Deep Noise Suppression*. In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 10731–10735. IEEE, 2024.

[39] G. Close, T. Hain, and S. Goetze. *Hallucination in Perceptual Metric-Driven Speech Enhancement Networks*. In 2024 31st European Signal Processing Conference (EUSIPCO), pages 21–25. IEEE, 2024.

[40] S. Venkataramani, R. Higa, and P. Smaragdis. *Performance based cost functions for end-to-end speech separation*. In 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pages 350–355. IEEE, 2018.

[41] M. Kolbaek, Z.-H. Tan, and J. Jensen. *On the relationship between short-time objective intelligibility and short-time spectral-amplitude mean-square error for speech enhancement*. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27(2):283–295, 2018.

[42] ITU-T. *Rec. P.862.2: Corrigendum 1, Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs*. International Telecommunication Union-Telecommunication Standardisation Sector, 2017.

[43] K. Tan and D. Wang. *A convolutional recurrent neural network for real-time speech enhancement*. In Interspeech, volume 2018, pages 3229–3233, 2018.

[44] S. Braun, H. Gamper, C. K. Reddy, and I. Tashev. *Towards efficient models for real-time deep noise suppression*. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 656–660. IEEE, 2021.

[45] J. Kim, M. El-Khamy, and J. Lee. *T-gsa: Transformer with gaussian-weighted self-attention for speech enhancement*. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6649–6653. IEEE, 2020.

[46] F. Dang, H. Chen, and P. Zhang. *DPT-FSNet: Dual-path transformer based full-band and sub-band fusion network for speech enhancement*. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6857–6861. IEEE, 2022.

[47] N. Raviv, O. Schwartz, and S. Gannot. *Low resources online single-microphone speech enhancement with harmonic emphasis*. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 8807–8811. IEEE, 2022.

[48] Z.-Q. Wang, G. Wichern, and J. Le Roux. *On the compensation between magnitude and phase in speech separation*. IEEE Signal Processing Letters, 28:2018–2022, 2021.

[49] D. Griffin and J. Lim. *Signal estimation from modified short-time Fourier transform*. IEEE Transactions on acoustics, speech, and signal processing, 32(2):236–243, 1984.

[50] Y. Masuyama, K. Yatabe, Y. Koizumi, Y. Oikawa, and N. Harada. *Deep griffin–lim iteration*. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 61–65. IEEE, 2019.

[51] Z. Průša, P. Balazs, and P. L. Søndergaard. *A noniterative method for reconstruction of phase from STFT magnitude*. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(5):1154–1164, 2017.

[52] Y. Masuyama, K. Yatabe, Y. Koizumi, Y. Oikawa, and N. Harada. *Phase reconstruction based on recurrent phase unwrapping with deep neural networks*. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 826–830. IEEE, 2020.

[53] L. Thieling, D. Wilhelm, and P. Jax. *Recurrent phase reconstruction using estimated phase derivatives from deep neural networks*. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 7088–7092. IEEE, 2021.

[54] Y. Masuyama, K. Yatabe, K. Nagatomo, and Y. Oikawa. *Online phase reconstruction via DNN-based phase differences estimation*. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31:163–176, 2022.

[55] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli. *wav2vec 2.0: A framework for self-supervised learning of speech representations*. Advances in neural information processing systems, 33:12449–12460, 2020.

[56] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed. *Hubert: Self-supervised speech representation learning by masked prediction of hidden units*. IEEE/ACM transactions on audio, speech, and language processing, 29:3451–3460, 2021.

[57] A. T. Liu, S.-W. Li, and H.-y. Lee. *Tera: Self-supervised learning of transformer encoder representation for speech*. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29:2351–2366, 2021.

[58] S. wen Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K. tik Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H. yi Lee. *SUPERB: Speech Processing Universal PERformance Benchmark*. In Proc. Interspeech 2021, pages 1194–1198, 2021. doi:10.21437/Interspeech.2021-1775.

[59] H. Sato, R. Masumura, T. Ochiai, M. Delcroix, T. Moriya, T. Ashihara, K. Shinayama, S. Mizuno, M. Ihori, T. Tanaka, et al. *Downstream task agnostic speech enhancement with self-supervised representation loss*. In Proc. Interspeech 2023, pages 854–858, 2023.

[60] R. Sutherland, G. Close, T. Hain, S. Goetze, and J. Barker. *Using Speech Foundational Models in Loss Functions for Hearing Aid Speech Enhancement*. In 2024 31st European Signal Processing Conference (EUSIPCO), pages 421–425. IEEE, 2024.

[61] Z. Huang, S. Watanabe, S.-w. Yang, P. García, and S. Khudanpur. *Investigating self-supervised learning for speech enhancement and separation*. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6837–6841. IEEE, 2022.

[62] B. Irvin, M. Stamenovic, M. Kegler, and L.-C. Yang. *Self-supervised learning for speech enhancement through synthesis*. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1–5. IEEE, 2023.

[63] Z. Wang, X. Zhu, Z. Zhang, Y. Lv, N. Jiang, G. Zhao, and L. Xie. *SELM: Speech enhancement using discrete tokens and language models*. In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 11561–11565. IEEE, 2024.

[64] J. Kong, J. Kim, and J. Bae. *Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis*. Advances in neural information processing systems, 33:17022–17033, 2020.

[65] A. Polyak, Y. Adi, J. Copet, E. Kharitonov, K. Lakhotia, W.-N. Hsu, A. Mohamed, and E. Dupoux. *Speech resynthesis from discrete disentangled self-supervised representations*. In Proc. Interspeech 2021, pages 3615–3619, 2021.

[66] A. Pasad, B. Shi, and K. Livescu. *Comparative layer-wise analysis of self-supervised speech models*. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1–5. IEEE, 2023.

3

Improved CEM for Speech Harmonic Enhancement in Single Channel Noise Suppression

Classical speech enhancement methods estimate underlying clean speech signals with the Gaussian distribution assumptions of speech and noise. This estimate can be further improved by iteratively refining it in a two-stage speech enhancement framework. According to the source-filter model, which is independent of the speech distribution assumption, speech signals can be decomposed into excitation signals and envelope signals, where speech harmonics are captured in the excitation signals and characterised by a few bins in the cepstrum. Therefore, by boosting these harmonic-related bins of initial speech estimate, cepstral excitation manipulation (CEM) has been proposed to enhance speech harmonics.

In this chapter, we investigate this source-filter-model based two-stage method and propose two modifications after a close examination of the enhanced speech signals. One challenge is to find an appropriate harmonic amplification factor to obtain sufficient harmonic enhancement while keeping the naturalness of the enhanced speech. We propose to employ an adaptive amplification factor for each frame, calculated by residual amplitude estimation (RAE). Furthermore, instead of discarding cepstral coefficients corresponding to the fine structure, we propose to preserve them by cepstral convolution (CC). In such a manner, the parameters of CEM are dependent on input signal statistics. The ablation studies show that the two modifications are complementary and yield adaptive method that highlights

the speech harmonics.

⋆ ⋆ ⋆

Y. Song, and N. Madhu.

Published in IEEE/ACM Transactions on Audio Speech and Language Processing, July 2022.

Abstract The periodic nature of voiced speech is often exploited to restore speech harmonics and to increase inter-harmonic noise suppression. In particular, a recent paper proposed to do this by manipulating the speech harmonic frequencies in the cepstral domain. The manipulations were carried out on the cepstrum of the excitation signal, obtained by the source-filter decomposition of speech. This method was termed Cepstral Excitation Manipulation (CEM). In this contribution we further analyse this method, point out its inherent weakness and propose means to overcome it. First of all, it will be shown by both illustrative examples and theoretical analysis that the existing method underestimates the excitation, especially at low signal-to-noise ratio (SNR) conditions. This inherent weakness leads to speech harmonic weakening and vocoding due to the insufficient noise suppression in the inter-harmonic regions. Then, we propose two modifications to improve the robustness and performance of CEM in low SNR cases. The first modification is to use an instantaneous amplifying factor adapted to the signal, instead of a pre-defined constant, for the excitation cepstrum. The second modification is to smooth the excitation cepstrum to preserve additional fine structure, instead of discarding it. These modifications result in better preservation of speech harmonics, more refined fine structure and higher inter-harmonic noise suppression. Experimental evaluations using a range of standard instrumental metrics conclusively demonstrate that our proposed modifications clearly outperform the existing method, especially in extremely noisy conditions.

3.1 Introduction

Single channel speech enhancement is widely used in several communications applications to suppress background noise and improve speech quality. Many efforts have been made in the topic and this is still an active field of research. The majority of statistical speech enhancement algorithms apply a gain function to the short time-frequency representation of speech. Based on the estimation of the power spectral density of the noise and the underlying speech, the so-called *a priori* SNR and *a posteriori* SNR are, typically, the two crucial parameters required for the gain function calculation.

Under individual independent Gaussian assumptions of noise and speech spectral coefficients, Ephraim and Malah derived the classical minimum mean-square error short-time spectral amplitude (MMSE-STSA) estimator [1] and minimum mean-square error log-spectral amplitude (MMSE-LSA) estimator [2], where the *a priori* SNR is estimated by the decision-directed (DD) approach. One well-known drawback of minimum mean square error (MMSE) estimator is musical noise in the enhanced audio. The recursive smoothing of the DD approach tackles it to some degree, but also leads to speech distortions. This drawback has been handled in various ways in the literature. For example, [3] proposed the two-step noise reduction (TSNR) technique to not only solve the one-frame delay introduced by the bias of *a priori* SNR estimation but also combined it with *harmonic regeneration noise reduction* (HRNR) to restore the distorted harmonics by applying a non-linearity in the time domain. Cepstro-temporal smoothing has been proposed in [4] to reduce the musical tones. It essentially applies a post-filter to the MMSE gain estimate in the cepstrum domain. The recursive smoothing of high-order cepstral coefficients of the gain function reduces abrupt changes in its fine structure, thereby lowering musical noise. This technique can also be applied to the cepstral coefficients of the enhanced spectrum directly. For instance, a selective cepstro-temporal smoothing scheme was proposed in [5]. Typically, these approaches target a better enhancement of the speech harmonics in voiced speech regions.

On the other hand, the speech production process can be abstracted as a source-filter model that captures the harmonic structure in speech with low mathematical complexity [6]. Thereby, the speech signal is decomposed into an envelope and an excitation signal. There have been analysis-synthesis speech enhancement frameworks based on this decomposition, leading to learning-based envelope enhancement solutions, e.g., [7–9]. In addition to the envelope enhancement, a synthetic excitation signal is also introduced in [8, 9], to address the musical noise problem. But it has also been reported that synthetic speech lacks naturalness. For this reason, and since the source-filter decomposition is sensitive to noise, the approaches are usually combined with the aforementioned spectral amplitude estimators, resulting in a TSNR framework.

More recently, there have been attempts to improve speech quality by directly manipulating the excitation signal. In [10, 11], the cepstral representation of the excitation signal is adopted to highlight the periodic structure of speech. Using this representation, it is possible to get a clearer speech estimation and even to restore low-amplitude or lost harmonics by amplifying the periodic structure of voiced speech in the cepstral domain. Consequently, this approach is termed cepstral excitation manipulation (CEM). A benchmark of CEM against the relevant state-of-the-art in [10] indicated the potential of this approach to further improve the enhanced speech quality. This method was, further, combined with envelope

estimation methods in [11], thereby integrating the benefits of envelope and excitation improvement for speech enhancement. This demonstrated that in addition to being used as a stand-alone component, CEM could also be piggy-backed onto other speech enhancement frameworks – which makes this idea versatile.

In [10], the excitation of speech is replaced by either an idealised, synthetic excitation or by selecting one from pre-trained cepstral excitation templates. The non-template method has the advantage that it is capable of recovering lost harmonics in a relatively simple manner: all but two of the cepstral coefficients of the excitation signal are set to zero (the process termed 'cepstral nulling'). The zeroth coefficient, corresponding to the energy term, is preserved, along with the cepstral coefficient that has maximum amplitude within the range of allowed speech fundamental frequencies. This latter coefficient is further amplified by a pre-determined constant, thereby implicitly emphasising the spectral peaks at the fundamental frequency and its harmonics. However, as discussed in [10], this method shares the disadvantage of the lack of naturalness in the enhanced audio with the analysis-synthesis framework. This is due to the fact that most of the cepstrum has been discarded, which means the loss of speech fine structure. The amendment provided by [10] is to introduce cepstral excitation templates in either a speaker-dependent or a speaker-independent manner. However, this solution adds to the complexity of the system because of the template training and extra models required. Another problem of this method comes from the constant pitch amplifying factor. When the harmonics are corrupted by noise, the dynamic range of the boosted speech excitation signal is insufficient with a pre-determined amplifying factor. In such cases, the noise at the inter-harmonic frequencies is poorly suppressed, which results in a vocoder-like effect. Thus, classical CEM fails to maintain the sharpness of harmonics in the enhanced output signals, especially in low SNR conditions.

In this contribution, we focus on the above shortcomings of CEM and propose means to address them. First, instead of using a fixed pitch amplifying factor, the dynamic range of the synthetic excitation is instantaneously and adaptively estimated based on the input signal spectrum. Second, the excitation signal cepstrum is selectively averaged along the quefrency, which preserves the spectral fine structure better and results in a more natural estimate of the underlying speech spectrum. This can serve as an alternative to speaker excitation templates trained on large corpora as in [10]. Lastly, these two improvements can be combined to yield a robust, improved CEM approach with better performance in all SNR conditions.

The paper is structured as follows: the baseline CEM method, together with the speech enhancement framework, is introduced in Section 3.2. Representative examples are first shown in Section 3.3 to demonstrate the weakness of classical CEM in excitation synthesis and its consequence. Our modifications aimed at addressing these shortcomings are next presented. The proposed methods are thoroughly evaluated in Section 3.4 and the conclusions are presented in Section 3.5.

3.2 Cepstral excitation manipulation (CEM) baseline

We consider an additive mixture of an underlying speech signal $s(n)$ mixed with the background noise signal $v(n)$. The goal of speech enhancement is to get a clean speech estimate $\widetilde{s}(n)$ of better quality and/or intelligibility given the noisy observation $y(n) = s(n) + v(n)$. With an M-point windowed Fourier Transform, the mixture can be represented in the short-time Fourier transform (STFT) domain as the summation of the short term spectra of speech S and of noise V[1]:

$$Y(l,m) = S(l,m) + V(l,m)\,, \tag{3.1}$$

where l is the frame index and m is the frequency bin index.

The gain function method is adopted in this framework where, for each frame l, the estimate of the underlying clean speech spectrum is obtained by multiplying the noisy input spectrum with a gain function $G(l,m)$:

$$\widetilde{S}(l,m) = G(l,m)Y(l,m)\,. \tag{3.2}$$

The $G(l,m)$ are typically (see, e.g., [12]) obtained as functions of the following two parameters: the *a priori* SNR $\xi(l,m)$ and the *a posteriori* SNR $\gamma(l,m)$, respectively defined as:

$$\xi(l,m) = \frac{\lambda_s(l,m)}{\lambda_v(l,m)}\,, \tag{3.3}$$

and

$$\gamma(l,m) = \frac{|Y(l,m)|^2}{\lambda_v(l,m)}\,, \tag{3.4}$$

where $\lambda_s(l,m)$ and $\lambda_v(l,m)$ represent the power spectral densitys (PSDs) of the speech and noise signals, respectively. However, since the true values of $\lambda_s(l,m)$ and $\lambda_v(l,m)$ are not available, their estimates, $\widehat{\lambda_s}(l,m)$ and $\widehat{\lambda_v}(l,m)$, are substituted into the above formulae to compute the *a priori* and *a posteriori* SNRs for the calculation of the gain function.

3.2.1 Overview of CEM-based Speech Enhancement Framework

To improve the speech estimate $\widehat{\lambda_s}(l,m)$, especially for the voiced speech segments, the two-stage speech enhancement framework is proposed in [10], and is summarised below.

In the first stage, a preliminary noise reduction is applied, resulting in an initial speech estimate $\widehat{S}(l,m)$. This is obtained by applying the MMSE-LSA gain function together with the DD approach [2]. The noise PSD $\widehat{\lambda_v}(l,m)$ is estimated by the minimum statistics (MS) approach [13].

[1] We follow the standard convention: uppercase letters indicate quantities in the spectral domain; lowercase variables are time-domain signals. Since the cepstrum, defined as the inverse transform of the logarithmic spectrum, is also a quasi-temporal representation and lowercase letters are adopted for cepstral variables as well.

Using linear prediction coding (LPC) analysis [6], $\widehat{S}(l,m)$ is decomposed into the *envelope* $H(l,m)$ and the *residual* $\widehat{R}(l,m)$, which is also termed the speech excitation signal.

The key idea of [10] lies in the manipulation of this excitation signal in the *cepstral* domain. First, the excitation signal R is converted to cepstrum where fundamental frequency can be easily detected. The speech harmonics are selectively boosted as detailed below. Applying the original speech envelope $|\widehat{H}(l,m)|$ to this enhanced excitation signal $|\overline{R}|(l,m)$, an idealised speech estimate $|\overline{S}(l,m)|$ can be obtained. A new *a priori* SNR $\widetilde{\xi}_l$ is then calculated from this harmonic-enhanced estimate $|\overline{S}(l,m)|$ to obtain the final gain function for that frame. This method is named as CEM_{ID} because it replaces the original excitation with an idealised one. Along with the traditional MMSE-LSA approach, it forms a baseline for our work. The signal flow graph of CEM_{ID} is graphically illustrated in Figure 3.1. The details of the cepstral manipulation of CEM are now presented.

3.2.2 CEM_{ID} in detail: F_0 Detection

To analyse the periodicity of voiced frames, the excitation signal amplitude $|\widehat{R}(l,m)|$ is transformed into the cepstral domain by a Q-point discrete cosine transformation of type II (DCT-II):

$$\begin{aligned} c_l^r(q) &= \text{DCT}\{\ln(|\widehat{R}(l,m)|)\} \\ &= \sum_{m=0}^{M/2} \ln(|\widehat{R}(l,m)|) \cdot \cos\left[\frac{\pi q}{Q}(m+0.5)\right], \end{aligned} \tag{3.5}$$

where $q = \{0, 1, \ldots, Q-1\}$ denotes the quefrency bin index. Since the amplitude spectrum is symmetric, only half of the residual spectrum $|\widehat{R}(l,m)|$ is required for cepstrum calculation and Q is set to $M/2+1$. The fundamental frequency and its harmonics correspond to a peak in the cepstrum. For the signal sampled at f_s, the relationship between frequency f and its corresponding quefrency bin is $f = f_s/q$. Therefore, the fundamental frequency F0 at frame l can be obtained by finding its corresponding quefrency bin q_{F0} where the excitation cepstrum $c_l^r(q)$ achieves its maximum in the allowed quefrency range. The estimated fundamental frequency is then given by

$$\text{F0}(l) = \frac{f_s}{q_{\text{F0}}(l)}, \tag{3.6}$$

where

$$q_{\text{F0}}(l) = \underset{q \in \mathcal{Q}}{\text{argmax}}\{c_l^r(q)\}. \tag{3.7}$$

Given that the fundamental frequency of human speech usually falls in the range from 50 to 500 Hz [6], the search boundary in the quefrency domain is constrained correspondingly as $\mathcal{Q} = [q_{f=500}, \ldots, q_{f=50}]$.

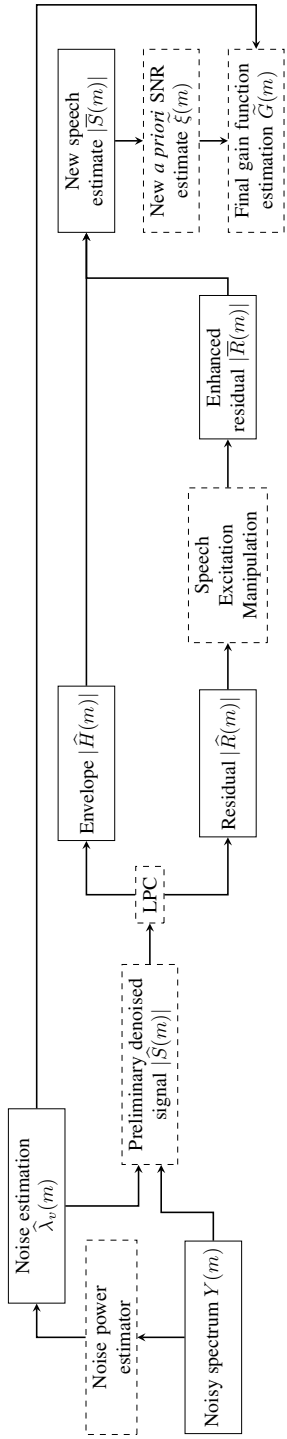


Figure 3.1: Block diagram of the gain function calculation in two-stage noise reduction. Dashed boxes represent manipulation blocks whereas solid rectangular boxes indicate data contained. Please note that all terms are in the STFT domain, where the frame index *l* has been dropped for conciseness.

3.2.3 CEM_{ID} in detail: Excitation Manipulation

Following the identification of $q_{\text{F0}}(l)$, the original excitation is completely replaced by a synthesised excitation $\overline{c^r}_l(q)$: $c_l^r(0)$, as an indication of energy level, is preserved in the idealised excitation $\overline{c^r}_l$. The cepstral peak is *amplified* by a predetermined constant $\alpha_{c^r}(>1)$. The rest of the cepstrum is discarded, i.e.,

$$\overline{c^r}_l(q) = \begin{cases} c_l^r(0), & q = 0 \\ \alpha_{c^r} \cdot c_l^r(q), & q = q_{\text{F0}},\ \alpha_{c^r} > 1 \\ 0, & \text{otherwise} \end{cases} . \tag{3.8}$$

This excitation manipulation helps with speech enhancement in two ways. On the one hand, the speech harmonics are emphasised by scaling up the harmonic-related cepstral peak. On the other hand, the remaining noise is removed by nulling the excitation cepstrum.

The new speech residual amplitude $|\overline{R}|(l,m)$ spectrum can be acquired by the inverse DCT-II (iDCT) of the idealised excitation cepstrum $\overline{c^r}_l$:

$$\begin{aligned} |\overline{R}|(l,m) &= \exp(\text{iDCT}\{\overline{c^r}_l\}) \\ &= \exp\left(\frac{\overline{c^r}_l(0)}{Q} + \frac{2}{Q}\sum_{q=1}^{Q-1} \overline{c^r}_l(q) \cdot \cos\left[\frac{\pi q}{Q}(m+0.5)\right]\right) . \end{aligned} \tag{3.9}$$

The synthesis procedure described by Eq. (3.9) could lead to false peaks or an undesired rising tendency at the edges of the spectrum $|\overline{R}|$. This is addressed by a cosine decay in [10] at the spectral edges to avoid these artefacts in the enhanced signal, namely by linearly extending the spectrum from the trough before the first peak and from the trough following the peak of the last harmonic to the respective edges.

3.2.4 Speech Estimation

With the idealised excitation $|\overline{R}|(l,m)$ and the speech envelope $|\widehat{H}(l,m)|$, the speech amplitude spectrum is synthesised as $|\overline{S}(l,m)| = |\overline{R}|(l,m) \cdot |\widehat{H}(l,m)|$. Instead of directly using this synthetic spectrum as clean speech estimate, it is proposed to use this idealised spectrum to *re-estimate* the *a priori* SNR $\widetilde{\xi}_l(l,m)$ as:

$$\widetilde{\xi}_l(l,m) = \frac{|\overline{S}(l,m)|^2}{\widehat{\lambda_v}(l,m)} . \tag{3.10}$$

Using this $\widehat{\xi}(l,m)$ and the previously computed *a posteriori* SNR $\widehat{\gamma}(l,m)$, the final *gain* function $\widetilde{G}(l,m)$ is computed in the standard manner (e.g., MMSE-LSA). Applying this gain function yields the final clean speech spectrum estimate: $\widetilde{S}(l,m) = \widetilde{G}(l,m)Y(l,m)$. The clean speech estimate in the time domain is

obtained by applying inverse short-time Fourier transform (iSTFT) and overlap-add synthesis.

3.3 Improved Excitation Manipulation

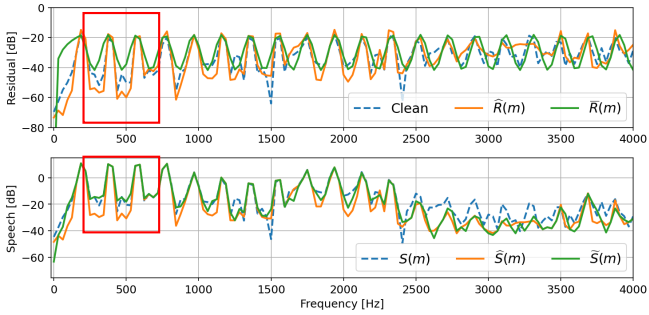
3.3.1 Analysis of the drawbacks of CEM

Figure 3.2 presents the residual spectrum and speech estimation by CEM_{ID} of a voiced frame corrupted by white noise at SNRs of 15 dB (Figure 3.2a) and of -5 dB (Figure 3.2b). The speech excitation idealisation is reasonably accurate when SNR $= 15$ dB; however, CEM_{ID} is clearly influenced by residual noise for the same speech segment when SNR $= -5$ dB. Since CEM_{ID} is driven by the excitation manipulation and the envelopes are not modified, the deterioration of CEM_{ID} at -5 dB comes solely from the (insufficient) excitation synthesis. Compared with the clean speech spectrum, the estimated speech spectrum (green curve) at -5 dB (Figure 3.2b) is strongly overestimated between the harmonics. From the upper panel, it can be observed that the idealised excitation is even less clear than the excitation spectrum of the preliminary denoised signal (orange curve). As the result, the enhanced spectrum loses the 'sharpness' of its harmonics in voiced frames, especially for the first few harmonics (e.g., the harmonics inside the red square of Figure 3.2b). When being used to calculate the new *a priori* SNR, this weakened periodic structure results in poorer inter-harmonic noise suppression, leading to a vocoding effect in the final speech estimate.

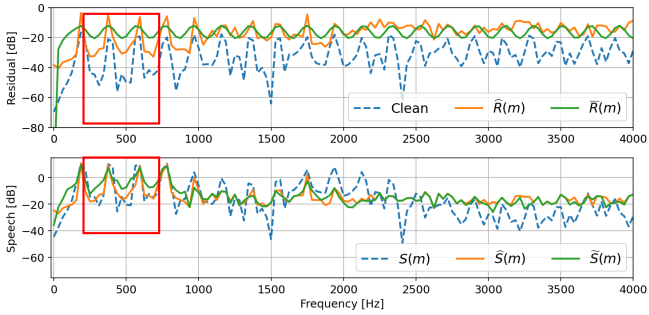
To understand this inherent weakness of CEM from an analytic perspective, we take a closer look at the enhanced excitation spectrum, which is written in the log domain as

$$\begin{aligned}
\ln(|\overline{R}|(l,m)) &= \text{iDCT}\{\overline{c^r}_l\} \\
&= \frac{\overline{c^r}_l(0)}{Q} + \frac{2}{Q}\sum_{q=1}^{Q-1}\overline{c^r}_l(q)\cdot\cos[\frac{\pi q}{Q}(m+0.5)] \\
&= \frac{\overline{c^r}_l(0)}{Q} + \frac{2}{Q}\overline{c^r}_l(q_{\text{F0}})\cdot\cos[\frac{\pi q_{\text{F0}}}{Q}(m+0.5)] + \\
&\quad \frac{2}{Q}\sum_{q\in[1,2,\ldots,Q-1]}^{q\neq q_{\text{F0}}}\overline{c^r}_l(q)\cdot\cos[\frac{\pi q}{Q}(m+0.5)] \\
&= \frac{\overline{c^r}_l(0)}{Q} + \text{iDCT}\{\mathcal{F}(c^r_{\text{pitch}})\} + \text{iDCT}\{\mathcal{G}(c^r_{\text{rest}})\}
\end{aligned} \quad , \tag{3.11}$$

where $c^r_{\text{pitch}} = [0,0,\ldots,0,c^r_l(q_{\text{F0}}),0,\ldots,0]$ and $c^r_{\text{rest}} = [0,c^r_l(1),c^r_l(2),\ldots,c^r_l(q_{\text{F0}}-1),0,c^r_l(q_{\text{F0}}+1),\ldots,c^r_l(q_{\text{F0}})]$. $\mathcal{F}(\cdot)$ and $\mathcal{G}(\cdot)$ are the manipulating functions on the respective cepstrum components. The three terms correspond to the log-



(a) With white noise, SNR $= 15$ dB



(b) With white noise, SNR $= -5$ dB

Figure 3.2: Speech enhancement by CEM on a voiced frame corrupted by white noise at different SNRs. In each figure, the clean reference, the preliminary denoised result ($\widehat{\cdot}$), the synthesised result ($\overline{\cdot}$), and the final estimate ($\widetilde{\cdot}$) are presented. The *upper panel* of each sub-figure compares the *synthetic excitation* spectrum with that of the preliminary denoised and the clean speech, whereas the lower panel compares the *final clean speech estimate* based on the idealised excitation with that of the preliminary denoised and the clean signal. The method suffers from vocoding, especially in the frequency region delineated by the red square. The artefact results from the amplitude underestimation of the synthetic excitation. This vocoding effect reduces when SNR increases.

spectrum energy, the harmonics, and the fine structure of the excitation. Accordingly, different manipulations are applied to the three terms in CEM_{ID}: the log-spectrum energy term remains untouched, $\mathcal{F}(\cdot)$ amplifies the harmonic term with a constant factor α_{c^r} to emphasise the periodic structure, and $\mathcal{G}(\cdot)$ sets the fine structure term to $[0, 0, \ldots, 0]$ for extra noise suppression and to reduce musical noise.

In [10], CEM_{ID} was designed to overestimate harmonic amplitudes by a fixed, pre-determined amplifying factor α_{c^r}. However, as demonstrated by the performance degradation of CEM_{ID} from the high-SNR condition to the low-SNR condition in Figure 3.2, the α_{c^r} suggested by the authors could be insufficient in certain

cases. To understand this, recall that cepstral coefficients are obtained by the inner product of the logarithmic amplitude spectrum and cosine basis functions, computed over the entire spectrum. However, when the harmonics are less pronounced (e.g., due to noise), the excitation signal loses its periodic structure in these regions. In such a case, the cepstral coefficient corresponding to $\mathrm{F0}$ can still stand out because of the recognisable harmonics (typically in the lower frequency regions); however, its value is attenuated because the harmonic regions are averaged with the unstructured ones. This issue becomes more prominent for wideband speech since harmonic structure is typically more compromised at the higher frequencies regions, leading to more distorted STFT bins that attenuate the $\mathrm{F0}$-related coefficient. Thus, no matter what *constant* value is chosen as the amplifying factor, there is always the risk of insufficient excitation in $\mathrm{CEM_{ID}}$.

The other problem of $\mathrm{CEM_{ID}}$ is that cepstrum nulling leads to artificiality in the enhanced speech, as has been noticed in [10]. Eq. (3.8) has showed that $\mathrm{CEM_{ID}}$ does *not* preserve any fine spectral structure. It is therefore proposed in [10] to improve the naturalness of the enhanced speech by the template method, where each template is particular to a range of fundamental frequencies. From the perspective of Eq. (3.11), this modification replaces the nulling function $\mathcal{G}(\cdot)$ of $\mathrm{CEM_{ID}}$ by the $\mathrm{F0}$-related templates which introduces more coefficients into c^r_{rest}.

By reformulating the speech excitation signal synthesis into three individual, interpretable components in the above analysis, it is easier to understand and appreciate the two limitations of $\mathrm{CEM_{ID}}$ when used for speech harmonic enhancement. In the following sections, we will propose our modifications targeting these two shortcomings. We will leave the energy term intact, and enhance the other two terms separately.

3.3.2 Residual Amplitude Estimation (RAE)

The first modification is to replace the constant scaling factor α_{c^r} of $\mathrm{CEM_{ID}}$ with a *data-adaptive* factor. Since the energy of voiced speech is typically concentrated in the low-frequency region, as shown in Figure 3.2b, the preliminary noise reduction performs better in the first few low-frequency harmonics, where the SNR is higher. Therefore, we propose to deduce the speech excitation dynamic range from these clearer harmonics. The other consequence of this energy distribution is the decay of the excitation dynamic range as frequency increases. Therefore, *one* amplifying factor that suits low-frequency harmonics will overestimate the high-frequency harmonics and lead to annoying artefacts in the enhanced audio. Motivated by these observations, we propose the residual amplitude estimation (RAE) for $\mathcal{F}(c^r_{\mathrm{pitch}})$ in Eq. (3.11). Our RAE consists of three components: an adaptive amplifying factor τ for the excitation dynamic range, an amplitude decay function $\omega_l(m)$ to avoid overestimation in the high-frequency region, and a cosine function to model the

harmonic structures. Based on these three components, the harmonic term of the synthetic excitation can thus be written in the log-domain as:

$$\text{iDCT}\{\mathcal{F}(c^r_{\text{pitch}})\} = \tau \cdot \omega_l(m) \cdot \cos[\frac{\pi q_{\text{F0}}}{Q}(m+0.5)]\,. \tag{3.12}$$

3.3.2.1 Adaptive Amplifying Factor τ

From Section 3.3.1, we see that the height of the cepstral peak $c^r_l(q_{\text{F0}})$, which may be seen as an analogue of the signal energy at the harmonics of F0, is affected by the noise, especially in the higher frequencies. Since speech energy for voiced segments is typically concentrated in the lower frequency regions, an analysis of the harmonic peaks in these regions would provide a better basis for estimating the amplification factor. Therefore, we introduce topographic prominence [14], which provides local information of peaks, for a data-dependent estimate of the scaling factor. A peak is defined as a local maximum and its prominence describes how much the peak stands out from its neighbourhood. The prominence is defined as the vertical distance between the peak and its lowest contour line. Given the excitation spectrum $|R(l,m)|$, the set of local peaks is first identified. Then, a set of prominences $P = \{p_1, p_2, \dots\}$ can be obtained by calculating the prominence for each peak. Figure 3.3 shows an example to calculate the ith prominence p_i according to the following steps:

- Extend the peak value horizontally until it crosses the signal or reaches the analysis interval boundary.
- Define the bases of the peak as the lowest values of the signal on each side.
- Define the maximum of the two bases as the contour line.
- The prominence p_i of the ith peak is defined by the vertical difference between the contour line and the peak.

Prominence measures the 'local height' of each peak. Since Eq. (3.12) models the periodic structure by a cosine function, the prominence of the excitation signal is twice the amplitude of this cosine (i.e., the dynamic range of the excitation) in an ideal case. Due to the energy concentration of speech in low frequencies, we assume a high SNR in the first few harmonics and thus true excitation can be recovered after preliminary denoising. The scaling factor τ is then deduced from the prominences whose corresponding peak frequencies are below 1000 Hz:

$$\tau = \frac{\max(p_i | p_i \in P, f_{p_i} \leq 1000\text{Hz})}{2}\,, \tag{3.13}$$

where f_{p_i} is the frequency of the ith peak. The analysis interval of prominence is set to $2 \cdot \text{F0}(l)$ to ensure that the prominence p_i measures the true local height of the harmonics.

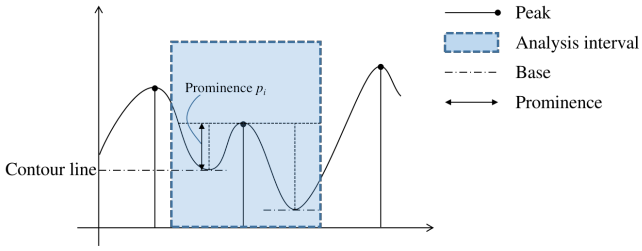


Figure 3.3: Topographic prominence: the local peak is extended horizontally to both sides until crossing the signal (the left end) or reaching the pre-determined analysis interval boundary (the right end). Since the left base is higher than the right one, it is used as the contour line to calculate prominence.

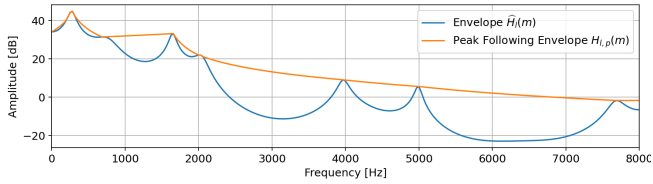
We note that in practice noisy fluctuations in the excitation could be recognised as false peaks. However, only peaks at speech harmonics and their vicinity are desired. To ensure this, the minimum distance between two detected peaks is set as $0.8{\cdot}M_{\mathrm{F0}}(l)$, where $M_{\mathrm{F0}}(l)$ is the frequency bin index of the fundamental frequency of the current frame. In this distance, only one major peak can be identified.

3.3.2.2 Amplitude Decay $\omega_l(m)$

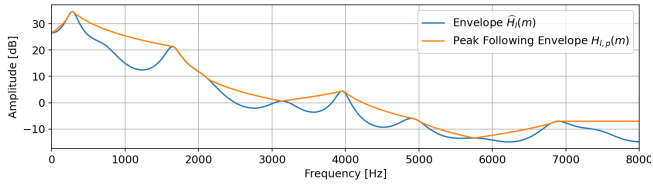
An extra weighting rule in the frequency domain is introduced to ensure that residual amplitudes are properly tapered down in low-energy bands to generate more natural speech. Intuitively, the dynamic range of harmonics in low-energy bands should be small as well. The speech envelope $|\widehat{H}(l,m)|$ from LPC, which captures speech formants, is a good indicator of the spectral amplitude trend. Therefore, we introduce the weighting rule $\omega_l(m)$ based on $|\widehat{H}(l,m)|$ to address the mismatch between the synthetic harmonic dynamic range and the speech spectral energy. We first calculate the peak-following envelope $H_l^p(m)$ of $|\widehat{H}(l,m)|$ by linear interpolation between two adjacent peaks of $|\widehat{H}(l,m)|$ in the linear domain as shown in Figure 3.4. Then, max normalisation and clipping as shown in Eq. (3.14) are applied to convert the peak following envelope $H_l^p(m)$ into the weighting rule $\omega_l(m)$. The excitation will be decayed below fundamental frequency after cosine edge decay, so no weighting rule is applied to these frequencies, i.e., $\omega_l(m) = 1$. When frequency exceeds F0, the lower boundary of $\omega_l(m)$ is set as 0.1 to prevent a total harmonic suppression.

$$\omega_l(m) = \begin{cases} 1, & m \le M_{\mathrm{F0}}(l) \\ \max\left\{ \dfrac{H_l^p(m)}{\max\limits_{m\in\{0,\ldots,M/2\}} H_l^p(m)}, 0.1 \right\}, & m > M_{\mathrm{F0}}(l) \end{cases} . \tag{3.14}$$

For the same reason described in section 3.2.3, extra spectrum decay at the excitation spectrum edges is also required in our modification.



(a) An speech active frame, clean



(b) The same speech active frame with pub noise, SNR $= 5$ dB

Figure 3.4: Examples of the peak following envelope for the same speech segment under different SNRs.

Figure 3.5 compares enhanced speech spectra by the proposed model (Eq. (3.12) – Eq. (3.14)) and the baseline method CEM_{ID}. It can be observed that CEM_{ID} blurs the initial four harmonics due to the insufficient excitation, while the proposed modification is able to suppress the inter-harmonic noise to a satisfactory level. This indicates that the proposed method is able to restore sharpened harmonics even in adverse conditions, which is beneficial to speech quality.

3.3.3 Cepstrum Smoothing (CC)

Given the transformation between spectrum and cepstrum in Eq. (3.5) and Eq. (3.9), we see that the low-quefrency cepstral coefficients describe the coarse structure (envelope) of the spectrum, while the high-quefrency coefficients include both speech fine structure and noise-related fluctuations. By nulling all the cepstral coefficients in c^r_{rest}, musical noise is clearly reduced in CEM_{ID}. However, as discussed in [10], this improvement comes at the cost of speech naturalness, because the excitation fine structure is lost in this operation.

Aiming at more naturalness, we propose to improve the CEM_{ID} by preserving more cepstral coefficients. Instead of cepstrum nulling during excitation synthesis, the excitation cepstrum is *smoothed* along the quefrency-axis so that more details are preserved while noise is suppressed. We choose to leave coefficients whose quefrencies are lower than a certain threshold quefrency q_{low} unchanged – thereby preserving the general shape – and to average the remaining coefficients with rectangular moving-average windows whose lengths are proportional to the *bit-length*

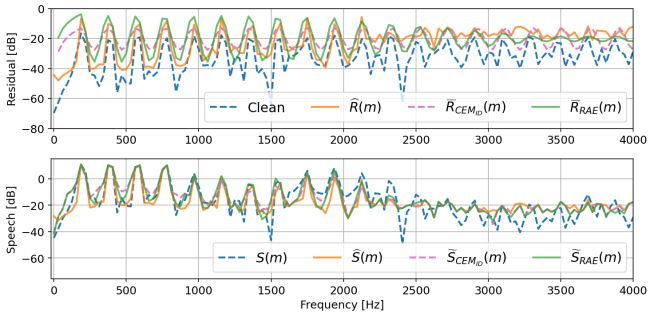


Figure 3.5: Demonstration of the benefit of an adaptive pitch amplification factor. The top panel shows the improved speech excitation, and the bottom one shows the benefit of the improved excitation on the final speech estimate for CEM_{ID} and RAE. For comparison, the result of the preliminary denoising and the noisy input frame are also provided. The input is the same voiced frame as in Figure 3.2, mixed with white noise at 0 dB. Apart from the clean reference and the preliminary denoising results, the signals enhanced by CEM_{ID} are also provided for comparison. In generating the synthetic excitation, all cesptral coefficients except bin 0 and the bin corresponding to fundamental frequency are discarded. The cepstral value corresponding to the fundamental frequency is scaled up by a constant pitch amplifying factor $\alpha_{c^r} = 2$ for CEM_{ID} (from [10]) and by an adaptive factor, computed as proposed in Eq. (3.12) – (3.14) for the RAE.

of quefrency bin index q:

$$\bar{c}_{\text{rest},l}(q) = \begin{cases} c^r_{\text{rest}}(q), & q \leq q_{low} \\ \sum_{i=q-n}^{q+n} \frac{n-|q-i|+1}{n^2} c^r_{\text{rest}}(q), & q > q_{low} \end{cases}, \tag{3.15}$$

where $n = \lfloor \log_2(q) \rfloor$. By choosing a window size proportional to the quefrency index, we maintain the averaging interval (for a given sampling rate) even if the DFT analysis window length changes. Rather than define arbitrary functions for window sizes based on the quefrency index, the choice of using the bitlength was an empirical solution that gave the best results based on several trials. We term this method as cepstral convolution (CC).

Figure 3.6 compares the proposed cepstral smoothing scheme with CEM_{ID}. It can be observed that our method retains more spectral structure, especially in low frequencies. Another advantage of preserving more cepstral coefficients is to partially compensate for the quantisation error of the pitch detection and the attenuation of the cepstral coefficient corresponding to F0 in low SNRs.

The two modifications (RAE and CC) enhance speech harmonics from complementary aspects. The result of the combined estimation is shown in Figure 3.7, where the peaks at the harmonics follow that of the clean speech and the valleys

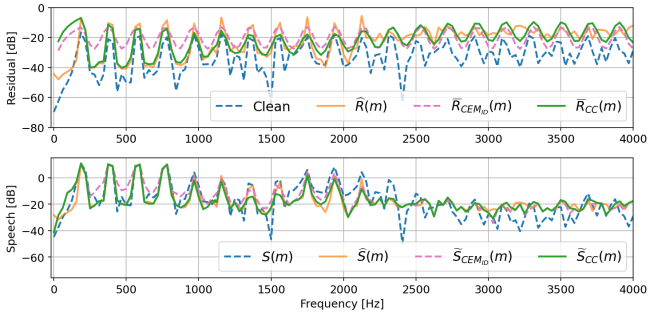


Figure 3.6: Demonstration of the benefit of preserving cepstral fine structure. The input is the same voiced frame as in Figure 3.2, mixed with white noise at 0 dB SNR. To generate the synthetic excitation with harmonic enhancement, a constant amplification factor of 2 is applied to c^r_{pitch}. The c^r_{rest} is set to zeros as proposed for CEM_{ID}; or smoothed by Eq. (3.15) for the CC approach.

are well accentuated, resulting in a better inter-harmonic noise suppression. Please also refer to Figure 3.8 for an appreciation of how these modifications help with the final goal of speech enhancement.

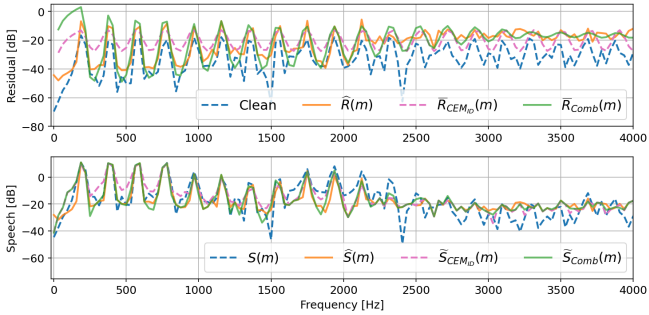


Figure 3.7: Demonstration of the benefit of the combined method (applying RAE and CC) to generate the enhanced excitation. The input is the same voiced frame as in Figure 3.2, mixed with white noise at 0 dB SNR.

3.4 Evaluation

The examples in Figure 3.5 – Figure 3.7 were chosen to visually demonstrate the benefits of the proposed modifications within the framework of CEM_{ID}. Now we present a more rigorous evaluation of the proposed improvements.

We use the MMSE-LSA gain function with the DD approach as preliminary noise reduction. The four different excitation manipulation methods discussed

previously (summarised in Table 3.1) are systematically compared. The baseline approach, CEM_{ID}, is implemented as proposed in [10] with cepstrum nulling and a constant harmonic amplifying factor of $\alpha_c = 2$. The first variant is residual amplitude estimation (RAE) which replaces the constant harmonic amplifying factor by a data-adaptive factor. Our second variant, cepstral convolution (CC), still adopts the constant amplifying factor, but smooths cepstral coefficients, i.e., retaining more spectral information, as explained in section 3.3.3. Compared with the baseline CEM_{ID}, the results of these two methods show the respective improvement accrued due to each *individual* modification. Finally, the joint benefit of combining both modifications is evaluated.

Table 3.1: Evaluated methods

Method	c^r_{rest} manipulation	Harmonic synthesis
CEM_{ID}	Eq. (3.8) (nulling)	Eq. (3.8) with $\alpha_{c^r} = 2$
RAE	Eq. (3.8) (nulling)	Eq. (3.12) – (3.14), adaptive estimation
CC	Eq. (3.15) (smoothing)	Eq. (3.8) with $\alpha_{c^r} = 2$
Comb	Eq. (3.15) (smoothing)	Eq. (3.12) – (3.14), adaptive estimation

3.4.1 Experimental Setup

The four methods were evaluated on the PTDB-TUG database [15]. The database contains clean utterances from 20 speakers (10 males and 10 females). For each speaker, five sentences were randomly chosen from the corpus. Five different signals from the ETSI noise database [16] were mixed with the clean speech at six different SNRs: $\{-5$ dB, 0 dB, 5 dB, 10 dB, 15 dB, 20 dB$\}$. The noise signals were: *white noise*, *car* (stationary, low-frequency noise), *highway* (non-stationary, low-frequency noise), *buccaneer 1* (narrow-band noise), and *pub* (babble noise). All the speech and noise signals were down-sampled to 16 kHz. To mix the signals at the chosen SNR, the level of the clean speech was measured by the active speech level (according to ITU-T P.56[17]), and that of noise by the long-term root-mean-square (RMS) value.

For all methods, the frame length is 512 points with a 50% overlap between frames. A square-root von Hann window is used for the analysis and the synthesis. The discrete Fourier transform (DFT) length is $M = 512$ samples. As the benchmark approach, the parameters of CEM_{ID} are identical to [10]. The order of LPC and the parameter q_{low} for our proposed method CC in Eq. (3.15) are set to 20 and 10, respectively.

To avoid gain function overflow, the *a posteriori* SNR is limited between -40 dB and 40 dB, and the lower boundary of *a priori* SNR is -25 dB. For both preliminary denoising and final speech estimation, the gain function is lim-

ited between -15 dB and 0 dB. The smoothing factor α of the DD approach is 0.98 for preliminary noise reduction.

3.4.2 Noise estimation

Instead of using the MS noise estimator as proposed in [10], the minimum mean square error-speech presence probability (MMSE-SPP) approach with fixed priors [18] is adopted in our work. It has been noted in [18] that MS suffers from noise floor overestimation and delay, while MMSE-SPP is capable of un-biased and fast noise tracking. Note that this noise estimator is used for *all* the evaluated approaches in Table 3.1. As suggested in [18], we assume an equal *a priori* probability for speech presence and absence $P(\mathcal{H}_0) = P(\mathcal{H}_1)$ without prior knowledge, and the optimal *a priori* SPP is set to 15 dB for MMSE-SPP.

3.4.3 Quality Measures

First, the same measures as in [10] are employed for a direct comparison, namely noise attenuation, speech-to-speech-distortion ratio, and ΔSNR.

Additionally, noise attenuation and ΔSNR of *the speech active frames* are also employed to highlight the benefits of the excitation manipulation methods. The quality of the different methods is evaluated by the white-box approach [19]. Having obtained the gain function to denoise the observed noisy signal, it is then *separately* applied to the noise component $v(n)$ and to the clean speech component $s(n)$. The measures are subsequently based on the filtered noise $v'(n)$ and the filtered speech $s'(n)$.

The segmental noise attenuation (NA) is calculated as

$$\mathrm{NA}_{\text{seg}} = 10 \log_{10} \left[\frac{1}{L} \sum_{l=0}^{L-1} \mathrm{NA}(l) \right] , \tag{3.16}$$

with

$$\mathrm{NA}(l) = \frac{\sum_{k=0}^{T-1} v(k + lT + \Delta)^2}{\sum_{k=0}^{T-1} v'(k + lT)^2} , \tag{3.17}$$

where T is the frame length, Δ is to compensate the sample delay of the filtered signal (after overlap-add synthesis) and L is the total number of frames in the signal. Higher NA indicates better noise reduction ability of the evaluated method.

The segmental speech-to-speech-distortion ratio (SSDR) measures the distortion introduced by speech enhancement. Higher SSDR suggests less distortion in speech. For frame l, the speech distortion is defined as

$$e(k + lT) = s'(k + lT) - s(k + lT + \Delta),\ k \in (0, T - 1) . \tag{3.18}$$

The single frame $\mathrm{SSDR}(l)$ is given by

$$\mathrm{SSDR}(l) = 10\log_{10}\left[\frac{\sum_{k=0}^{T-1} s(k+lT)^2}{\sum_{k=0}^{T-1} e(k+lT)^2}\right] . \tag{3.19}$$

Since the speech distortion should only be evaluated on the speech active frames, the measured SSDR is the average $\mathrm{SSDR}(l)$ on the set of speech active frames L_1:

$$\mathrm{SSDR} = \frac{1}{||L_1||}\sum_{l\in\{L_1\}} \mathrm{SSDR}(l)\,, \tag{3.20}$$

where $||L_1||$ is the cardinality of L_1.

In addition, the difference between noisy signal SNR, $\mathrm{SNR_{in}}$, and the denoised signal SNR, $\mathrm{SNR_{out}}$, provides us with global information about the SNR improvement of the methods. The speech level is calculated according to the active speech level measure from ITU P.56 [17] and the noise level is taken from the long-term RMS value of the noise signal. $\mathrm{SNR_{in}}$ is decided by the difference between the active levels of its two components, $s(n)$ and $v(n)$. After the noise reduction, $\mathrm{SNR_{out}}$ is obtained in the same manner by the two *filtered* components, $s'(n)$ and $v'(n)$.

$$\Delta\mathrm{SNR}(l) = \mathrm{SNR_{out}}(l) - \mathrm{SNR_{in}}(l)\,. \tag{3.21}$$

Further, to investigate the noise reduction ability in speech active frames, NA of speech active frames $\mathrm{NA_{act}}$ is calculated on the set of speech active frames L_1:

$$\mathrm{NA_{act}} = 10\log_{10}\left[\frac{1}{||L_1||}\sum_{l\in\{L_1\}} \mathrm{NA}(l)\right] . \tag{3.22}$$

Similarly, the SNR improvement in active frames $\Delta\mathrm{SNR_{act}}$ is given by

$$\Delta\mathrm{SNR_{act}} = \frac{1}{||L_1||}\sum_{l\in\{L_1\}} \left[\mathrm{SNR_{out}}(l) - \mathrm{SNR_{in}}(l)\right] . \tag{3.23}$$

The perceptual quality of the filtered clean speech components $s'(n)$ and of the denoised signals $\widetilde{s}(n)$ are evaluated by the wide-band perceptual evaluation of speech quality (WB-PESQ) [20]. The output of this metric is the mean opinion score - listening quality objective (MOS-LQO). PESQ MOS-LQO scores range from 1.04 to 4.64 and a higher score indicates better speech quality.Note that in the following we drop 'MOS-LQO' in the metric notations for conciseness, but the results are always on MOS-LQO scale rather than raw PESQ scores.

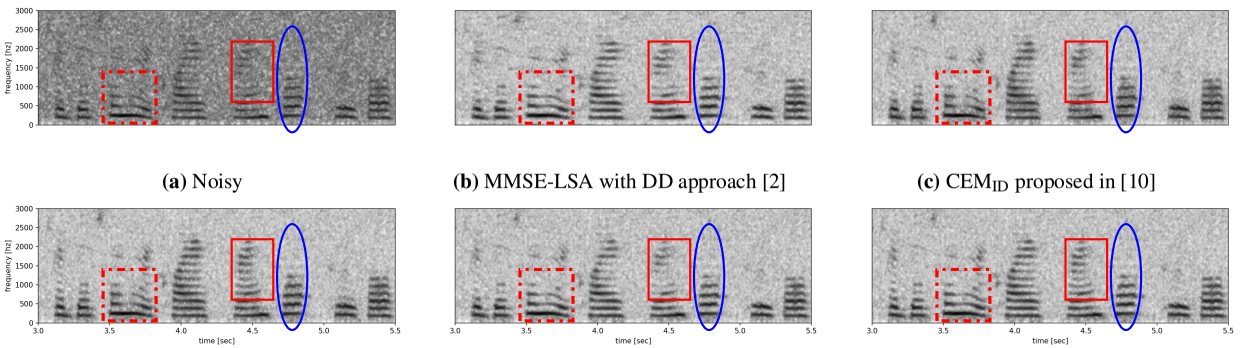
Two metrics, $\mathrm{PESQ_{st}}$ and ΔPESQ can be derived from the PESQ MOS-LQO score. ΔPESQ is defined as the PESQ score improvement of the enhanced signal compared to the noisy input. It is a comprehensive metric that takes all kinds of

artefacts in the processed signal into consideration. PESQ_{st} is the score of the filtered speech component. It illustrates the speech distortion introduced by the noise suppression gain function. We can see from the definition that PESQ_{st} is unable to detect insufficient noise reduction, e.g., it cannot reflect the gain function overestimation of CEM_{ID} in inter-harmonic frequencies; however, this overestimation indeed leads to a noticeable vocoding effect. Therefore, we additionally introduce $\Delta\text{PESQ}_{\text{act}}$ to compare the audio quality in speech *active* frames. For this, the PESQ scores of the noisy input and the *enhanced* signal $\widetilde{s}(n)$ are evaluated in the speech active frames $l \in L_1$. Before computing the PESQ, all the speech inactive frames $(l \notin L_1)$ in the noisy and the enhanced signal are replaced by the silence from the clean utterance. Thus, $\Delta\text{PESQ}_{\text{act}} = \text{PESQ}_{\widetilde{s},\text{act}} - \text{PESQ}_{y,\text{act}}$ reflects the speech quality improvement by the tested methods in the speech active frames.

It should be noted that PESQ was initially designed to measure speech quality degradation in telecommunication [20] and thus it is *not*, in general, a good metric to evaluate speech quality after noise suppression. To have a better idea of the speech quality, we also tested the methods by perceptual objective listening quality analysis (POLQA) metric, which is specifically designed for enhanced speech quality evaluation [21]. It allows for predicting speech quality over various distortions for wideband and super-wideband speech signals. Lastly, short-time objective intelligibility (STOI) [22] is employed to evaluate the intelligibility of the denoised signal.

3.4.4 Experimental Results and Discussion

We start with visual examples for an intuitive appreciation. Figure 3.8 shows the spectrograms of the noisy input, MMSE-LSA baseline approach, the CEM_{ID} baseline and the proposed excitation manipulation approaches. This is for the case when a clean speech utterance is mixed with white noise at 0 dB. Generally speaking, all four approaches recover the harmonics to a different degree. It can be observed when comparing Figure 3.8d and Figure 3.8e with the preliminary denoising result Figure 3.8b and the original CEM_{ID} in Figure 3.8c that our methods, adaptive harmonic enhancement and cepstral convolution, effectively address the previously discussed weaknesses of CEM_{ID}, and the combination of the two (Figure 3.8f) takes the advantages of both proposed modifications to yield an even superior result. For ease of exposition, three regions, where the respective contributions and advantages of the proposed approaches can be appreciated best, are highlighted. **Dash dot square**: this region shows the benefit of preserving more fine structure in the synthesised excitation. Compared to Figure 3.8b, the second and the third harmonics in Figure 3.8c and Figure 3.8d become weaker and discontinuous in time, indicating the drawback of methods purely focusing on the emphasising the F_0 and its harmonics, and neglecting the fine structure. In contrast, preserving this



(a) Noisy **(b)** MMSE-LSA with DD approach [2] **(c)** CEM_{ID} proposed in [10]

(d) Residual amplitude estimation (RAE) **(e)** Cepstral convolution (CC) **(f)** Comb

Figure 3.8: Example spectra contrasting the different methods for speech harmonics enhancement in a white noise condition at SNR $= 0$ dB. The spectra on the first row depict the noisy utterance and the results of the two baseline approaches. The spectra on the second row present the results of the proposed methods. **Dash dot square**: this region shows the benefit of preserving more fine structure. **Solid square**: this region highlights the effect of emphasising the fundamental and harmonic frequencies properly. **Oval**: this region best illustrates the drawbacks of a fixed amplification factor and cepstral nulling followed in CEM_{ID}. Detailed discussion can be found in Section 3.4.4.

structure by cepstral convolution (Figure 3.8e) yields a more time-continuous harmonic structure. The combination of adaptive harmonic enhancement and cepstral convolution (Figure 3.8f) yields the best result by combining the advantages of the two manipulations. **Solid square**: this region highlights the effect of *adequately* emphasising the frequencies corresponding to F_0 and the harmonics. Whereas CEM_{ID} is able to boost the low-amplitude harmonics to a certain extent, the benefit of an adaptive amplification factor is evident (Figure 3.8d and Figure 3.8f). **Oval**: this region best illustrates the drawbacks of a fixed amplification factor and cepstral nulling followed in CEM_{ID}. The insufficient inter-harmonic noise suppression is clearly visible and leads to an audible vocoding effect. RAE shows a slight improvement, while CC and the combined method generate the best results. In the following, the baselines and the proposed modifications are thoroughly evaluated by the metrics introduced in section 3.4.3. The results are grouped by SNRs of the input signals for detailed performance comparison.

3.4.4.1 Overall Average of Instrumental Metrics

Tables 3.2 and 3.3 show the measures averaged over the whole test data. Table 3.2 shows only the metrics employed in [10], whereas Table 3.3 presents the metrics, additionally, on speech active frames.

Table 3.2: Evaluation results: instrumental metrics utilised in [10], averaged on the whole test set

method	NA[dB]	SSDR[dB]	ΔSNR[dB]	PESQ_{st}
LSA	11.88	12.23	8.52	3.66
CEM_{ID}	12.20	12.89	8.44	3.67
RAE	11.72	13.59	8.36	3.66
CC	12.04	13.27	8.88	3.59
Comb	11.37	13.86	8.50	3.66

Table 3.3: PESQ MOS-LQO and other instrumental metrics on speech active frames, averaged on the whole test set

method	NA_{act}[dB]	$\Delta\text{SNR}_{\text{act}}$[dB]	ΔPESQ	$\Delta\text{PESQ}_{\text{act}}$
LSA	9.12	6.45	0.38	0.54
CEM_{ID}	8.91	5.97	0.39	0.55
RAE	8.58	5.98	0.41	0.60
CC	9.24	6.82	0.41	0.58
Comb	8.66	6.44	0.42	0.61

According to Table 3.2, CEM_{ID} yields higher NA than the other speech enhancement methods, and higher SSDR than the preliminary noise reduction methods. However, a different result is seen when focusing on the speech active frames (Table 3.3). Here CEM_{ID} does not provide benefit (in terms of NA and SSDR) over the baseline MMSE-LSA approach. This divergence indicates that, in terms of NA, CEM_{ID} benefits mostly from the extra noise reduction ability in silent frames when discarding the majority of the cepstral coefficients of the excitation signal. In terms of SSDR, this indicates the effect of inadequate harmonic emphasis and discarding spectral fine structure.

Comparing with CEM_{ID}, the proposed methods, RAE and CC, introduce less speech distortion, as indicated by a higher SSDR. Cepstral convolution provides good noise reduction in both the global sense and speech active frames. The combined method provides the best speech estimate in terms of speech quality (PESQ MOS-LQO). It shows the lowest NA but similar NA_{act} to CEM_{ID}, which suggests the overall metrics of the combined method are influenced by its performance in silent regions. This trade-off is expected since we also enhanced silent and unvoiced frames and thus generated false harmonics in them.

We described the excitation dynamic range underestimation of CEM_{ID} in section 3.3.1 with an example of Figure 3.8c; however, it is difficult to observe the degradation of this vocoding effect from the metrics in Table 3.2. The reason is that SSDR and PESQ_{st} are both evaluated on the filtered speech component, so insufficient inter-harmonic gain suppression will not cause artefacts in the filtered speech component (in contrast, it would actually be beneficial for these metrics!) This weakness of CEM_{ID} is, however, reflected by the comprehensive PESQ metrics: ΔPESQ and PESQ_{act}. Compared to the MMSE-LSA baseline, we see only a small benefit from CEM_{ID} in these metrics, whereas the proposed methods yield the best scores.

The SNRs of the test set lie in a wide range. To better appreciate the contributions of the various methods we now consider the results grouped by SNR.

3.4.4.2 ΔPESQ and $\Delta\text{PESQ}_{\text{act}}$

Figure 3.9 provides the results on ΔPESQ and $\Delta\text{PESQ}_{\text{act}}$ at different SNRs (mean improvement and the 95% confidence intervals). We provide the results in this manner because ΔPESQ indicates the overall performance over the noisy utterance (factoring in, thereby, possible degradations introduced by the methods due to errors in F_0 estimation and over-estimation of the harmonic amplitudes), whereas $\Delta\text{PESQ}_{\text{act}}$ is computed on *speech-active* frames. Thereby, $\Delta\text{PESQ}_{\text{act}}$ can better highlight the benefits of the proposed methods on the parts of the signal where they are expected to contribute most prominently.

The following insights are obtained. Firstly, in terms of overall quality, we may conclude that while CEM_{ID} yields an average ΔPESQ improvement compared to

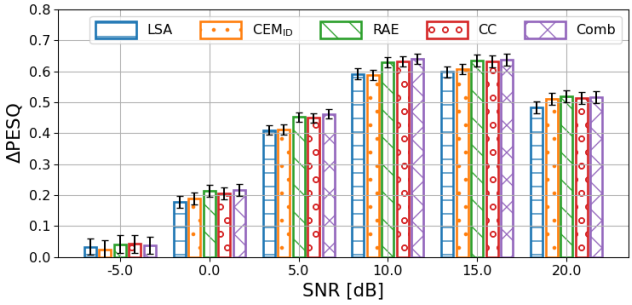
the baseline MMSE-LSA approach, this is only true for high SNRs (10 dB, 15 dB and 20 dB). At lower SNRs, CEM_{ID} introduces more distortion than this baseline. However, if we consider the confidence intervals, it may be disputed whether this difference is statistically significant. In contrast, each of our proposed improvements consistently provides a better ΔPESQ compared to MMSE-LSA and CEM_{ID}. Especially at SNRs from 5dB to 15dB, the difference may be considered statistically significant.

When we restrict the metric evaluation to the speech active frames ($\Delta\text{PESQ}_{\text{act}}$), CEM_{ID} shows a consistent improvement over the MMSE-LSA baseline for a wider range of input SNRs, but it is again debatable whether this difference is significant (strongly overlapping confidence intervals). In contrast, again, our proposed modifications have higher scores, with the differences being significant over the same SNR range as for ΔPESQ. On the basis of the *PESQ scores*, however, it is difficult to conclude whether the *combination* of RAE and CC (which has the highest score in all conditions) is a significant improvement over the two modifications considered individually. But, this is at least a first indication that the RAE and CC provide *complementary* improvements.

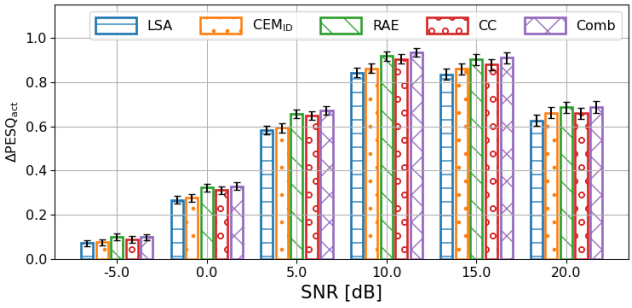
Meanwhile, the improvement in PESQ MOS-LQO of CEM_{ID} gradually approaches that of RAE as SNR increases, which indicates that the proposed RAE is a better candidate for the synthesised excitation, since CEM_{ID} can provide a good estimation under high SNRs.

3.4.4.3 POLQA

We also evaluated the performance of the five methods using POLQA, which is the new industry standard metric for benchmarking the voice quality for voice communications applications. The evaluation was performed on a subset of 500 gender- and noise-balanced samples with SNRs from -5 dB to 15 dB, which we believe are the most essential SNRs to observe the difference between methods according to the results of PESQ and PESQ_{act}. As shown in Figure 3.10, POLQA shows a similar trend as PESQ MOS-LQO: CEM_{ID} degrades speech quality in low SNRs, while the proposed methods are able to improve it under all conditions. In terms of POLQA we see that RAE and CC are consistently better than both baselines: MMSE-LSA and CEM_{ID}, and this performance is significant from an SNR of 5dB onwards. However, RAE and CC, compared to each other, seem to offer the same performance in terms of POLQA. It is now interesting to see that the *combined* method is again better than both RAE and CC, and this difference is significant. This is a strong indication of the complementary nature of the improvements offered by RAE and CC. The results of POLQA analysis have also been confirmed by a listening test by two experts. A Spearman's correlation coefficient of 0.91 between the expert scores and those of POLQA was found. These results are more reliable indicators of the quality improvements obtained, and reinforce



(a) Average ΔPESQ at different SNRs



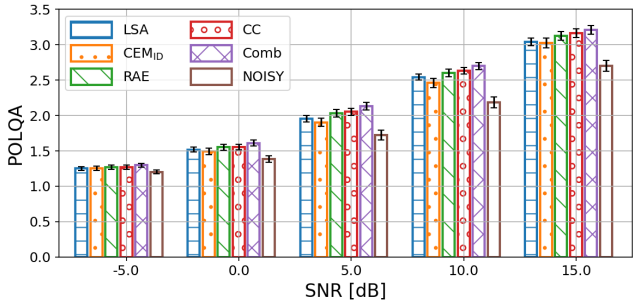
(b) Average $\Delta\text{PESQ}_{\text{act}}$ at different SNRs

Figure 3.9: Improvements in PESQ MOS-LQO on the whole signal (Figure (a)), and on the speech active frames (PESQ_{act}) (Figure (b)) for the different methods. The scores are averaged over the different noise types at each SNR. The error bars represent the 95% confidence interval.

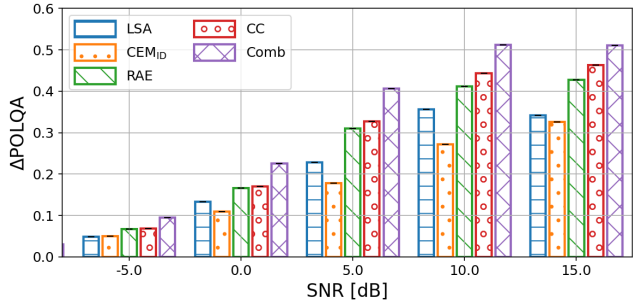
our conclusions based on the other metrics.

3.4.4.4 NA and SSDR

Figure 3.11 and Figure 3.12 demonstrate the effect of the proposed methods on the filtered individual components. Higher NA indicates better noise suppression of the method while higher SSDR indicates less distortion being introduced on the resultant speech during the processing. As expected, noise reduction as well as speech distortion decreases when SNR increases for all methods. It should be noted that in Figure 3.11 NA decreases only 1 dB from the lowest SNR cases to the highest SNR cases, while NA_{act} decreases by 5 or 6 dB from one extreme to another. This means a higher overall gain function in the speech active frames, which is expected in speech enhancement.



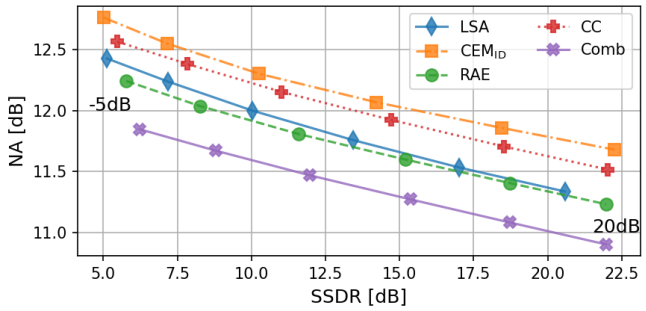
(a) Average POLQA at different SNRs



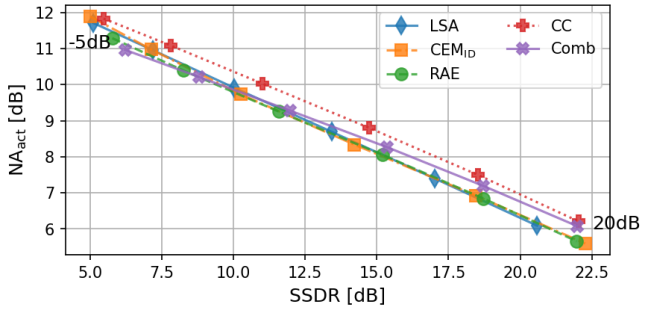
(b) Average ΔPOLQA at different SNRs

Figure 3.10: POLQA and ΔPOLQA of signals enhanced by the different methods, averaged over noise types at each SNR. The proposed methods (RAE, CC and combined methods) are able to improve speech quality in all cases. The combined method shows the highest improvement. The error bars represent the 95% confidence interval.

It is difficult to appreciate the difference among the methods because Figure 3.11 shows mainly the magnitude change. Therefore, choosing the preliminary denoising approach (MMSE-LSA) as the baseline, Figure 3.12 illustrates the improvement of NA, NA_{act} and SSDR of all excitation manipulation methods compared to this baseline. The performance difference on NA (Figure 3.12a) and NA_{act} (Figure 3.12b) indicates that CEM_{ID} strongly benefits from extra noise reduction of speech *inactive* frames at low SNRs, as observed from the overall average of these metrics. However, this advantage of NA comes at the cost of extra speech distortion (the lowest SSDR among four methods), which is also more noticeable in low SNR cases. The combined method scores lower than other methods in noise reduction. Note that CEM_{ID} or our modifications are always carried out without



(a) Average SSDR-NA of each method at different SNRs



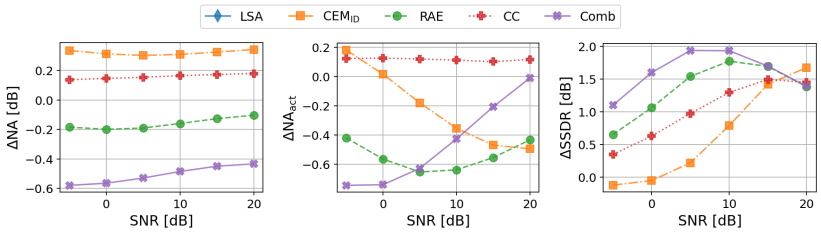
(b) Average SSDR-NA$_{act}$ of each method at different SNRs

Figure 3.11: SSDR-NA relationship of signals enhanced by different methods, averaged among all noise types at each SNR

explicit voiced and unvoiced detection. The lower noise reduction score of the combined approach may, therefore, be due to the over-amplification under the harmonic structure assumption in unvoiced and speech inactive frames. In terms of SSDR, the improvement of CEM$_{ID}$ increases sharply as the SNR increases. SSDR of CEM$_{ID}$ at -5 dB is even lower than MMSE-LSA, while our RAE yields more than 1 dB higher SSDR than the LSA baseline in that case. This comparison confirms our hypothesis that using a constant amplifying factor leads to noticeable artefacts due to the underestimation of residual dynamic range. In contrast, both RAE and CC are able to reduce speech distortion. The latter performs better in terms of noise reduction while the former in terms of speech preservation. The combined approach subsequently takes advantage of the complementary nature of these improvements and yields the best result.

3.4.4.5 STOI

The STOI scores, grouped by SNRs, are shown in Figure 3.13. The proposed methods exceed CEM$_{ID}$, but compared to noisy input, only the combined method



(a) NA **(b)** NA of the speech active frames **(c)** SSDR

Figure 3.12: Average improvement on NA, NA_{act} and SSDR of methods over preliminary denoising results (by decision-directed approach). The difference between Figure 3.12a and Figure 3.12b suggests that CEM_{ID} benefits from extra noise reduction in silent regions. Figure 3.12c shows that the proposed methods successfully improve speech quality.

is able to slightly improve the score in extreme conditions (SNRs of -5 dB, 0 dB, and 5 dB).

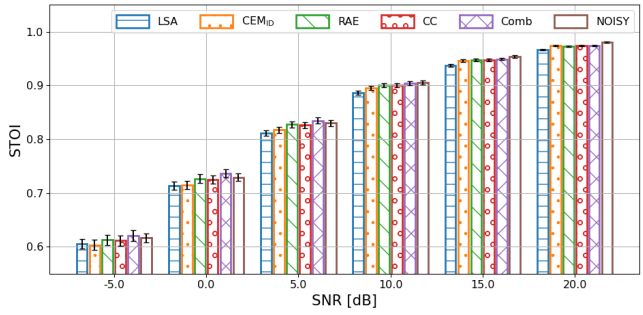


Figure 3.13: STOI of noisy signals and enhanced signals, averaged across different noise types at each SNR. The error bars represent the 95% confidence interval.

3.5 Conclusions

In this paper we have investigated the CEM approach proposed by [10] in detail. By reformulating the excitation synthesis problem, we were able to get a better insight into the inherent weakness of this approach, namely that the enhanced audio may lose its harmonic sharpness due to dynamic range underestimation and the loss of fine structure in the synthesised excitation signal spectrum. Based on our findings, we then proposed two modifications that are able to enhance the harmonic structure of voiced speech in a more natural and robust way. The proposed modifi-

cations include residual amplitude estimation and cepstral convolution smoothing. The evaluation results on multi-noise conditions show that the proposed modifications are better able to restore lost harmonics and sharpen the existing ones in voiced frames. Each modification, individually, improves over CEM_{ID}. The two modifications are, also, complementary. This is evident from the fact that the combined method scores higher than each modification individually. The improvement is still robust at low SNRs.

There is still room for further improvement. For example, a finer pitch estimation method could be beneficial. Current F0 estimation is based on peak-picking on the discrete quefrency bins and then calculating the corresponding frequencies. This could introduce a quantisation error when the actual fundamental frequency falls between the adjacent bins. Secondly, since we assume a harmonic structure for all frames, there is the risk of stronger musical noise, which has been reflected by the difference between metrics on active frames and these on the whole signal. This can be solved by introducing a voice activity detection module, and applying the proposed method to voiced speech frames only.

Lastly, we note that CEM need not be seen as a stand-alone method. In our work, we consider a statistical noise suppression framework within which CEM is integrated. However, in practice, CEM can be piggy-backed onto any denoising framework which can output an estimate of the gain function and noise floor. This also opens the possibility to integrate CEM within DNN-based frameworks, allowing for a marriage of model-based approaches and data-driven approaches, with all the ensuing benefits thereof. These are directions we will consider for the future.

We urge the reader to listen to the audio examples at https://yanjuesong.github.io/Improved-CEM-samples/.

Acknowledgment

The authors would like to thank the experts N.M.P. Neumann and J.G. Beerends at TNO Netherlands for helping with the POLQA scores and for conducting the listening tests.

References

[1] Y. Ephraim and D. Malah. *Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator*. IEEE Transactions on acoustics, speech, and signal processing, 32(6):1109–1121, 1984.

[2] Y. Ephraim and D. Malah. *Speech enhancement using a minimum mean-square error log-spectral amplitude estimator*. IEEE transactions on acoustics, speech, and signal processing, 33(2):443–445, 1985.

[3] C. Plapous, C. Marro, and P. Scalart. *Improved signal-to-noise ratio estimation for speech enhancement*. IEEE Transactions on Audio, Speech, and Language Processing, 14(6):2098–2108, 2006.

[4] C. Breithaupt, T. Gerkmann, and R. Martin. *Cepstral smoothing of spectral filter gains for speech enhancement without musical noise*. IEEE Signal processing letters, 14(12):1036–1039, 2007.

[5] C. Breithaupt, T. Gerkmann, and R. Martin. *A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing*. In 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, pages 4897–4900. IEEE, 2008.

[6] P. Vary and R. Martin. *Digital speech transmission: Enhancement, coding and error concealment*. John Wiley & Sons, 2006.

[7] T. Rosenkranz. *Modeling the temporal evolution of LPC parameters for codebook-based speech enhancement*. In 2009 Proceedings of 6th International Symposium on Image and Signal Processing and Analysis, pages 455–460. IEEE, 2009.

[8] R. Chen, C.-F. Chan, and H. C. So. *Model-based speech enhancement with improved spectral envelope estimation via dynamics tracking*. IEEE Transactions on Audio, Speech, and Language Processing, 20(4):1324–1336, 2011.

[9] T. Mellahi and R. Hamdi. *LPC-based formant enhancement method in Kalman filtering for speech enhancement*. AEU-International Journal of Electronics and Communications, 69(2):545–554, 2015.

[10] S. Elshamy, N. Madhu, W. Tirry, and T. Fingscheidt. *Instantaneous a priori SNR estimation by cepstral excitation manipulation*. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(8):1592–1605, 2017.

[11] S. Elshamy and T. Fingscheidt. *DNN-Based Cepstral Excitation Manipulation for Speech Enhancement*. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27(11):1803–1814, 2019.

[12] R. C. Hendriks, T. Gerkmann, and J. Jensen. *DFT-Domain based single-microphone noise reduction for speech enhancement*. Synthesis Lect. Speech and Audio Proc. Morgan & Claypool, 2013. doi:10.2200/S00473ED1V01Y201301SAP011.

[13] R. Martin. *Noise power spectral density estimation based on optimal smoothing and minimum statistics*. IEEE Transactions on speech and audio processing, 9(5):504–512, 2001.

[14] B. Kashyap, M. Horne, P. N. Pathirana, L. Power, and D. Szmulewicz. *Automated topographic prominence based quantitative assessment of speech timing in cerebellar ataxia*. Biomedical Signal Processing and Control, 57:101759, 2020.

[15] G. Pirker, M. Wohlmayr, S. Petrik, and F. Pernkopf. *A pitch tracking corpus with evaluation on multipitch tracking scenario*. In Twelfth Annual Conference of the International Speech Communication Association, 2011.

[16] ETSI. *ETSI noise database*. https://docbox.etsi.org/stq/Open/EG%2020 2%20396-1%20Background%20noise%20database/Binaural_Signals. (Last accessed 04/2022), 2006.

[17] ITU-T. *Rec. P.56: Objective Measurement of Active Speech Level*. International Telecommunication Union-Telecommunication Standardisation Sector, 2011.

[18] T. Gerkmann and R. C. Hendriks. *Unbiased MMSE-based noise power estimation with low complexity and low tracking delay*. IEEE Transactions on Audio, Speech, and Language Processing, 20(4):1383–1393, 2011.

[19] T. Fingscheidt, S. Suhadi, and S. Stan. *Environment-optimized speech enhancement*. IEEE transactions on audio, speech, and language processing, 16(4):825–834, 2008.

[20] ITU-T. *Rec. P.862.2: Corrigendum 1, Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs*. International Telecommunication Union-Telecommunication Standardisation Sector, 2017.

[21] ITU-T. *Rec. P.863: Perceptual objective listening quality prediction (POLQA)*. International Telecommunication Union-Telecommunication Standardisation Sector, 2018.

[22] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen. *A short-time objective intelligibility measure for time-frequency weighted noisy speech*. In 2010

IEEE International Conference on Acoustics, Speech and Signal Processing, pages 4214–4217. IEEE, 2010.

4

Aiding Speech Harmonic Recovery in DNN-Based Single Channel Noise Reduction Using Cepstral Excitation Manipulation (CEM) Components

Although thoroughly examined, the gain of optimising the statistical two-stage framework is limited. The analysis of the two-stage speech enhancement framework in the appendix A indicates that its performance bottleneck lies in the initial speech estimate. Therefore, in the following chapters, we adopt deep neural network (DNN)-based systems to better model speech distribution. In this chapter, the convolutional recurrent U-net architecture for speech enhancement (CRUSE) is chosen as the baseline system, whose architecture and training scheme have been systematically optimised for single channel speech enhancement. However, it is still challenging to identify and capture weak harmonics, especially in low signal-to-noise ratio (SNR) regions. This is similar to the problem we have addressed in Chapter 3. Given the findings on speech harmonic preservation in statistical methods, we propose to incorporate the source-filter model into the deep learning methods to tackle this challenge.

Instead of explicitly integrating the source-filter model into one specific architecture, we supplement the knowledge by including one extra loss term that focusses on harmonic reconstruction accuracy in the cepstrum. It measures the distance at the harmonic-related cepstral bin between the clean reference and the

enhanced signal. Thereby, the source-filter model is integrated with no extra computational cost at the inference stage. Inspired by the success of cepstral excitation manipulation (CEM) in boosting harmonics, we further amplify the harmonics of the clean reference by a constant factor in the cepstrum. It is a general solution to the harmonic over-suppression problem in DNN-based methods despite their loss functions and training targets.

⋆ ⋆ ⋆

Y. Song, and N. Madhu.

Published in the proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), June 2023.

Abstract Weak harmonics of voiced speech segments are often lost during the process of noise suppression – especially at low SNRs. This leads to a distortion in the harmonic structure, and an accompanying loss in quality. In this paper, inspired by previous work on speech harmonic enhancement using statistical methods, we present a loss function component we term cepstral excitation manipulation (CEM) loss, which is constructed based on the fundamental frequency-related cepstral coefficients. This component can be introduced to the training of state-of-the-art architectures and its benefit is benchmarked, here, on CRUSE. Experiments show that the proposed loss function component nicely supplements standard loss functions and the harmonic structure is better preserved. On average, the best system improves by 0.4 on PESQ and 0.47 on DNSMOS compared to the noisy input. Substantial improvements are primarily in low SNRs (-5 dB to 5 dB) – the range for which harmonic recovery is most required.

4.1 Introduction

Speech enhancement is a key component in communications and other audio-based applications. From statistical methods [1–3] to deep learning models [4–7], significant progress has been achieved in this topic. However, single-microphone speech enhancement remains challenging in low signal-to-noise ratio (SNR) scenarios. One problem with the enhanced signal in these conditions is the loss of the weak harmonics, which has been noted both in classical statistical frameworks [2, 3] and data-driven deep learning methods [5].

Compared to traditional model-based methods, data-driven approaches using deep neural networks (DNNs) have shown a great advantage in the quality of the enhanced signals. These models are usually trained in a supervised manner to either directly predict the underlying clean speech (e.g. [6, 7]) or, alternatively, a gain

function (e.g., [4, 5]) that preserves the speech components and attenuates noise. Furthermore, as speech structure can be better exploited in the short-time Fourier transform (STFT) domain, a gain prediction or spectral mapping in this representation is a frequently chosen training target for DNNs. Recently the convolutional recurrent U-net architecture for speech enhancement (CRUSE) has established itself as a state-of-the-art. The network is trained to predict a real-valued time-frequency (TF) gain function (mask), given the STFT representation of the noisy input. As demonstrated in [4], this architecture yields a considerable improvement to the noisy signals and allows for a range of trade-offs in terms of complexity and quality due to its compact and efficient architecture. In this architecture, the short-term context information is extracted in the encoder, made up of successive convolutional layers , whereas the long-term temporal modelling is handled by the recurrent layers in the bottleneck.

The powerful modelling ability of DNNs enables them to automatically learn to map the input into meaningful feature representations for denoising, so DNNs are often trained in an end-to-end manner, without explicitly exploiting any prior knowledge of speech. The common loss function for speech enhancement networks, the mean-square error (MSE), averages errors in all TF bins. Thereby, poor reconstruction of weak harmonics is not highly penalised. When DNNs are poor, this leads to a distorted harmonic structure in the enhanced signal with weakened and lost harmonics, as well as low noise suppression in the inter-harmonic regions.

Recent research has attempted to address this. In [5], this problem is tackled by a new loss function that explicitly increases the weight of the prediction loss in the harmonic frequency bins. Consequently, the focus of the network is driven more towards the error in these TF bins in the training. No extra computation is required in the inference stage. The fundamental frequency and its harmonics are identified by a sub-band-based auto-correlation analysis. Experiments show that emphasising the loss in these bins improves the harmonic preservation. This method is strictly limited to a training target in the TF domain. Moreover, this loss function increases the penalisation in the harmonics, but is not capable of increasing the contrast. In other words, the poor suppression of inter-harmonic regions is not really addressed by this loss function.

A similar problem exists in the statistical methods. It is not easy to separate weak harmonic components that are masked by background noise. Since the classical gain functions such as [1] are calculated independently for each TF bin, once the statistical model fails to separate a weak harmonic, there is no chance to recover it from the spectral context. Therefore, with an *explicit* speech production modelling, the exploiting of the source-filter decomposition provides a solution to empahsise the harmonics and to further suppress the inter-harmonic regions [2, 3, 8]. An enhancement of speech harmonics in the cepstral domain, dubbed the cepstral excitation manipulation (CEM), was first proposed in [2] within a two-stage framework.

The distorted excitation signals were partially restored by boosting the F0-related cepstral coefficient of the excitation signal with a constant factor. As an extension to this idea, in [3], we propose an adaptive amplifying factor for higher spectral contrast.

In [8], a DNN is used to directly estimate the speech excitation signal (CEM-DNN). The final gain function is then obtained by the statistical spectral weighting rule [1] based on the speech synthesised from the DNN-predicted excitation. The model of [3] can be piggy-backed onto this system to further improve the excitation estimate. However, as the explicit goal of the approach is to incorporate this within a classical model-based speech enhancement (with an estimated noise floor), the potential of the approach is limited. It is also difficult to combine this explicit model with end-to-end style DNNs.

In this work, encouraged by our previous results on CEM, we present a cepstral loss function component to train speech enhancement DNNs with an explicit emphasis on speech harmonic prediction errors. This component places no specific requirement on the domain of the training target. It also allows for an amplification of the harmonics in a controlled way - which was one of the benefits of the model-based CEM approach. To emphasise speech harmonics, the proposed loss component compares the network prediction to an *amplified* clean reference, in the cepstral domain. An adaptive amplification factor, proportional to the strength of the excitation, is also investigated. By thus formulating the loss in the cepstral domain, we implicitly encourage sinusoidal modelling of the excitation in the frequency domain which, as shown in [3], can not only preserve weak harmonics, but also improve noise suppression *between* the harmonics. We note, further, that this loss component can be mixed with any other major loss functions for network training. The benefit of the proposed loss component is benchmarked on the CRUSE net architecture [4], trained to predict the ideal ratio mask (IRM). Experimental validation using standard instrumental metrics confirms that a better estimate of speech harmonics is obtained when training with this extra component, and the benefit is more pronounced at low SNRs where the weak harmonics are prone to be distorted.

4.2 Signal model and CRUSE architecture

4.2.1 Signal model

We assume that the background noise $v(k)$ corrupts the observed microphone signal $y(k)$ in an additive way: $y(k) = s(k) + v(k)$, with k being the discrete time sample index. Using an M-point windowed FFT on overlapped and windowed signal segments, we obtain the STFT domain representation: $Y(l, m) = S(l, m) + V(l, m)$, where m is the frequency bin index and l is the frame index.

The goal of the network is to estimate the IRM $\left(\widehat{G}(l,m)\right)$ from the noisy input. The IRM is defined as:

$$G(l,m) = \left(\frac{|S(l,m)|^2}{|S(l,m)|^2 + |V(l,m)|^2}\right)^{\gamma}, \tag{4.1}$$

where γ is usually 0.5. The clean speech estimate is then given by:

$$\widehat{S}(l,m) = \widehat{G}(l,m)Y(l,m)\,, \tag{4.2}$$

from which the time-domain estimated signal $\widehat{s}(k)$[1] is generated.

4.2.2 CRUSE architecture

We adopt the most efficient architecture reported by [4] for all our experiments. It is a 4-layer U-net with 1×1 *convolutional* skip connection. Other than replacing the grouped gated recurrent units (GRUs) in the bottleneck layer by a single GRU layer, other parameters are kept the same as suggested, and summarised in Table 4.1. The input to the network is the log power spectra of the noisy signal – as originally proposed. However, we modify the training target from the linear mask to the

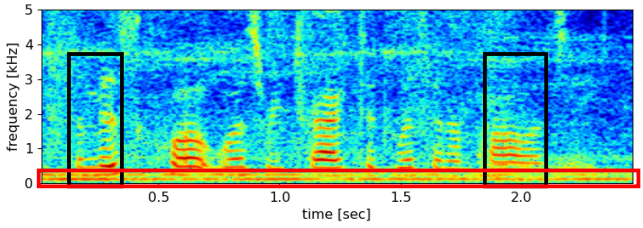
Table 4.1: Parameters of the CRUSE net. Since encoder and decoder have *symmetric* structures, only the encoder structure is enumerated.

Encoder parameters	
Channels	16, 32, 64, 128
Kernel size (Time, Frequency)	(2, 3) for all layers
Stride (Time, Frequency)	(1, 2) for all layers
Default activation functions	Leaky ReLU, slope$= 0.03$
Skip connection	1×1 convolution
Number of training epochs	10

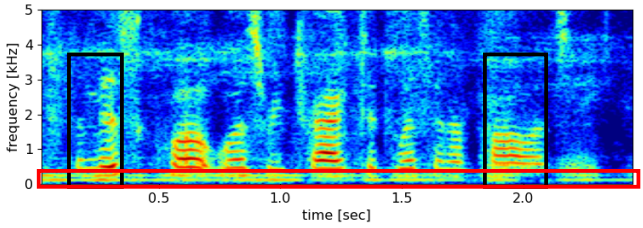
bounded *logarithmic* mask (log-mask) for a better capture of low mask values [9]. Accordingly, the activation function of the final layer of the decoder has been changed from the sigmoid function to a linear function with clipping between -40 and 3 dB. Empirically, we found no big difference between the network trained by the MSE of spectrum estimate and the MSE of log-mask estimate. Since a lower boundary of mask estimate benefits the enhanced speech quality, we choose to calculate the MSE loss on the log-mask, i.e., $L_{MSE}(l) = |\log\{G(l,m)\} - \log\{\widehat{G}(l,m)\}|^2$. The log-compression avoids the numerical issue triggered by small IRM values in the training stage, and accelerates learning, which finally

[1] We follow the convention that estimated values are noted by $\widehat{\cdot}$

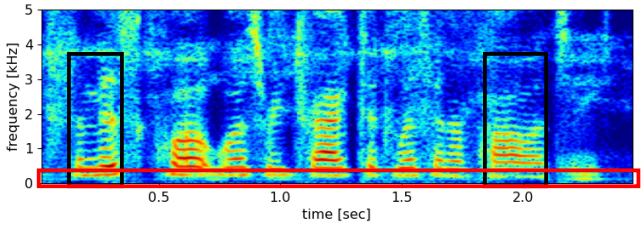
leads to a better inference. This is depicted in Figure 4.1, which demonstrates the speech estimate obtained by the same CRUSE architecture, trained with different targets. Comparing Figure 4.1b and 4.1c, it is clear that when using the MSE on the logarithm of the masks as a training target, the resulting model predicts a cleaner spectrum at low SNR regions (red box) and better inter-harmonic contrast (black boxes). Thus, we take the MSE on the log-mask as the *baseline* loss function.



(a) Noisy input



(b) Denoised by linear mask



(c) Denoised by log mask

Figure 4.1: Benefit of log-compressed mask. Noisy signal: kitchen noise, 5 dB. Log-compressed mask yields a cleaner estimate both in terms of noise suppression (red rectangle) as well as speech preservation (black boxes).

4.3 Exploitation of excitation information

4.3.1 F0 detection

To obtain the speech excitation information for the training first, the cepstral representation is first obtained from the spectra of the clean speech signals ($S(l,m)$) by a Q-point inverse discrete Fourier transform (iDFT):

$$c_s(l,q) = \text{IDFT}\big(\log(|S(l,m)|)\big) \tag{4.3}$$

where q is the quefrency bin index. Next, from the location of the cepstral peak, the F0-related quefrency bin $q_{\text{F0}}(l)$ in each frame is estimated. The peak-search is constrained within the range of allowable fundamental frequencies for human speech $m_0 \in [50, 300]$Hz, leading to a search range of $\mathcal{Q} \in \left\{ \lfloor \frac{f_s}{300} \rfloor, \ldots, \lfloor \frac{f_s}{50} \rfloor \right\}$ in quefrency ($\lfloor \cdot \rfloor$ denotes the floor function). Thus:

$$q_{\text{F0}}(l) = \underset{q}{\text{argmax}}\{c_s(l,q)\},\ q \in \mathbb{Q}\,, \tag{4.4}$$

Note: although the true $q_{\text{F0}}(l)$ can lie between two quefrency bins, our previous work [3] indicated that the cepstral peak is accurate enough for the purpose of improved excitation generation.

4.3.2 CEM loss function component

The amplitude of $c_s(l, q_{\text{F0}})$ indicates the strength of the voicing and, consequently, the harmonicity of the spectrum. Thus, the prediction error in harmonics can be indicated by the difference between the predicted and clean cepstra at $q_{\text{F0}}(l)$. To further emphasise the harmonics, we boost this cepstral coefficient by a factor $\beta(>1)$. Concisely:

$$\mathcal{L}_{\text{CEM}}(l) = \left|\beta \cdot c_{\text{target}}(l, q_{\text{F0}}(l)) - \widehat{c}(l, q_{\text{F0}}(l))\right|. \tag{4.5}$$

Note that c_{target} and $\widehat{c}$ can be calculated from the spectral estimate as: $c_{\text{target}} = \text{IDFT}\{\log(|S(l,m)|)\}$ and $\widehat{c} = \text{IDFT}\{\log(|\widehat{G}(l,m) \cdot Y(l,m)|)\}$. Or, it can also be computed on the predicted masks directly: $c_{\text{target}} = \text{IDFT}\{\log(G(l,m))\}$ and $\widehat{c} = \text{IDFT}\{\log(\widehat{G}(l,m))\}$.

As reported in [3], an adaptive amplifying factor performs better than the constant one. Therefore, if β in $\mathcal{L}_{\text{CEM}}(l)$ is related to the strength of the harmonics, the network might capture speech harmonics better. When a harmonic indicator $\eta(l) \in [0,1]$ is given for each frame, the amplifying factor can be decided by $\beta(l) = \eta(l) + 1$.

The final loss function is a weighted summation of the normal MSE and our CEM loss: $L(l) = \mathcal{L}_{\text{MSE}}(l) + \lambda\mathcal{L}_{\text{CEM}}(l)$.

4.3.3 Harmonic indicator

We employ the voicing detection scheme suggested in [10] as the harmonic indicator $\eta(l)$. The clean speech reference is low-pass filtered to the narrowband (0 to 4 kHz) first because voiced frames are more structured in this bandwidth. Then, the peak $\rho(l)$ of the normalised auto-correlation of the narrowband speech excitation signal in the allowable F0 range is taken as the detection feature. Using logistic regression, $\rho(l)$ is further converted to a harmonic indicator. Based on the distribution of auto-correlation function peaks, we choose the sigmoid parameters so that $\eta(l)$ is higher than 0.99 when $\rho(l) = 0.35$, and lower than 0.01 when $\rho(l) = 0.2$.

4.3.4 Summary

The proposed loss function component quantifies the prediction error in harmonic. For this purpose, the clean speech reference, the network output (mask prediction or clean speech estimate) and the training target are converted to the cepstrum. From the clean speech cepstrum, q_{F0} of each frame is identified (Eq. (4.4)). Then, the error is measured by the difference between the *boosted target* and the predicted output at this cepstral bin (Eq. (4.5)). By considering the loss in the cepstral domain, we implicitly impose a sinusoidal modelling of the harmonic spectrum. By amplifying the target value before loss computation, we not only target improved preservation of weak harmonics but also achieve better inter-harmonic noise suppression.

Compared to the harmonic weighted MSE loss proposed in [5], our method is more flexible in the choice of the training target and the major loss function. The harmonic weighted MSE loss is only designed for linear, STFT-domain loss, while ours can also be appended to time-domain loss functions. In addition, the implicit sinusoidal modelling of the harmonic spectrum allows boosting the harmonics in a controlled way (by choosing the amplifying factor β) to further improve inter-harmonic *contrast*.

4.4 Experimental evaluation

We choose the basic CRUSE net (trained with the MSE-loss on log-mask) as our baseline. In addition to that, we also train a CRUSE net with the harmonic-weighted loss function (dubbed 'HAMSE') of [5], where the threshold is set to $\theta = 0.4$ as suggested, the recursive smoothing factor $\alpha = 0.9$, and a searching frequency band of $K = 7$ bins for all frequencies. We verified that these parameters generate a reliable harmonic mask for clean speech. The harmonic weighted MSE loss is originally proposed for linear domain prediction, and we observe a similar improvement from the linear CRUSE baseline when replacing the normal MSE function with HAMSE. Yet, the performance of HAMSE on the linear CRUSE

is still much worse than the log-mask CRUSE baseline. Therefore, we skip this variant due to space limitations.

For the proposed loss function, the optimal parameters for different training targets are set empirically by trials and summarised in Table 4.2. Variants where the CEM loss is computed on the estimated masks are denoted as $\star$`maskCEM`. λ in Eq. (4.5) is chosen so that the CEM loss counts for around 10% of the total loss for a converged basic CRUSE. Note that even when CEM loss is calculated on the estimated speech $|\widehat{S}(l,m)|$, the MSE loss is still based on the log-mask.

Table 4.2: Experimented CEM loss parameters.

denotation	β	CEM loss on:
Basic	-	-
2CEM	2	$\lvert\widehat{S}(l,m)\rvert$
adpCEM	$1+\eta(l)$	$\lvert\widehat{S}(l,m)\rvert$
maskCEM	1.2	$\widehat{g}(l,m)$
adpmaskCEM	$1+\eta(l)$	$\widehat{g}(l,m)$
HAMSE	Harmonic weighted MSE loss as proposed in [5]	

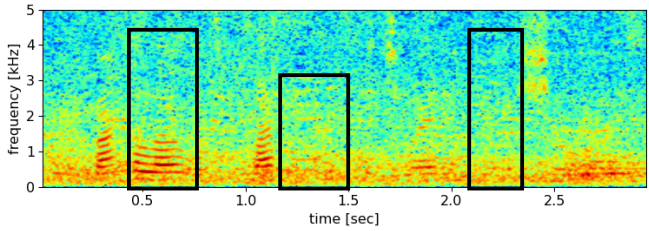
For training, the dataset is synthesised with TIMIT training set and ETSI noise at 6 SNRs: {-5 dB, 0 dB, 5 dB, 10 dB, 15 dB, 20 dB}, which generates about 66 hours of training audio in each SNR. The networks are evaluated on a fully *unseen* set, MS-SNSD, from Microsoft [11]. The MS-SNSD test set speech, and part of the training set and test set noise[2] are utilised. The noise signals are filtered using a second-order Butterworth highpass filter with cutoff frequency at 100 Hz before mixing. All samples are 16 kHz. We choose $M = \mathbb{Q} = Q = 512$, and a square-root von Hann window with 50% overlap as both the analysis and synthesis window.

4.5 Results and discussion

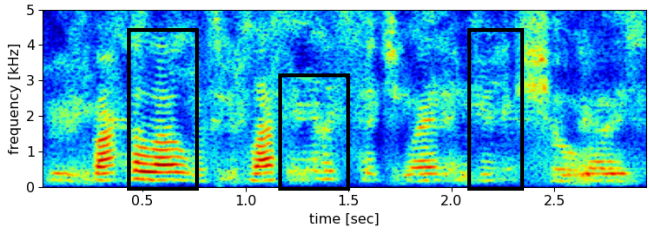
First, we present a set of spectrograms to intuitively demonstrate the benefit of the proposed CEM loss. Figure. 4.2 shows the enhanced signals by two nets (basic CRUSE and maskCEM CRUSE) for the speech mixed with cafe noise at 0 dB SNR. It may be seen that the net trained with the proposed loss function can better preserve the harmonic structure compared to the baseline.

For the comprehensive benchmark, we employ three widely-used objective metrics: the wide-band perceptual evaluation of speech quality (WB-PESQ) [12], scale-invariant signal-to-distortion ratio (SI-SDR) [13] and deep noise suppression mean opinion score (DNSMOS) [14]. We also present an evaluation on the basis

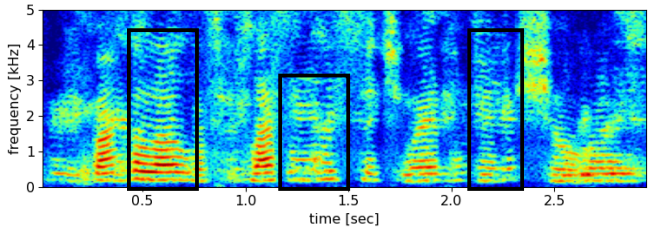
[2] `AirConditioner, Airport, Babble, Bus, Cafe, CafeTeria, Car, Field, Hallway, Kitchen, Metro, Square, VacuumCleaner`



(a) Noisy input



(b) Denoised by basic CRUSE



(c) Denoised by maskCEM CRUSE

Figure 4.2: Comparison of the denoised signals by the baseline CRUSE network and the net trained with mask-based cem loss function. Noisy signal: Cafe noise, 0 dB. Better speech harmonic structures can be observed in the signal processed by the proposed network in the framed areas.

of short-time objective intelligibility (STOI) [15]. However, as this modification is aimed at improving the preservation of speech harmonics (thus, mainly the quality), we do not expect major improvement on this metric, which primarily looks at *envelope* fidelity.

The evaluation results are grouped by input SNRs in Figure 4.3 where we see, firstly, that all methods improve the quality and intelligibility compared to the noisy input. The baseline method is already quite performant across all metrics. Compared to the baseline, the main benefit of the proposed loss function seems to be in the low SNRs (-5 dB, 0 dB, and 5 dB). The baseline is comparable to the proposed methods in higher SNRs. This is expected and in line with the previous observations in [3]. Our loss is designed for better capture of speech harmonics, which is

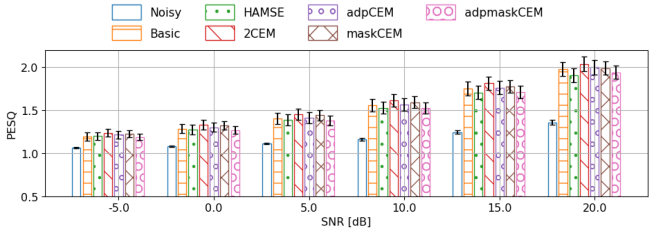
mostly impaired in low SNRs. The similar performance to baseline in high SNRs proves that the proposed loss function does not harm the generalisation ability of the networks. Among the CEM loss variants, the constant amplifying schemes yield better speech quality (PESQ, DNSMOS) than their adaptive counterparts (2CEM VS adpCEM, maskCEM VS maskadpCEM) - an interesting, counterintuitive result for which we have (as yet) no satisfactory explanation. We conjecture that the networks have difficulty in learning voiced/unvoiced discrimination when this is not explicitly incorporated in the loss function. Based on STOI scores, we note that there is no deterioration in the speech intelligibility in all benchmarked methods. Thus, the harmonic enhancement does not destroy envelope fidelity.

For the harmonic weighted MSE, although a larger improvement has been observed in the linear domain, its performance is not ideal on log-mask. Only DNSMOS shows improvement compared to the baseline net, which is comparable to adpCEM but not good as maskCEM in low SNRs, and the scores of other metrics even *deteriorate* at higher SNRs.

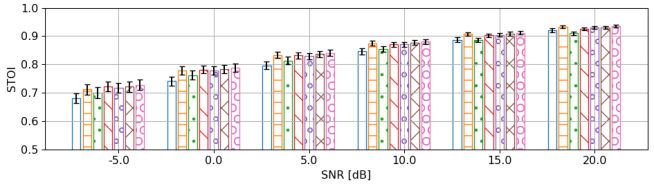
The overall best system is maskCEM, which demonstrates substantial scores on SI-SDR and DNSMOS metrics - indicating an evident improvement in speech quality.

4.6 Conclusions

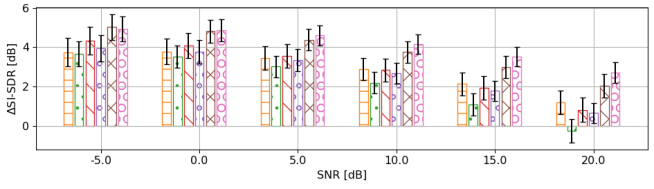
We proposed a loss function component, CEM loss, for the improved preservation of harmonic structure during speech enhancement. By quantifying the estimation errors of harmonic frequencies in the cepstral domain, a compact representation is obtained - which not only allows for better harmonic reconstruction but also permits improved inter-harmonic noise suppression. Incorporating this loss component results in better preservation of speech harmonics, especially in low SNR regions, where weak harmonics are prone to be lost. The improvement is reflected in both audio quality and the improved objective metrics in the comprehensive evaluation. Audio samples are available at: https://yanjuesong.github.io/CEM-loss-samples/, and we encourage the reader to judge the improvements for themselves. We are also open to feedback on this.



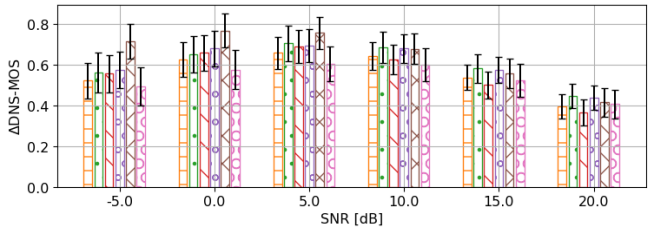
(a) PESQ



(b) STOI



(c) ΔSI-SDR



(d) ΔDNSMOS (overall)

Figure 4.3: Evaluation of PESQ, STOI, SI-SDR, and DNSMOS-overall, grouped by input SNRs. Averaged on the whole testset, and 95% confidence interval is given by the error bars.

References

[1] Y. Ephraim and D. Malah. *Speech enhancement using a minimum mean-square error log-spectral amplitude estimator*. IEEE transactions on acoustics, speech, and signal processing, 33(2):443–445, 1985.

[2] S. Elshamy, N. Madhu, W. Tirry, and T. Fingscheidt. *Instantaneous a priori SNR estimation by cepstral excitation manipulation*. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(8):1592–1605, 2017.

[3] Y. Song and N. Madhu. *Improved CEM for speech harmonic enhancement in single channel noise suppression*. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30:2492–2503, 2022.

[4] S. Braun, H. Gamper, C. K. Reddy, and I. Tashev. *Towards efficient models for real-time deep noise suppression*. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 656–660. IEEE, 2021.

[5] N. Raviv, O. Schwartz, and S. Gannot. *Low Resources Online Single-Microphone Speech Enhancement with Harmonic Emphasis*. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 8807–8811. IEEE, 2022.

[6] A. Pandey and D. Wang. *A new framework for CNN-based speech enhancement in the time domain*. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27(7):1179–1188, 2019.

[7] K. Tan and D. Wang. *A convolutional recurrent neural network for real-time speech enhancement*. In Interspeech, volume 2018, pages 3229–3233, 2018.

[8] S. Elshamy and T. Fingscheidt. *Improvement of Speech Residuals for Speech Enhancement*. In 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pages 219–223. IEEE, 2019.

[9] A. Bohlender, A. Spriet, W. Tirry, and N. Madhu. *Neural networks using full-band and subband spatial features for mask based source separation*. In European Signal Processing Conf. (EUSIPCO), pages 346–350, 2021.

[10] N. Madhu and M. Krini. *Spectral refinement with adaptive window-size selection for voicing detection and fundamental frequency estimation*. In 2020 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), pages 1–6. IEEE, 2020.

[11] C. K. Reddy, E. Beyrami, J. Pool, R. Cutler, S. Srinivasan, and J. Gehrke. *A Scalable Noisy Speech Dataset and Online Subjective Test Framework*. Proc. Interspeech 2019, pages 1816–1820, 2019.

[12] ITU-T. *Rec. P.862.2: Corrigendum 1, Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs*. International Telecommunication Union-Telecommunication Standardisation Sector, 2017.

[13] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey. *SDR–half-baked or well done?* In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 626–630. IEEE, 2019.

[14] C. K. Reddy, V. Gopal, and R. Cutler. *DNSMOS P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors*. In ICASSP, 2022.

[15] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen. *A short-time objective intelligibility measure for time-frequency weighted noisy speech*. In 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, pages 4214–4217. IEEE, 2010.

5

Phase reconstruction in single channel speech enhancement based on phase gradients and estimated clean-speech amplitudes

All investigations so far focus on magnitude estimation for speech enhancement. Phase enhancement is known to be a challenge in single channel scenario. With deep neural networks (DNNs), one straightforward solution is to extend real training targets of the networks to the complex domain. However, since phase spectrogram is sensitive to time shift and shows no discernable pattern, the effect of DNNs in the complex domain remains debatable. In this chapter, we try to address this challenge based on phase derivatives. Although the phase spectrogram shows no pattern, its difference along the time- or frequency axis has a clear structure, and is related to the logarithm of the magnitude. This serves as the foundation of derivative-based phase retrieval solutions, which aims to find a matched phase for a known, clean speech magnitude. To the best of our knowledge, this relationship is not widely explored in speech enhancement.

Since this method requires clean speech magnitudes to retrieve a matched phase, we utilise the DNNs trained in the previous chapter to obtain the clean magnitude estimates as an approximation, based on which a matched phase can be obtained by the derivative-based phase retrieval algorithm. Two problems, however, remain to be solved. First, a direct combination of the estimated magnitude

and the retrieved phase sounds artificial. Considering that the noisy phase has been widely accepted as a reliable phase estimate, this contrast indicates that the noisy phase is, in fact, an effective estimate, especially in high signal-to-noise ratio (SNR) regions. Therefore, we propose to fuse the retrieval result with the noisy phase to reconstruct a balanced phase estimate. Secondly, the influence of distorted magnitude spectrograms on phase derivative prediction networks is unclear, which is investigated by our ablation studies.

⋆ ⋆ ⋆

Y. Song, and N. Madhu.

Published in the proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), April 2024.

Abstract Phase *gradients* can help enforce phase *consistency* across time and frequency, further improving the output of speech enhancement approaches. Recently, neural networks were used to estimate the phase gradients from the short-term amplitude spectra of *clean* speech. These were then used to synthesise phase to obtain a plausible time-domain signal. However, using purely synthetic phase in speech enhancement yields unnatural-sounding output. Therefore we derive a closed-form phase estimate that combines the synthetic phase with that of the enhanced speech, yielding more natural output. Secondly, we empirically evaluate the benefit of (re-)training the phase gradient estimation networks on the amplitude spectra of the *estimated* clean-speech signal. Lastly we apply our proposed phase enhancement to the output of a *phase-aware* speech enhancement DNN, verifying if an *independent* phase estimator brings additional advantage. Results show that, compared to the baseline, the proposed approach further improves the DNSMOS scores by ≈ 0.1 on average, and significantly in the first quartile on broadband, quasi-stationary noises, where phase enhancement is expected to have maximum benefit. Training phase gradient estimators on estimated speech spectra is additionally beneficial here. Our method even improves the performance of the phase-aware approach, indicating its feasibility as a generic post-processor for speech enhancement.

5.1 Introduction

Phase estimation of the underlying speech signal is a challenging problem in single-channel speech enhancement (SE), because of the lack of discernible structure in phase spectra and the added complication of the 2π periodicity. Classical speech enhancement in the short-time Fourier transform (STFT) domain, therefore, focussed on the estimation of the clean speech amplitude and retained the phase of

the noisy input. Typically this was justified by assuming a uniform distribution for phase in $[0, 2\pi]$ [1], or by assuming a (complex) Gaussian distribution for the signal [2], whereby the noisy phase was the minimum mean square error (MMSE) optimal estimate. Yet, as discussed in [3], the importance of phase estimation in speech enhancement cannot be ignored.

Prior work: Interestingly, while the phase spectra do not show structure, the phase *derivatives* demonstrate clear patterns. This was first systematically studied and exploited in [4, 5] to blindly estimate the phase of speech *harmonics* during voiced speech. The core idea here was to enforce the theoretically derived phase gradient across time frames, for all frequency bins containing speech harmonics, and to ensure consistency across *frequency* by factoring in the influence of the window function. However, the approach requires a robust estimate of the fundamental frequency (F0) in each frame. Unfortunately, F0 estimation accuracy degrades at low signal-to-noise ratios (SNRs), rendering the approach less effective at exactly the conditions where phase enhancement would be most beneficial [6]. Secondly, the underlying assumption is that the spectrum is purely harmonic during voiced speech. However, when we have mixed excitation or spectra where higher harmonics are not integer multiples of F0, phase reconstruction leads to a metallic-sounding output due to over-excitation.

Similar to classical approaches, deep neural network (DNN)-based SE methods operating in the STFT domain either estimate a denoising mask or perform spectral mapping to clean speech amplitudes. To include phase, mask estimation has been extended into the complex domain as, e.g., the complex ideal ratio mask [7] or phase sensitive mask [8]. Spectral mapping has been similarly extended to estimate the complex coefficients of the clean speech in the rectangular [9] or polar form [10, 11]. However, as shown in [10], the complex extensions make only small changes to the phase, compared to the magnitude-only methods, implying that the DNNs do not sufficiently learn the distribution of the clean speech phase, but likely achieve a local optimum in the MMSE sense, as in the statistical methods.

STFT-based methods are attractive for practical applications, due to their interpretability and tunability. To solve the associated phase enhancement problem, therefore, the use of temporal and spectral phase derivatives (gradients) to estimate the phase of clean speech have come under renewed focus. Specifically interesting - as they are closely related to the denoising problem - are approaches that can estimate the phase given only the *clean* magnitude spectra. Such approaches are based on implicit relations between the spectral and temporal phase gradients and the STFT amplitude spectra. In [12], phase is retrieved in a non-causal and purely signal theoretic framework, by integrating the gradients across time and frequency. In contrast, in [13, 14], DNNs are first trained to predict the mapping from amplitude spectra of clean speech to the phase derivatives along time and frequency. At inference, the phase at any particular STFT bin is obtained by integrating the

estimated phase derivatives along time or frequency. The estimated phases from each direction of integration are fused to obtain a single, consistent phase value for that STFT bin, whereas [13] heuristically weights the estimate from each direction, in proportion to the magnitude along that path. [14] proposes an elegant, analytic solution by posing the fusion as an optimisation problem, to be solved independently at each STFT bin.

Contributions: We first note that the goal in [14] is to *synthesise* the phase of clean speech, given its *clean* STFT amplitudes. This is the inspiration for our work. Yet, in speech enhancement, the audio sounds too robotic when applying the fully synthetic estimate to the estimated magnitude. We derive a closed-form phase estimate, combining the synthetic phase with that of the enhanced speech, leading to more natural output. Secondly, whereas the phase-gradient estimators of [14] are trained on amplitude spectra of clean speech, we empirically evaluate the benefit of training these estimators in matched conditions, using the *estimated* amplitude spectra. Lastly, we study the added value of our proposed approach in *phase-aware* enhancement approaches. This would indicate the feasibility of the approach as a generic post-processor in STFT-based speech enhancement.

The paper is organised as follows: the signal model and the baseline enhancement networks are discussed in Sec. 5.2, along with the concept of phase derivatives. Next, we briefly summarise the approach of [14] and derive the analytic solution for the phase estimate in the speech enhancement context (Sec. 5.3). Following this, we evaluate the proposed approach on the DNS challenge test set and, thereby, also draw conclusions on the questions raised previously. The key take-aways are summarised in the conclusion.

5.2 STFT domain speech enhancement

Assuming an additive mixing model at the microphone, the observed signal is represented in the STFT domain as:

$$Y(l,m) = H_{\mathrm{RIR}}(m)S(l,m) + V(l,m) \tag{5.1}$$

where the clean speech $S(l,m)$ is degraded by the background noise $V(l,m)$ and possible reverberation introduced by the room transfer function $H_{\mathrm{RIR}}(m)$. The l and m are the frame index and the frequency bin index, respectively. Speech enhancement is obtained by estimating a time-frequency (TF) mask (or *gain*) $G(l,m)$ which, applied to $Y(l,m)$, yields the clean speech spectrum estimate:

$$\widehat{S}(l,m) = G(l,m)Y(l,m)\,, \tag{5.2}$$

from which the time-domain signal $\widehat{s}(k)$ is obtained by inverse Fourier transform and overlap-add. The estimated mask, $G(l,m)$, can be real-valued or complex-

valued (phase-aware extensions). In the former case, the noisy phase is used in the final speech estimate.

5.2.1 DNN baselines: CRUSE and Complex CRUSE (C-CRUSE)

Convolutional, recurrent encoder-decoder networks with skip connections (commonly called UNets) are widely used for single- and multi-channel speech enhancement in the STFT domain as they offer a good balance between computational efficiency and performance. Specifically, we select convolutional recurrent U-net architecture for speech enhancement (CRUSE) [15] as the baseline for predicting the real-valued TF mask from the noisy magnitude.

For the *phase-aware* baseline, we extend CRUSE as follows: the input features are formed by concatenating the real and imaginary parts of the noisy spectrogram along the channel dimension. Two output channels are obtained containing, respectively, the real and imaginary part of the complex mask. The final mask is then: $G(l,m) = G_R(l,m) + jG_I(l,m)$. To allow phase to be modelled in the full range $[0, 2\pi]$, the hyperbolic tangent activation function is used in the final layer. We dub this extension as C-CRUSE.

5.2.2 Phase derivatives

In the polar form, the complex spectrogram, $S(l,m)$, of clean speech can be written in terms of the amplitude $A(l,m)$ and phase $\Phi(l,m)$ as:

$$S(l,m) = A(l,m)\exp\left(j\Phi(l,m)\right).$$

The phase derivative along frequency and time is then approximated as:

$$\Delta_f\Phi(l,m) = \Phi(l,m) - \Phi(l,m-1) \quad \text{and} \tag{5.3}$$

$$\Delta_t\Phi(l,m) = \Phi(l,m) - \Phi(l-1,m)\,. \tag{5.4}$$

As the STFT $S(l,m)$ is computed on windowed, overlapped, time-segments, there is an additional offset term in the phase that is proportional to the frame shift (N^{hop}). Because of the 2π periodicity, this offset can distort the structure in the temporal phase difference. To avoid this, [5] proposes to modulate $S(l,m)$ into the baseband. If M is the frame length, the baseband-modulated phase $\Psi(l,m)$ is given by:

$$\Psi(l,m) = \Phi(l,m) + \psi_0(l,m)\,, \tag{5.5}$$

with $\psi_0(l,m) \equiv -2\pi lm\frac{N^{\text{hop}}}{M}$. The *baseband* phase difference is then:

$$\Delta_t\Psi(l,m) = \Psi(l,m) - \Psi(l-1,m)\,. \tag{5.6}$$

Note that it is easy to compute $\Delta_t\Phi(l,m)$ from the baseband phase difference $\Delta_t\Psi(l,m)$. Finally, if either $\Phi(l,m-1)$ or $\Phi(l-1,m)$ are available and the

corresponding phase differences from (5.3) and (5.4) can be estimated, $\Phi(l,m)$ can be computed.

5.3 Phase estimation

5.3.1 Estimating $\Delta_f\Phi(l,m)$ and $\Delta_t\Psi(l,m)$

As shown in [12], phase derivatives are connected to the log magnitude spectra. While it is possible to estimate these derivatives analytically, it would require limiting assumptions on the evolution of phase across time and frequency and exhibit the same drawbacks as [5]. In contrast, data-driven methods [13, 14], where DNNs are utilised to learn this relationship, offer more robust estimates.

Two separate UNets are used to predict the two phase differences, respectively. The details of the networks are provided in section 5.4. Training targets are the phase differences of clean speech, $\Delta_f\Phi(l,m)$ and $\Delta_t\Psi(l,m)$. Due to the periodic nature of phase, cosine loss functions, defined below, are adopted.

$$\mathcal{L}_f = \sum_{l,m} \left(1 - \cos\left(\widehat{\Delta_f\Phi}(l,m) - \Delta_f\Phi(l,m)\right)\right) \tag{5.7a}$$

$$\mathcal{L}_t = \sum_{l,m} \left(1 - \cos\left(\widehat{\Delta_t\Psi}(l,m) - \Delta_t\Psi(l,m)\right)\right) \tag{5.7b}$$

5.3.2 Phase retrieval from clean speech amplitudes

As independent networks are employed to predict the phase differences across time and frequency, we obtain two estimates of $\Phi(l,m)$ - one for each integration path. A final, consistent phase estimate is obtained by fusing the individual results. This is formulated very elegantly in [14] as an optimisation problem, allowing $\Phi(l,m)$ to be computed recursively and in an analytical manner, once the phase differences themselves are estimated. We briefly summarise this, using the same notation as [14], before deriving our extension.

Define $V(l,m) \equiv \frac{S(l,m)}{S(l-1,m)}$, which, clearly, is linked to the temporal phase difference. Inserting $\widehat{\Delta_t\Phi}(l,m)$, and the *known* clean speech amplitudes $A(l,m)$, we obtain an estimate of $V(l,m)$ as:

$$\widehat{V}(l,m) = \frac{A(l,m)}{A(l-1,m)} \exp\left(j\,\widehat{\Delta_t\Phi}(l,m)\right). \tag{5.8}$$

Denote by $\mathbf{z}_l = \left[z(l,0), z(l,1), \ldots, z(l,M')\right]^T$, the complex spectrum *estimate* of clean speech for the $M' = M/2$ positive frequencies of frame l. The reason for using $z(l,m)$ instead of $S(l,m)$ will become clear presently. Given the clean

speech spectral estimate $\widehat{\mathbf{S}}_{l-1} = [\widehat{S}(l-1,0), \widehat{S}(l-1,1), \ldots, \widehat{S}(l-1,M')]^T$ of the prior frame, $\mathbf{z}_l$ can be obtained by minimising:

$$\mathcal{J}_t(\mathbf{z}_l, \widehat{\mathbf{S}}_{l-1}, \widehat{\mathbf{V}}_l) = \left\|\mathbf{z_l} - \widehat{\mathbf{V}}_l \odot \widehat{\mathbf{S}}_{l-1}\right\|^2_{\mathbf{\Lambda}_l}. \tag{5.9}$$

$\widehat{\mathbf{V}}_l = [\widehat{V}(l,0), \widehat{V}(l,1), \ldots, \widehat{V}(l,M')]^T$, and $\odot$ is the Hadamard product. The notation $\left\|\mathbf{e}\right\|^2_{\mathbf{\Lambda}}$ is the weighted inner product: $\mathbf{e}^H\mathbf{\Lambda}\mathbf{e}$.

To obtain $\mathbf{z}_l$ from the phase gradient across *frequency*, we define $U(l,m) \equiv \frac{S(l,m)}{S(l,m-1)}$. Given $\widehat{\Delta_f \Phi}(l,m)$, we estimate $U(l,m)$ as:

$$\widehat{U}(l,m) = \frac{A(l,m)}{A(l,m-1)} \exp\left(j\,\widehat{\Delta_f \Phi}(l,m)\right). \tag{5.10}$$

Defining the sparse $(M'-1) \times M'$ matrix $\mathbf{D}_l$ as:

$$D_l(m,m') = \begin{cases} -\widehat{U}(l,m+1) & m' = m \\ 1 & m' = m+1 \\ 0 & \text{otherwise} \end{cases}, \tag{5.11}$$

it is clear that $\mathbf{z}_l$ can be estimated by minimising:

$$\mathcal{J}_f(\mathbf{z}_l, \widehat{\mathbf{U}}_l) = \left\|\mathbf{D}_l\mathbf{z}_l\right\|^2_{\mathbf{\Gamma}_l}. \tag{5.12}$$

Minimising the combined cost functions in (5.9) and (5.12) :

$$\mathcal{J}(\mathbf{z}_l) = \left\|\mathbf{z}_l - \widehat{\mathbf{V}}_l \odot \widehat{\mathbf{S}}_{l-1}\right\|^2_{\mathbf{\Lambda}_l} + \left\|\mathbf{D}_l\mathbf{z}_l\right\|^2_{\mathbf{\Gamma}_l}. \tag{5.13}$$

yields $\widehat{\mathbf{z}}_l = \left(\mathbf{\Lambda}_l + \mathbf{D}_l^H\mathbf{\Gamma}_l\mathbf{D}_l\right)^{-1}\mathbf{\Lambda}_l\left(\widehat{\mathbf{V}}_l \odot \widehat{\mathbf{S}}_{l-1}\right)$, which is the jointly optimal estimate. From this, the phase estimate is obtained as:

$$\widehat{\Phi}(l,m) = \angle\widehat{z}(l,m). \tag{5.14}$$

where $\angle z$ calculates the phase of a complex coefficient z. This is combined with the amplitude $A(l,m)$ to yield $\widehat{S}(l,m)$.

The weights $\mathbf{\Lambda}$ and $\mathbf{\Gamma}$ indicate the reliability of the two predicted phase differences. Since the network prediction accuracy is related to the spectral magnitude, it is proposed in [14] to define diagonal weighting matrices as:

$$\mathbf{\Lambda}_l(i,i) = (A(l-1,i)A(l,i))^p \quad \text{and} \tag{5.15}$$

$$\mathbf{\Gamma}_l(i,i) = \gamma \cdot (A(l,i)A(l,i+1))^p, \tag{5.16}$$

where p is the magnitude compression factor, and γ is the extra factor to balance two different estimates.

5.3.3 Phase reconstruction for speech enhancement

When applying the above approach for estimating the phase for speech enhancement, two points should be considered: first, the magnitude spectra used for predicting the phase gradients and estimating the phase are obtained from the preceding speech enhancement system. Thus, they are imperfect and possibly contain artefacts due to residual noise and speech distortion. Hence, it might be advantageous to train the DNNs for phase gradient estimation on these estimated speech amplitudes. Second, using the purely synthetic estimate from (5.14) leads to an output that sounds unnatural compared to the speech in the noisy mixture. Therefore, to obtain natural-sounding audio, the initial phase available from the speech enhancement stage should be incorporated in the phase estimator. We propose to do this by including an additional term in the cost function in (5.13), which penalises large deviations from the estimated speech spectrum obtained after the speech enhancement stage. Denoting the enhanced speech at frame l from the baseline speech enhancement by $\widetilde{\mathbf{S}}_l = [\widetilde{S}(l,0), \widetilde{S}(l,1), \ldots, \widetilde{S}(l,M')]^T$, the *spectral deviation* cost to be added to (5.13) can be expressed as:

$$\mathcal{J}_s(\mathbf{z}_l, \widetilde{\mathbf{S}}_l) = \left\|\mathbf{z}_l - \widetilde{\mathbf{S}}_\mathbf{l}\right\|^2_{\mathbf{\Omega}_l}. \tag{5.17}$$

Since only the current frame is relevant to this distance, we propose to construct $\mathbf{\Omega}_l$ as the diagonal matrix:

$$\mathbf{\Omega}_l(i,i) = \omega(\widetilde{A}(l,i))^{2p}, \tag{5.18}$$

consistent with the definition of $\mathbf{\Lambda}$ and $\mathbf{\Gamma}$. Further ω is a hyperparameter to adjust the contribution of this cost component. This leads to the following estimate of $\mathbf{z}_l$ in the context of speech enhancement:

$$\widehat{\mathbf{z}}_l = \left(\mathbf{\Lambda}_l + \mathbf{D}_l^H \mathbf{\Gamma}_l \mathbf{D}_l + \mathbf{\Omega}_l\right)^{-1} \left(\mathbf{\Lambda}_l\left(\widehat{\mathbf{V}}_l \odot \widehat{\mathbf{S}}_{l-1}\right) + \mathbf{\Omega}_l \widetilde{\mathbf{S}}_l\right) \tag{5.19}$$

Computing the enhanced phase as in (5.14), we obtain the final clean speech estimate as:

$$\begin{aligned} \widehat{S}(l,m) &= \left|\widetilde{S}(l,m)\right| \exp\left(j\,\widehat{\Phi}(l,m)\right) \\ &= \left|G(l,m)Y(l,m)\right| \exp\left(j\,\widehat{\Phi}(l,m)\right) \end{aligned} \tag{5.20}$$

The system is summarised by the block diagram in Figure 5.1.

5.4 Experimental evaluation and discussion

For CRUSE and C-CRUSE, we adopt the four layer encoder-decoder structure with grouped gated recurrent units (GRUs) (true to [15]), to predict the mask for

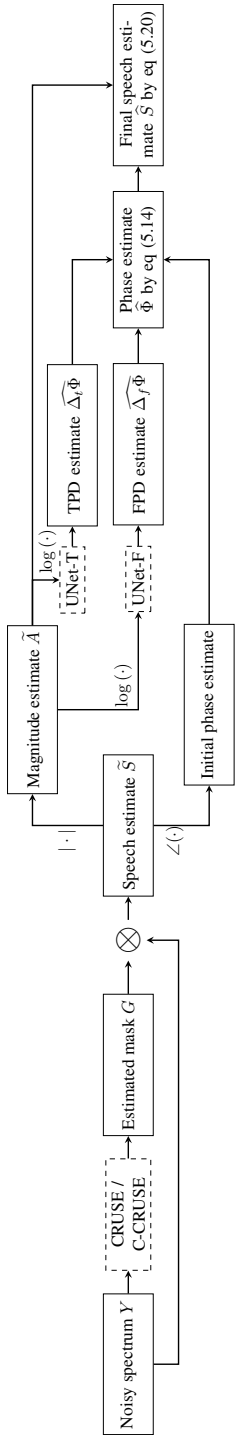


Figure 5.1: Block diagram of the proposed speech enhancement system with phase reconstruction. Dashed boxes represent neural networks whereas solid boxes indicate data contained. The frame index l and the frequency bin index m have been dropped for conciseness.

the initial noise reduction and dereverberation. The networks were trained on the DNS challenge 2021 wideband dataset [16]. We synthesised 140 hours of training data from English speech, with 50% of them in reverberant conditions ($T_{60} \in [0.3$ s – 1.3 s$]$). The SNR of the training set varied between -5 dB and 20 dB. Audio was sampled at 16 kHz. The STFT employs 75% overlapped frames and a square-root Hann window of length $M = 512$ for analysis & synthesis. The enhanced speech signals were evaluated by segmental SNR [17], short-time objective intelligibility (STOI) [18], and DNSMOS P.835 [19].

The UNets for phase gradient estimation comprise three convolutional layers in the encoder and decoder, respectively. Kernels of dimension 2×3 (time, frequency), and strides of 1×2 were used at all layers. The number of channels of each layer were: $16, 32, 32$, which resulted in a 992 unit fully connected layer at the bottleneck. All convolutional layers were followed by the leaky ReLU function with an $\alpha = 0.003$ negative slope. Two sets of UNets were trained: 1) the first set, as originally proposed for phase retrieval, learn the relationship between the *clean* magnitude spectra and the phase derivatives. They are agnostic of the speech enhancement stage; 2) the second set is adapted to the speech enhancement context, and learn to predict the phase derivatives of the clean speech from the *estimated* magnitude spectra. Both approaches have unique potential advantages: enhancement-agnostic networks need no retraining when switching other components in the pipeline, while networks trained specific to a certain speech enhancement system might provide better performance, due to matched conditions. We denote the networks as 'SE-Agnostic' and 'SE-Matched' in the sequel.

Optimal values for the compression factor p and the weights γ and ω, were obtained by grid search. Based on the deep noise suppression mean opinion scores (DNSMOSs) on a subset (with wideband, quasi-stationary noise) of development data of the DNS2020 challenge [20], the optimal parameters were $[p = 0.3, \gamma = 10, \omega = 5]$ for SE-Agnostic, and $[p = 0.5, \gamma = 10, \omega = 5]$ for SE-Matched.

5.4.1 Results & Discussion

The optimised system is evaluated on the DNS2021 synthetic test set [16]. Averaged metrics are given in Table 5.1. Compared to the noisy input, a significant improvement is provided by all approaches, and on all metrics. As an *upper bound* on achievable performance, we also present results where the *oracle, clean* phase is used with the estimated speech amplitudes. Comparing CRUSE and C-CRUSE, we see a marginal benefit of estimating the phase in the enhancement stage - in line with previous works.

More importantly, the proposed phase enhancement improves all quality metrics, compared to *both* corresponding baselines (CRUSE & C-CRUSE). Only STOI is constant for all approaches. This is expected: phase enhancement should not

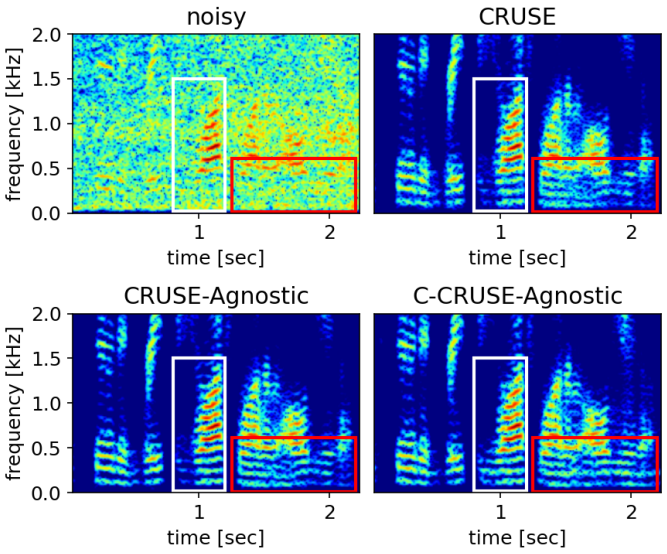


Figure 5.2: Comparison of the denoised signals by CRUSE and with the proposed phase reconstruction. Noisy signal: Street noise, -2 dB. Note the clearer harmonic structure after phase reconstruction by the proposed method in the highlighted area. C-CRUSE in combination with phase reconstruction even manages to pick up very weak harmonic structure (white box) and gives a more continuous harmonic spectrum (red box)

affect the speech *envelope*, which is important for intelligibility. This *sanity check* ensures we do not improve quality at the cost of intelligibility. We also see that the phase-enhanced outputs are *comparable* to using oracle phase – a pleasing result.

We expect the proposed phase reconstruction to offer maximum benefit with stationary, broadband noise conditions, as such noise typically results in vocoding artefacts between the harmonics after speech enhancement – which phase reconstruction can ameliorate. In such cases, we expect differences between the various configurations to be more evident. Thus, we split the test set into two subsets: **a)** mixtures with stationary or short-term stationary noise, such as car, traffic, babble; **b)** mixtures with sparse, transient noise, such as footsteps, typing, etc. The distribution of the DNSMOS scores are shown in Figure 5.3 for both subsets. We now see that on *subset a*, SE-Matched has a bigger margin over the SE-Agnostic. When the noise is less stationary and sparse (*subset b*), using SE-Agnostic is better. We reason that in such cases there are fewer contiguous regions where the speech and noise overlap. Then, SE-Agnostic networks, being trained on clean speech, yield more accurate phase estimates. The averaged results in Table 5.1 may indicate only a small achievable improvement by using the proposed phase reconstruction

Table 5.1: Averaged instrumental metrics on test set. Best results in bold. Phase enhancement consistently improves all metrics compared to baselines and is comparable to using *oracle* phase.

Method	segSNR [dB]	STOI	DNSMOS	
			OVRL	SIG
Noisy	6.87	0.87	2.53	3.33
CRUSE	13.74	0.93	3.10	3.36
CRUSE-Agnostic	14.30	0.93	3.17	3.43
CRUSE-Matched	14.19	0.93	3.17	3.44
C-CRUSE	13.92	0.93	3.14	3.40
C-CRUSE-Agnostic	**14.45**	0.93	**3.20**	**3.45**
CRUSE-OraclePhase	14.51	0.94	3.17	3.43
C-CRUSE-OraclePhase	14.77	0.94	3.20	3.45

on CRUSE/C-CRUSE. However, Figure 5.3 shows that phase reconstruction manages to boost the signal quality in poor SNR conditions – reflected by the decreased spread and higher minimum in the scores!

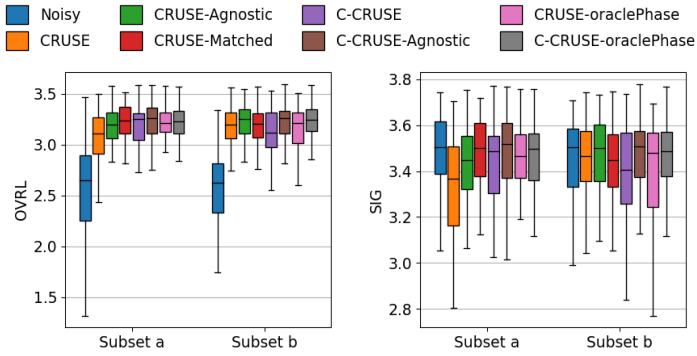


Figure 5.3: DNSMOS score distribution, separately on broadband/quasi-stationary and transient/sparse noise subsets from DNS 2021 test set.

5.5 Conclusions

We proposed a phase reconstruction method for speech enhancement, based on phase gradients. Using independent DNNs to predict spectral and temporal phase derivatives from the estimated amplitude spectra (from a preceding speech enhancement stage), we obtain two estimates of the phase. A closed-form, analytic solution was derived to fuse these estimates in an MMSE-optimal manner. We further introduced an additional cost term that incorporated phase informa-

tion present in signal after the speech enhancement stage – which led to a more natural-sounding output. Experimental results validate the quality improvement brought by the proposed phase enhancement - with the performance of the proposed method being comparable to using oracle phase. The proposed phase estimator is also beneficial when used with phase-aware speech enhancement, indicating its feasibility as a generic post-processor in STFT-based speech enhancement frameworks. Lastly, training the phase derivative estimator DNNs specific to the preceding speech enhancement stage is beneficial when noise is (short-term) stationary and broadband. For sparse, transient noises, training the DNNs on clean-speech spectra gives more accurate results. Audio samples can be found at https://aspireugent.github.io/diff-based-phase-reconstruction-SE/.

References

[1] P. Vary and R. Martin. *Digital speech transmission: Enhancement, coding and error concealment*. John Wiley & Sons, 2006.

[2] Y. Ephraim and D. Malah. *Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator*. IEEE Transactions on acoustics, speech, and signal processing, 32(6):1109–1121, 1984.

[3] K. Paliwal, K. Wójcicki, and B. Shannon. *The Importance of Phase in Speech Enhancement*. Speech Communication, 53(4):465–494, apr 2011.

[4] T. Gerkmann, M. Krawczyk, and R. Rehr. *Phase estimation in speech enhancement—unimportant, important, or impossible?* In IEEE 27th Convention of Electrical and Electronics Engineers in Israel, pages 1–5. IEEE, 2012.

[5] M. Krawczyk and T. Gerkmann. *STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement*. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 22(12):1931–1940, 2014.

[6] P. Vary. *Noise suppression by spectral magnitude estimation -mechanism and theoretical limits*. Signal Processing, 8(4):387–400, 1985.

[7] D. S. Williamson, Y. Wang, and D. Wang. *Complex ratio masking for monaural speech separation*. IEEE/ACM transactions on audio, speech, and language processing, 24(3):483–492, 2015.

[8] M. Kolbæk, D. Yu, Z. Tan, and J. Jensen. *Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks*. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(10):1901–1913, 2017.

[9] K. Tan and D. Wang. *Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement*. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28:380–390, 2019.

[10] D. Yin, C. Luo, Z. Xiong, and W. Zeng. *Phasen: A phase-and-harmonics-aware speech enhancement network*. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 9458–9465, 2020.

[11] A. Li, C. Zheng, G. Yu, J. Cai, and X. Li. *Filtering and refining: A collaborative-style framework for single-channel speech enhancement*. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30:2156–2172, 2022.

[12] Z. Průša, P. Balazs, P. Søndergaard, and L. Peter. *A noniterative method for reconstruction of phase from STFT magnitude*. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(5):1154–1164, 2017.

[13] L. Thieling, D. Wilhelm, and P. Jax. *Recurrent phase reconstruction using estimated phase derivatives from deep neural networks*. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 7088–7092. IEEE, 2021.

[14] K. N. Y. Masuyama, K. Yatabe and Y. Oikawa. *Online phase reconstruction via DNN-based phase differences estimation*. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31:163–176, 2022.

[15] S. Braun, H. Gamper, C. K. Reddy, and I. Tashev. *Towards efficient models for real-time deep noise suppression*. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 656–660. IEEE, 2021.

[16] C. K. Reddy, H. Dubey, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan. *ICASSP 2021 deep noise suppression challenge*. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6623–6627. IEEE, 2021.

[17] K. Eneman, A. Leijon, S. Doclo, A. Spriet, M. Moonen, and J. Wouters. *Auditory-profile-based Physical Evaluation of Multi-microphone Noise Reduction Techniques in Hearing Instruments*. In Advances in Digital Speech Transmission, pages 431 – 458. John Wiley & Sons Ltd., New York, 2008.

[18] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen. *A short-time objective intelligibility measure for time-frequency weighted noisy speech*. In 2010 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 4214–4217. IEEE, 2010.

[19] C. K. Reddy, V. Gopal, and R. Cutler. *DNSMOS P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors*. In 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021.

[20] C. K. Reddy, V. Gopal, R. Cutler, R. C. E. Beyrami, H. Dubey, S. Matusevych, A. A. R. Aichner, S. Braun, et al. *The INTERSPEECH 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results*. arXiv preprint arXiv:2005.13981, 2020.

6

On the Disentanglement and Robustness of Self-Supervised Speech Representations

In the previous chapters, we have explored how to integrate domain knowledge expressed in explicit models into speech enhancement systems. Self-supervised learning (SSL), which can be regarded as a deep neural network (DNN)-based knowledge discovery system, emerged in recent years. These models trained with a large amount of data provides a machine-learning-based way to extract speech-related information. Although no manual annotation of data is provided in the training stage, experiments have validated that SSL models discover essential underlying speech-related patterns and extract meaningful representations that can be used for other tasks. Accordingly, a task-specific pipeline with SSL models can be divided into a (pre-trained) upstream SSL model to extract general speech representations, and a downstream model trained to complete a certain task. Therefore, SSL models can also be regarded as a prior knowledge encoding system. In the following two chapters, we investigate how to utilise SSL models in speech enhancement.

Numerous SSL models have been published in the past few years, making the choice of the appropriate model a research question by itself. Consequently, before dividing into the application of SSL models, a comprehensive understanding of them is necessary. Two questions are considered in Chapter 6 related to the application of SSL models in speech enhancement: a) what information can be

easily discovered from extracted representations, and b) how robust SSL models are against common interferences in the real world. Existing research on these questions remains mainly empirical: the downstream task models of the same architecture are trained with different upstream SSL models, to select the best-performing combination. Although a high-performance system can be obtained, this exhaustive search is not ideal given the huge time and energy consumption. In Chapter 6, we propose a framework to analyse these two aspects of SSL models directly based on their extracted embeddings. The analysis provides us the evidence to select SSL models in different scenarios. This chapter is the foundation for future investigation of SSL models in speech enhancement.

⋆ ⋆ ⋆

Y. Song, D. Kim, H-G. Kang and N. Madhu.

Published in the proceeding of the IEEE International Conference on Electronics, Information, and Communication (ICEIC), January 2024.

Abstract This paper conducts an analysis of latent embeddings generated by a range of pre-trained, self-supervised learning (SSL) models. Departing from conventional practices that predominantly focus on examining these embeddings within the realm of speech recognition tasks, our study investigates the characteristics associated with speakers and their behaviour under the influence of input distortions. We establish a controlled setting with varying background noise levels and different room impulse response conditions to assess the robustness of these embeddings. We measure speaker-related information by utilizing repetitive sentences spoken by multiple speakers. The results demonstrate that the robustness of pre-trained SSL models is influenced by the type and severity of distortion, whereas the inclusion of speaker information is determined by the specific pre-training approach employed. This distinct perspective offers valuable insights into the versatility and limitations of SSL models.

6.1 Introduction

Deep learning has had a significant influence on the domain of speech processing. However, the availability of a massive amount of labelled data is essential for attaining top-tier model performance. The scarcity of labelled data pairs for training has prompted researchers to develop various self-supervised learning (SSL) methods. These methods eliminate the requirement for explicit target speech data during the model training process.

Based on the training methodology, self-supervised models can be categorized [1] into generative [2–6], contrastive [7–9], and predictive [10–12] models.

Generative models are designed to predict future information by relying on preceding data. On the other hand, contrastive models utilize anchor representations as part of their training process to distinguish positive samples from negative samples, which are obtained within the model, while predictive models compute the target with a completely separate model, which is more similar to teacher-student training. These pre-trained models, which are trained on massive data, offer the benefit of adaptability to downstream tasks. This adaptability involves the utilization of the model's hidden states or fine-tuning it for the specific downstream task. The characteristics of the hidden states in self-supervised learning models have led researchers to incorporate them into large language models. In [13, 14], the w2v-BERT model is employed to extract semantic information from input speech.

There is no doubt that SSL models enhance the performance of downstream tasks when training the back-end model with SSL representations, especially for tasks with limited annotated data. Yet, it remains to be answered what information is preserved in the latent embeddings and how these models perform in complex acoustic environments for real-world applications. Some investigation into these queries has been carried out in associated applications. The evaluation of SSL models through keyword spotting or speaker verification [6, 11] offers some insights into the information present in these embeddings. With the help of a speech synthesis system, the information disentanglement properties of speech representations from three distinct SSL models are investigated in [15] through the assessment of various tasks. Furthermore, speech processing universal performance benchmark (SUPERB) [16] is proposed as a universal benchmark framework to ensure a fair comparison of different SSL models. It provides the standard datasets and metrics for the training and evaluation of compact back-end models across a range of downstream tasks. Since the pre-trained SSL model remains fixed and only a small back-end is trained, the scores obtained reflect the quality of the embeddings considering the given task. It should be noted that all of these evaluations are conducted using clean speech data. When enhancing or separating speech, it is crucial to consider potential interference caused by noise or cross-talk. The framework of SUPERB is expanded even further in [17], showing the effectiveness of SSL models in speech enhancement and separation tasks.

Our research aims to improve the efficiency of selecting an appropriate SSL model and accelerating system design by introducing a novel analysis of SSL representations that addresses two key aspects. Firstly, our analysis will show what kind of information is preserved within different SSL representations. Once this information is identified, one can select the pre-trained model that aligns with the requirements of the downstream tasks, or alternatively, integrate additional components into the system to compensate for any missing information. While there have been some investigations about this topic, these studies have primarily focused on different layers of a pre-fixed SSL model [15], or have been limited to

specific aspects according to particular system designs [18]. The lack of a comprehensive understanding of the information preserved in embeddings often leads to the optimization of the entire system through an iterative process of trial and error. Considering the large amount of available SSL models and the multiple hidden layers within each model, it would be very inefficient, if not impractical, to explore all possible combinations exhaustively. Consequently, an indication on the information preserved by the latent embeddings, which can be inferred from the embeddings themselves, would be helpful.

Secondly, there is a gap between the SSL training data and the real acoustic environment encountered during the inference stage. Distortions such as noise and reverberation are not always included in the SSL pre-training, but they are common in numerous audio applications. These distortions differ from the interference introduced by masking methods or pseudo prediction methods, which are commonly employed in unsupervised training schemes. The mismatch between the training tasks and the application scenarios in the real-world application may lead to performance degradation of SSL models. SSL models have been reported to exhibit strong performance in adverse environments, such as separating [17] or enhancing speech [15, 17]. Nevertheless, it remains unclear to what extent the latent embeddings are affected by audio interference. In this work, we introduce an analysis of the robustness of SSL representations, which should be considered when selecting SSL models for real-world applications. In addition, the level of robustness could also serve as an indicator of how much fine-tuning is necessary for the system to operate as intended.

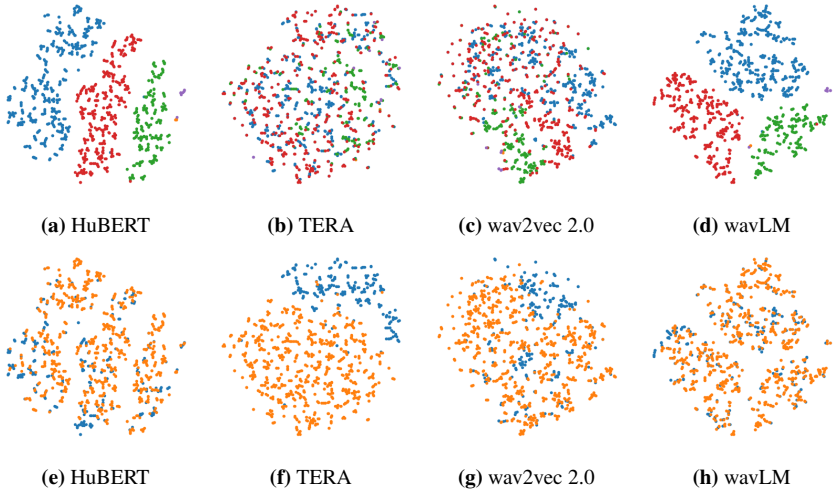
6.2 Experiments

6.2.1 Methodology

We select four well-known SSL models in the following analysis, each with a diverse training scheme, namely: hidden unit bidirectional encoder representations from transformers (HuBERT) [10], transformer encoder representations from alteration (TERA) [6], wav2vec 2.0 [8], and wavLM [11]. To investigate the content retained within self-supervised speech representations, we utilize the TIMIT dataset [19], given its comprehensive annotations at both phoneme and word levels. By employing sentences shared among different speakers, it becomes straightforward to create a controlled dataset to examine the information contained within the embeddings. If the embedding effectively preserves a specific type of information (such as phoneme information), then the representations containing similar information (extracted from the same phoneme, for instance) should exhibit proximity to each other. Thereby, the hyperspace should be separable based on the preserved information. Inspired by this intuition, we extract the latent embeddings from clean

TIMIT data and assess how closely their distribution correlates with various labels, which reveals the primary information encapsulated in the embeddings.

To illustrate how this analysis can shed light on the information preserved in SSL representations, we visualize all the averaged embeddings of the phoneme 'ao' from TIMIT sentence 'sa1', spoken by all speakers in the training set, using t-SNE [20] in Figure 6.1. The selection of this particular phoneme for illustration is due to its high frequency in the sentences, and its occurrence in multiple words across the sentence. The low-dimensional projection provided by t-SNE aims to retain local similarities as effectively as possible, making it a valuable tool for visualizing the distribution of latent embeddings in high-dimensional space. In the plot, every sample point corresponds to the average embedding of the consecutive frames of a single phoneme, and the same distribution is represented using two distinct labels: the word associated with the phoneme (top row), and the speaker's ID (bottom row). It is clear that certain models predominantly keep contextual information (forming a cluster according to the word information), while others are more influenced by the speaker information (with minimal overlap between male and female speakers). It is important to acknowledge that there is an inherent loss of information when projecting embeddings into a lower-dimensional space for visualization. Therefore, the absence of clear clustering based on one type of label does not necessarily imply the complete loss of that information.



(a) HuBERT **(b)** TERA **(c)** wav2vec 2.0 **(d)** wavLM

(e) HuBERT **(f)** TERA **(g)** wav2vec 2.0 **(h)** wavLM

Figure 6.1: The embedding distributions of all 'ao' sounds in 'sa1' from TIMIT training set, visualized by t-SNE. Each column presents the embeddings extracted by one SSL model. The plots in the first row are labeled by the words to which the phoneme belongs, and the second row by the speaker genders.

In the robustness analysis, the inclusion of background noise and reverberation

is considered to simulate the real-world recording conditions. The robustness of pre-trained SSL models is evaluated using the Valentini dataset [21] for speech, DEMAND noise dataset [22], and the MIT Impulse Response Survey dataset [23] for simulating distortion. We used one male and one female speaker, along with five different background noise types from various environments: living room, office, cafeteria, square, and bus noise, each at signal-to-noise ratio (SNR) levels of [-7, 0, 5, 10, 15] dB. With respect to the distortion, we utilize three distinct datasets: a noisy test set, a reverberation (reverb) test set, and a combined test set containing both noise and reverberation (all).

6.2.2 Metrics

To gauge the accessibility of specific information within the embeddings, we choose the training score of a logistic regression model. It indicates how easily the hyperspace can be separated in a linear manner when the embeddings are labelled based on that particular piece of information.

To measure the degree of embedding distortion resulting from audio interference, we employ two metrics to quantify this effect. The first metric relies on the mean-square error (MSE) between the clean and the distorted embeddings. In addition, we propose to normalize the embeddings by removing the mean and variance of the clean set before calculating the MSE. This normalization approach allows for a direct comparison of the robustness of different models, even when they embed the same audio into different latent spaces. Referring to the N-dimension embedding of one frame from the distorted signal as $\mathbf{e_s}$ and its clean reference as $\mathbf{e_x}$, the normalized MSE between the two can be computed as follows:

$$e(\mathbf{e_s}, \mathbf{e_x}) = \frac{1}{N}\big(\frac{\mathbf{e_s} - \mathbf{e_x}}{\sigma}\big)^T \cdot \big(\frac{\mathbf{e_s} - \mathbf{e_x}}{\sigma}\big)\,, \tag{6.1}$$

where σ is the variance of the *clean* data set embeddings.

The second measure we employ is the cosine similarity (CS) distance, which calculates the similarity between the clean and distorted embeddings in a polar coordinate system. With $\mathbf{e_s}$ and $\mathbf{e_x}$ represent the distorted embedding and its clean reference of one frame, respectively, the cosine similarity $\mathrm{sim}(\mathbf{e_s}, \mathbf{e_x})$ between them is defined as follows:

$$\mathrm{sim}(\mathbf{e_s}, \mathbf{e_x}) = \frac{\mathbf{e_s} \cdot \mathbf{e_x}}{||\mathbf{e_s}|| \cdot ||\mathbf{e_x}||}\,, \tag{6.2}$$

where $||\cdot||$ denotes the L2 norm.

6.3 Results

6.3.1 Preserved Information

We analyse the type of information extracted by the SSL models when the input is clean. The embedding is aggregated by averaging the embeddings of the same phoneme in consecutive frames. Since word, sentence, and speaker information is annotated in TIMIT, we systematically investigate these different types of information one by one. In Table 6.1, we provide the training scores, which represent the average prediction accuracy on the clean TIMIT training dataset.

Table 6.1: Logistic regression model training accuracy. Embeddings are extracted from TIMIT training set and averaged at phoneme level.

Data source	Target	Accuracy (%)			
		HuBERT	TERA	wav2vec2.0	wavLM
sa1	Phoneme	93.2	86.8	89.1	92.7
	Word	99.2	94.5	95.6	99.0
sx	Sentence	98.7	73.8	93.0	92.9
	Speaker	90.0	94.5	94.7	53.0

The first two rows of the table compare the significance of contextual information (predicting words) and phonetic information (predicting phonemes) within the embeddings. Since each speaker utters the same sentence ('sa1'), the cross-frame information (stemming from the transformer architecture) remains consistent for the same word. Therefore, the training scores, which signify the challenge of distinguishing phonemes or words in the latent space, are solely affected by the information preserved by the model. In all four models tested, classifying words is found to be a more straightforward task compared to classifying phonemes. This observation indicates that at the final hidden layer, all models tend to preserve a higher degree of contextual information compared to phonetic information. HuBERT exhibits the highest performance on both tasks.

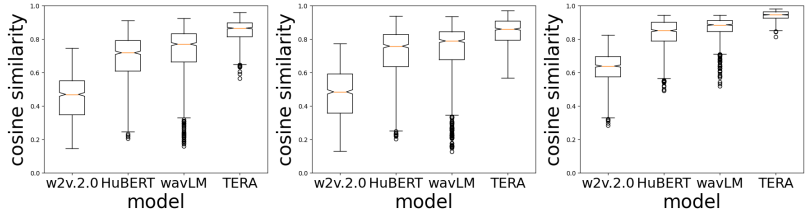
The following two rows involve a more comprehensive set of sentences, specifically all the phonetically-compact sentences ('sx') in the training dataset, which comprises a total of 330 unique sentences uttered by 462 speakers (5 sentences per speaker). This analysis provides a more extensive perspective on contextual information (i.e. sentence classification). Given the greater diversity of sentences, the training scores for speaker classification serve as an indicator of whether speaker information is preserved within the embeddings. The results clearly reveal a substantial contrast in the predominant information extracted by different models. It is challenging to deduce long-term contextual information from TERA representations, as evidenced by the low accuracy in sentence ID prediction from phonemes. However, TERA performs as the second-best model in terms of speaker information preservation, with only a slight 0.2% lower performance compared to

Table 6.2: Analysis of pretrained SSL model according to the SNR (dB) of noise distortion. ↓ means MSE is lower the better and ↑ means CS is higher the better.

Model		-7	0	5	10	15
wavLM	MSE ↓	0.967	0.593	0.430	0.352	0.295
	CS ↑	0.521	0.701	0.778	0.816	0.847
TERA	MSE ↓	0.452	0.334	0.265	0.212	0.166
	CS ↑	0.746	0.818	0.859	0.889	0.915

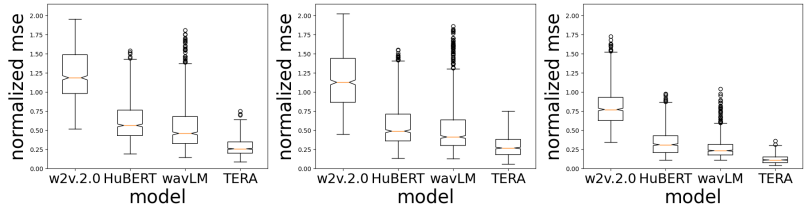
wav2vec 2.0 in speaker classification. On the contrary, wavLM, which ranks as the second-best in preserving contextual information with a 92.9% accuracy in sentence classification, significantly diminishes speaker information in its last hidden layer, achieving only 53.0% accuracy in speaker classification.

6.3.2 Distortion



(a) Noise+reverb **(b)** Noise **(c)** Reverb

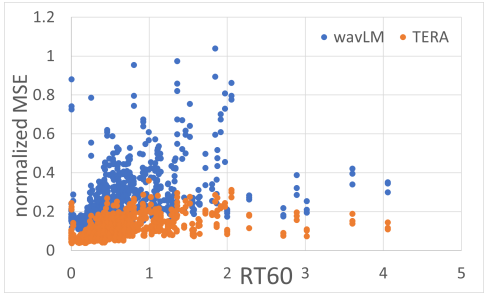
Figure 6.2: Cosine similarity between the clean and distorted embeddings from pre-trained SSL models. Three types types of distortions (noise+reverb, noise, reverb) are simulated.



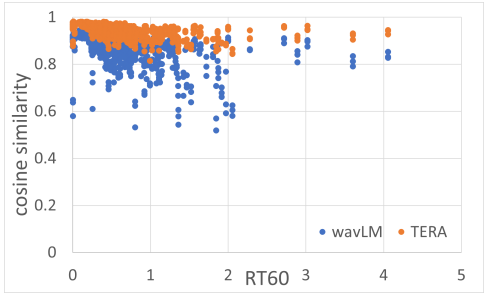
(a) Noise+reverb **(b)** Noise **(c)** Reverb

Figure 6.3: Standardized MSE between the clean and distorted embeddings from pre-trained SSL models. Three types types of distortions (noise+reverb, noise, reverb) are simulated.

To assess robustness with respect to various distortion types, we calculated the standardized mean squared error and cosine similarity for each pre-trained SSL



(a) Standardized MSE



(b) Cosine similarity

Figure 6.4: The influence of T60 on speech representations. Yellow is TERA and blue is wavLM.

model, as depicted in Figure 6.2 and Figure 6.3. Notably, the introduction of noise has a more detrimental impact on the robustness of the last hidden states in comparison to the distortion caused by reverberation. The further decline in robustness on the noise+reverb test set indicates that a combination of distortions exacerbates the quality of the embeddings. Among the pre-trained SSL models, TERA consistently exhibits the highest robustness across all test sets and measurements, with wavLM coming in second. The superior performance of TERA can be attributed to the augmentation techniques employed during training, which involved a range of masking and dropout methods. For a more in-depth analysis, we explored the results with consideration of the SNR and the room impulse response's RT60 parameter, as presented in Table 6.2 and Figure 6.4, focusing on top two models that exhibited similar performance: wavLM and TERA. In terms of additive noise, the distinction between the clean and distorted embeddings becomes more prominent when SNRs decrease. However, the differences observed in the context of reverberation were not as substantial, as illustrated in Figure 6.4. Interestingly, when evaluating the performance of TERA and wavLM on the reverberation test set, both models exhibited comparable mean performance. However, it is worth noting that TERA demonstrated significantly lower variance among samples compared

to wavLM, highlighting its outstanding robustness compared to other pre-trained models.

6.4 Conclusions

In this paper, we proposed to evaluate the quality of latent embeddings from SSL models directly through the embeddings themselves. This approach can furnish valuable insights for the selection of SSL models for specific downstream tasks, especially in the presence of noise and reverberation. We conducted an investigation on the latent embeddings from the last hidden layer of four well-known pre-trained models that were trained by various training schemes. Our analysis, based on embeddings from annotated clean speech, reveals that all four examined pre-trained SSL models tend to prioritize contextual information over phonetic information. The preservation of long-term contextual information and speaker information is contingent on the training scheme employed in SSL. For the practical application of SSL models in complex acoustic environments, we conducted a comparison of the robustness of the selected models' embeddings. When the input audio is distorted by noise or reverberation, the embeddings from TERA are least affected in terms of both the standardized MSE and cosine similarity. The quantitative results further highlight that additive noise has a more significant impact on the embeddings compared to reverberation.

References

[1] A. Mohamed, H. Lee, L. Borgholt, et al. *Self-supervised speech representation learning: A review*. IEEE J. Sel. Top. Signal Process., 2022.

[2] A. Van Den Oord, O. Vinyals, and K. k. *Neural discrete representation learning*. NeurIPS, 30, 2017.

[3] S. Pascual, M. Ravanelli, J. Serra, et al. *Learning problem-agnostic speech representations from multiple self-supervised tasks*. arXiv:1904.03416, 2019.

[4] M. Ravanelli, J. Zhong, S. Pascual, et al. *Multi-task self-supervised learning for robust speech recognition*. In ICASSP, pages 6989–6993, 2020.

[5] A. Liu, S. Yang, P. Chi, et al. *Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders*. In ICASSP, pages 6419–6423, 2020.

[6] A. Liu, S. Li, and H. Lee. *Tera: Self-supervised learning of transformer encoder representation for speech*. IEEE/ACM Trans. Audio, Speech, Language Process., 29:2351–2366, 2021.

[7] A. Baevski, S. Schneider, and M. Auli. *vq-wav2vec: Self-Supervised Learning of Discrete Speech Representations*. In ICLR, 2020.

[8] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli. *wav2vec 2.0: A framework for self-supervised learning of speech representations*. NeurIPS, 33:12449–12460, 2020.

[9] Y. Chung, Y. Zhang, W. Han, et al. *W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training*. In ASRU, pages 244–250, 2021.

[10] W. Hsu, B. Bolte, Y. Tsai, et al. *Hubert: Self-supervised speech representation learning by masked prediction of hidden units*. IEEE/ACM Trans. Audio, Speech, Language Process., 29:3451–3460, 2021.

[11] S. Chen, C. Wang, Z. Chen, et al. *Wavlm: Large-scale self-supervised pre-training for full stack speech processing*. IEEE J. Sel. Top. Signal Process., 16(6):1505–1518, 2022.

[12] A. Baevski, W. Hsu, Q. Xu, et al. *Data2vec: A general framework for self-supervised learning in speech, vision and language*. In ICML, pages 1298–1312, 2022.

[13] Z. Borsos, R. Marinier, D. Vincent, et al. *Audiolm: a language modeling approach to audio generation*. IEEE/ACM Trans. Audio, Speech, Language Process., 2023.

[14] A. Agostinelli, T. Denk, Z. Borsos, et al. *Musiclm: Generating music from text*. arXiv:2301.11325, 2023.

[15] K. Hung, S. Fu, H. Tseng, et al. *Boosting Self-Supervised Embeddings for Speech Enhancement*. In Proc. Interspeech 2022, pages 186–190, 2022. doi:10.21437/Interspeech.2022-10002.

[16] S. Yang, P. Chi, Y. Chuang, et al. *SUPERB: Speech Processing Universal PERformance Benchmark*. In Proc. Interspeech 2021, pages 1194–1198, 2021. doi:10.21437/Interspeech.2021-1775.

[17] Z. Huang, S. Watanabe, S. Yang, et al. *Investigating self-supervised learning for speech enhancement and separation*. In ICASSP, pages 6837–6841, 2022.

[18] A. Polyak, Y. Adi, J. Copet, et al. *Speech Resynthesis from Discrete Disentangled Self-Supervised Representations*. In Proc. Interspeech 2021, pages 3615–3619, 2021. doi:10.21437/Interspeech.2021-475.

[19] J. Garofolo, L. Lamel, W. Fisher, et al. *The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus*. NASA STI/Recon technical report n, 93:27403, 1993.

[20] L. Van der Maaten and G. Hinton. *Visualizing data using t-SNE.* J. Mach. Learn. Res., 9(11), 2008.

[21] Valentini-Botinhao, C. and others. *Noisy speech database for training speech enhancement algorithms and tts models*. 2017.

[22] J. Thiemann, N. Ito, and E. Vincent. *DEMAND: a collection of multi-channel recordings of acoustic noise in diverse environments*. In Proc. Meetings Acoust, pages 1–6, 2013.

[23] J. Traer and J. McDermott. *Statistics of natural reverberation enable perceptual separation of sound and space*. PNAS, 113(48):E7856–E7865, 2016.

7

Spectrum-Aware Neural Vocoder Based on Self-Supervised Learning for Speech Enhancement

In this chapter, we investigate how to utilise self-supervised learning (SSL) models, which automatically encode the prior knowledge on speech, for speech enhancement. We employ a re-synthesis framework: speech representations are extracted by the pre-trained SSL model, from which the underlying clean speech is then re-synthesised by a neural vocoder. Two major challenges are identified: the representations are distorted because of interference in the input signals, and some acoustic details that are important to high-fidelity signal reconstruction may be missing.

Since our analysis in Chapter 6 indicates that embeddings extracted by transformer encoder representations from alteration (TERA) are more robust against noise and reverberation than other models tested, we fix it as the upstream model to leverage this advantage. The neural vocoder employs the HiFi-GAN architecture, which is widely used for speech synthesis from other representations. Existing optimisation on HiFi-GAN for speech synthesis suggests providing more acoustic details benefits the synthesis performance. To include necessary acoustic details for speech synthesis, we propose to provide noisy spectrograms together with distorted embeddings to neural vocoders. This connection allows the neural vocoder to learn how to extract the relevant but missing details from noisy spectrogram directly. Since both SSL embeddings and noisy spectrograms are provided, the

optimal fusion method is decided by a series of ablation studies. Our experiments confirm that introducing noisy spectrograms into the neural vocoder improves the naturalness of enhanced speech signals.

⋆ ⋆ ⋆

Y. Song, D. Kim, H.-G. Kang and N. Madhu.

Published in the proceeding of the IEEE European Signal Processing Conference (EUSIPCO), August 2024.

Abstract Self-supervised learning (SSL) models for speech provide an efficient way to utilise raw, real-world data for acquiring versatile representations. Here, we investigate the benefits of employing such pre-trained SSL models for speech enhancement. Our approach involves customising a neural vocoder that produces enhanced speech using embeddings extracted from the noisy input by pre-trained SSL models. Specifically, we investigate the suitable incorporation of the noisy spectrogram in the network, to address possible loss of acoustic details in the embeddings. Through the exploration of different fusion techniques, we find that effectively incorporating both the SSL embeddings and noisy spectrogram into the neural vocoder results in a model that relies more on the noisy spectrogram for acoustic details and on the SSL embeddings for semantic information. Experimental results show that our proposed model yields a significant improvement of speech quality, compared to baseline models that rely solely on embeddings.

7.1 Introduction

A major obstacle in training deep learning networks using fully-supervised methods is the acquisition of sufficient, high-quality annotated data. Whereas sufficient audio data such as LibriSpeech [1] is available in the wild, they are usually unlabelled or only weakly-labelled. Self-supervised learning (SSL) models then emerge as an attractive alternative to utilise such large amounts of unlabelled data. SSL models are trained to extract versatile latent representations to enable various tasks such as automatic speech recognition, speaker identification and emotion recognition [2]. When integrating SSL models into a specific task, the final system is usually divided into an upstream component (SSL model) and a downstream component (front-end network tailored to the task).

For the task of speech enhancement, too, there has been some prior work on the use of SSLs. In [3], for example, a comparison was conducted among 13 SSL upstream models within the context of the same speech enhancement framework. A three-layer bidirectional long short-term memory (B-LSTM) network was

employed to predict a complex time-frequency-domain mask from the SSL representations of noisy signals. The clean speech spectrum was then estimated by applying the mask to the noisy spectrum. The same B-LSTM network predicts a better mask when using SSL embeddings as input, compared to using the noisy spectrum.

In [4], the vocoder architecture of HiFi-GAN [5] is explored to generate the clean speech signal directly from the distorted embeddings, extracted from noisy signals. Thus, the method was termed the 'denoising vocoder'. This variant, using modified contrastive predictive coding (CPC) [6] embeddings as input, was shown to outperform the version using log-Mel spectrograms. Similar to [3], both studies used a *learnable, weighted* sum of all hidden states of the embedding network, instead of only using the last hidden state. As verified in [3], this allows the downstream networks to make better use of the SSL models.

The HiFi-GAN of [4] was originally designed for the text-to-speech task, where the network learns to directly convert low-resolution mel spectrograms – predicted from the text – into high-fidelity waveforms by adversarial training. By replacing the real/fake labels of the discriminator by normalised objective metrics, e.g., as in MetricGAN [7] and MetricGAN+ [8], this architecture can be further adapted towards speech enhancement task.

Successful applications of SSL models across diverse speech tasks [2] highlight the presence of essential phonetic and semantic information in the learnt representations. Directly leveraging these representations – as in the aforementioned methods – can already yield improved denoising neural vocoders, particularly at the content-related level. However, it is also reported that the acoustic cues are missing in SSL embeddings. Incorporating task-relevant *a priori* information, such as fundamental frequency and speaker embedding, into the neural vocoder then aids in generating more natural speech, compared to using SSL embeddings alone [9].

For speech enhancement, this raises the interesting question of whether the downstream network can learn the required, additional information directly from the noisy spectrum. Thus, we systematically investigate means to effectively combine noisy spectrograms and speech representations (derived from pre-trained SSL models) for a HiFi-GAN-architecture-based denoising vocoder, and evaluate their benefits. Lastly, in addition to common speech quality metrics to benchmark the results, we also evaluate speaker *fidelity* by comparing the enhanced speech with the clean reference using a state-of-the-art speaker verification framework. The upper bound of the proposed system, that uses embeddings extracted from the underlying clean speech is also included in the benchmark.

Sec. 7.2 introduces the proposed generative framework, with the essential design considerations investigated. The experimental setup and ablation study settings are described in Sec. 7.3. Benchmarking results on various instrumental speech quality and intelligibility metrics are presented and analysed in Sec. 7.4.

Key take-aways are summarised in Sec. 7.5.

7.2 Methodology

7.2.1 Generative Model Framework

The goal of the system is to estimate the speech signal from the observed noisy and reverberant signal. Two features can be extracted from the observed signal, the spectrogram and the embedding, both of which contain distorted versions of the underlying target speech. While it has been shown in the aforementioned studies that the HiFi-GAN can estimate the underlying clean speech from either of the distorted features, here we investigate a complementary fusion of both features to achieve a better, consistent estimate. The proposed system is depicted in Figure 7.1. This fusion can be achieved in various ways, and using different versions of the input features – as we discuss below.

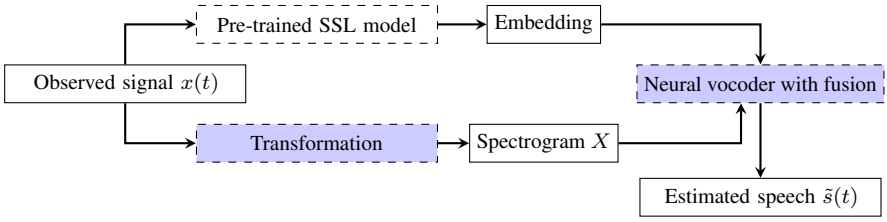


Figure 7.1: The diagram of the proposed system. The blocks with blue background are the components to be investigated in our work. The pre-trained SSL model is frozen.

7.2.2 Feature Extraction

For the first feature – the embedding – selecting an appropriate SSL model is important. In [10] we demonstrated that transformer encoder representations from alteration (TERA) [11] is robust to noise and reverberation. Such models can, therefore, extract better speech embeddings under adverse conditions - which is essential for application to speech enhancement. Thus TERA is chosen as the upstream SSL model in this work. We employ a pre-trained model, trained on 960-hours of LibriSpeech data from the s3prl toolkit[1]. The upstream model is frozen during the training of the downstream model.

The second feature – the observed signal spectrum – can be input either as magnitude or as log-magnitude. Both are examined, to see which representation

[1] https://github.com/s3prl/s3prl

allows for better denoising.

As the embedding dimensions of TERA and of the chosen spectral representation are different, additional layers are used to align the respective feature dimensionalities with our HiFi-GAN configuration. Following the approach outlined in [5], a 1D-convolutional layer is employed to transform the spectrum to the HiFi-GAN input dimension. The embeddings are projected by a fully-connected layer. The projections also allow implicit learnable weights of the two features.

7.2.3 Neural Vocoder: HiFi-GAN

HiFi-GAN consists of a generator and two discriminator components. The generator is responsible for upsampling the input to a time-domain signal using a stack of transposed 1D-convolutional layers and residual blocks at multiple scales. The discriminators are tasked with classifying the signal as either real (natural) or fake (synthesised).

We adjust the upsampling configuration to align with the frame length (25 ms) and frame shift (10 ms) of TERA. As we operate at 16 kHz, we configure the upsampling rates of the four residual blocks as $[8, 5, 2, 2]$, and the corresponding kernel sizes as $[16, 15, 4, 4]$, respectively. Other settings are identical to the original HiFi-GAN implementation[2] [5].

The total loss function for generator training consists of three components: a) the generator loss (the mean-square error (MSE) between the log-Mel spectra of the reference and the generated signals); b) the feature matching loss (MSE between the discriminator features of the reference and the generated signals); and c) the discriminator loss. Both the multi-scale discriminator (MSD) and the multi-period discriminator (MPD) are utilised as introduced in [5]. The training starts with a warmup stage, where only the generator is trained with the generator loss. Afterwards, all three losses are included.

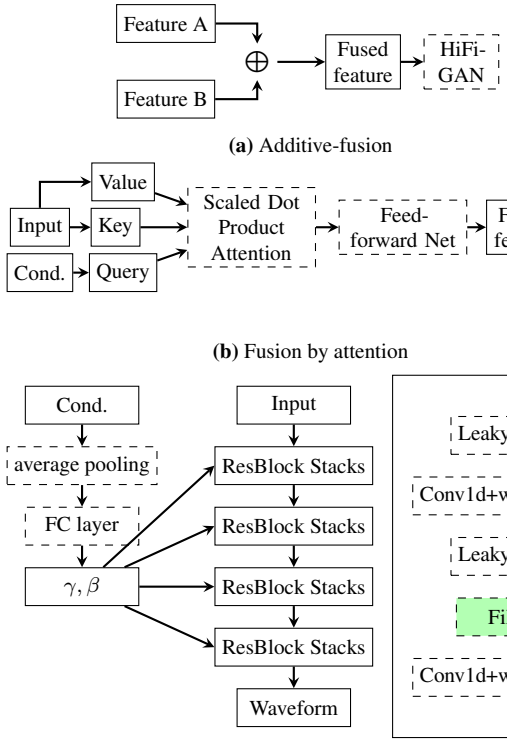
7.2.4 Fusion Methods

To fuse information from the TERA embeddings and the noisy input spectrum, we investigate three widely used strategies: addition, cross-attention transformer block [12], and the FiLM layer [13]. We depict their schematics in Figure 7.2.

Addition. The most naïve way to combine features is by a weighted sum. Using learnable weights, the network assigns relevant importance of each feature to the output.

Cross-attention using transformers. The transformer block uses the attention mechanism to combine different features. One feature is the *input* and the other, the

[2] https://github.com/jik876/hifi-gan/tree/master



(a) Additive-fusion

(b) Fusion by attention

(c) Fusion by FiLM layer

Figure 7.2: Different feature fusion methods, Cond.=conditioning. The feature fused by (a) addition or (b) attention block is converted into waveform by the followed neural vocoder, whereas the FiLM layer operates the feature maps of the neural vocoder by the extra conditioning information as shown in (c) (left). The architecture of the residual block with the FiLM layer is shown on the right.

conditioning. The conditioning provides auxiliary information to highlight the vital elements in the input term. The transformer block first calculates the multi-head cross-attention between the input X and the conditioning C. The query matrix is obtained from C, and the key- and the value matrices from X. Then, we employ a feed-forward network with two fully-connected layers, with layer normalisation as suggested in [12]. Two residual connections bypass the two modules, respectively. The number of multi-head attentions is set as 8.

FiLM layer. The FiLM layer [13] introduces the auxiliary information into the network by modifying the input feature maps by affine transformations, whose parameters depend on the feature used for *conditioning*. We apply this modification of the feature maps at the residual blocks in our vocoder. A vocoder residual

block consists of two convolutional layers, each preceded by an activation function (Leaky ReLU) and followed by weight normalisation. We insert the FiLM layer between the second activation function and the second convolutional layer. The conditioning feature is aggregated along both time- and feature-dimension by *mean-pooling*, and then linearly projected to a feature weight factor $\gamma_{i,c}$ and a bias $\beta_{i,c}$ to linearly transform the feature map from the cth channel and the ith layer of the neural vocoder.

7.2.5 Research Questions

Based on the above overview and the framework of Figure 7.1, the following questions now require to be addressed: **Q1.** How does the spectrum representation (i.e., magnitude or log-magnitude) affect the system performance? **Q2.** What is the relative contribution of representations extracted from each hidden state of TERA to the overall performance of the neural vocoder? **Q3.** How does the neural vocoder performance differ with the fusion method? **Q4.** For the cross-attention block and the FiLM layer, the input feature and conditioning feature play different roles in the system. Which feature is best suited to which role (i.e., spectrum as input and embedding as conditioning or vice-versa)?

7.3 Experiment Setup

For training and validation, we use the DNS 2021 challenge dataset [14] to synthesise 140 hours of noisy and reverberant audio. The signal-to-noise ratios (SNRs) are sampled from a uniform distribution ranging between -5 and 20 dB at 21 levels. Reverberation is simulated using synthetic room impulse responses (RIRs) SLR26 and SLR28 [15], where the reverberation time (RT60) is limited between 0.3 s to 1.3 s. To evaluate the performance of the proposed systems, we create a fully unseen test set using the CSTR VCTK speech corpus [16] and the NOISEX92 noise database [17], along with recorded RIRs from the MIT RIR database [18]. The SNRs are evenly distributed among $(-7, 0, 5, 10, 15)$ dB.

The network is trained by the AdamW optimiser with a learning rate of 0.0002 and $betas = [0.8, 0.99]$. Random cropping is adopted to further augment the training data: 10 second-utterances are cropped to 3 second segments from a randomly selected starting point after the warmup stage. The warmup stage continues for 20 epochs, and the system is further trained for a total of 200 epochs thereafter.

7.3.1 Metrics

Intrusive metrics have been reported to be less appropriate for assessing quality of generative models [19]. Therefore, we employ two non-intrusive metrics, deep

noise suppression mean opinion score (DNSMOS) [20] and non-intrusive objective speech quality assessment (NISQA)v2 [21], to evaluate the quality of the enhanced speech. Yet, with the reference signal available, one advantage of intrusive metrics is their sensitivity to 'hallucinated' input (a prominent artefact with generative models). Therefore, we also use the short-time objective intelligibility (STOI) metric [22], which is sensitive to perturbations of the speech envelope. Degraded STOI could, therefore, indicate compromised intelligibility.

Aside from speech quality metrics, we also evaluate how effectively the generative model preserves speaker information. For this we use the ECAPA-TDNN [23] speaker verification framework. Specifically, we compute the cosine similarity ($\in [0, 1]$) between the *speaker* embedding extracted from the clean reference and that extracted from the enhanced signal. A higher score indicates greater similarity between the two embeddings, and, consequently, speech synthesis with a more faithful capture of speaker characteristics.

7.3.2 Ablation Studies

Our *reference* proposed system employs the transformer block for feature fusion, followed by the neural vocoder. The cross-attention block takes the log-magnitude spectrum as the input feature and the (learnable) weighted sum of embeddings from all four TERA layers as the condition feature. As *baseline*, we train the neural vocoder to generate speech directly from the distorted TERA embeddings. This *Denoising Vocoder* network configuration is identical to [4].

To answer the questions listed in Sec. 7.2.5, we investigate each component individually *always* on the basis of the proposed *reference* system, where only one component is replaced each study. Each variant is independently trained on the same dataset with training settings identical to the reference system. The variants are listed in Table 7.1, and briefly described below. To address **Q1**, the log-magnitude spectrum input is replaced by the magnitude spectrum in variant 2. To benchmark the benefit of using information from multiple layers of the SSL embedding (**Q2**), variant 3 is trained using only the embedding from the *last* hidden state of TERA.

Next, we compare the attention block with the other two fusion methods (**Q3**): Variant 4 performs additive-fusion of the two features, and variant 5 uses FiLM layers. In attention blocks and FiLM layers, the input term and conditioning term serve different purposes by architecture design. To determine which feature is best suited to which role (**Q4**), we further train two networks. Variant 6 uses log-magnitude spectrum as the conditioning feature in the transformer block and variant 7 does the same for the FiLM.

Finally, to benchmark the upper bound of the proposed system, we train the neural vocoder for the ideal case, where the embeddings are extracted from the

Table 7.1: Evaluation results on the noisy, reverberant test set. The highest score of each metric is highlighted in bold, and the lowest by underline.

No.	Model Description	STOI	DNSMOS			NISQAv2					Speaker embedding cosine similarity
			OVRL	SIG	BAK	MOS	NOIS.	DIS.	COL.	LOUD.	
-	Distorted signals	0.770	1.814	2.458	2.005	1.659	1.697	3.073	2.328	2.505	0.600
-	Denoising vocoder (baseline)	0.808	3.086	3.379	**4.043**	3.097	3.601	3.325	2.953	3.726	0.551
1	Proposed reference	0.811	**3.054**	**3.405**	3.892	3.691	3.526	3.998	**3.494**	**3.992**	0.529
2	Magnitude spectrum feature	**0.819**	2.999	3.374	3.835	3.566	3.406	3.997	3.433	3.906	0.552
3	TERA - last hidden state	0.798	2.955	3.303	3.876	3.605	3.609	3.955	3.351	3.870	0.524
4	Additive-fusion	0.814	3.017	3.306	3.997	**3.768**	**3.932**	**4.032**	3.465	3.984	**0.584**
5	FiLM	0.739	2.696	3.005	3.827	2.828	3.409	3.408	2.614	3.434	0.387
6	Attention conditioned by spectrum	0.811	2.966	3.261	3.968	3.522	3.602	3.862	3.276	3.876	0.524
7	FiLM conditioned by spectrum	0.777	2.814	3.149	3.825	3.122	3.315	3.679	3.017	3.659	0.539
8	Clean embedding	0.949	3.121	3.429	3.977	4.161	3.876	4.330	3.979	4.238	0.899
-	Clean signals	1	3.668	3.951	4.209	4.550	4.251	4.596	4.301	4.476	-

clean speech and the log-magnitude spectrum from the noisy and reverberant input.

7.4 Results and Discussions

The evaluation results of all variants are summarised in Table 7.1. Compared to the distorted signals and the baseline, the proposed *reference* model and most of its variants show significant improvements in speech quality metrics, with some variants scoring higher in STOI as well.

Comparing variants 1 & 2, using the magnitude spectrum leads to a degradation in speech quality, especially in terms of background noise (BAK-DNSMOS / NOIS-NISQAv2). However, this change seems to help the neural vocoder in capturing the speech envelope more effectively, resulting in the highest score on STOI. This could be because the spectral nulls would be more strongly indicated in log scale, leading to better noise suppression and finer harmonics, but the stronger suppression could also lead to more aggressive attenuation of weak signal components (affecting STOI).

Comparing variants 1 & 3, by reducing the information included from the SSL model, all the metrics apart from DIS-NISQAv2 degrade. While it is commonly assumed that the final layer feature map contains the majority of speech-related information, this ablation study indicates that *other layers still contain some valuable information relevant to speech reconstruction*. We now examine the contri-

Table 7.2: Combination weights for TERA hidden state layers

Variant	**Layer1**	**Layer2**	**Layer3**	**Layer4**
1	-0.002	-0.011	0.036	0.098
2	0.003	0.016	-0.105	-0.248
4	0.017	0.025	-0.479	-1.229
5	0.015	0.116	-0.586	-1.495
6	0.015	0.105	-0.524	-1.275
7	-0.004	-0.166	0.037	0.157
8	-0.068	-0.048	-0.041	0.115

bution of each hidden state of TERA. Table 7.2 shows the learned weights used to combine different TERA hidden states for the different variants. Although there are differences in scale, the trend indicating the importance of the last two layers, particularly the last one, remains consistent. This differs from the trend reported in [3, 4] that the first few layers are given greater weight across all tested models. However, in [3, 4], the neural vocoder decodes only the distorted SSL embeddings. Thus, *we attribute this difference to the acoustic cues provided by the additional spectrum information in our method.* It is commonly assumed that the earlier layers contain more detailed phonetic information [3], while the final layer tends to encode more semantic information [2]. With the incorporation of auxiliary information, the model appears to prioritise the SSL embeddings for semantic information,

because the detailed phonetic information can be deduced from the spectrogram. This is in line with the findings in [9]: embeddings alone are insufficient to capture all phonetic details necessary for authentic speech reconstruction.

Regarding the fusion methods: additive-fusion (variant 4) of the projected spectrum and the embedding performs similarly to, or even better than the reference in terms of STOI, BAK-DNSMOS, and NOIS.-NISQAv2. Generally there is more residual noise in speech inactive frames in speech signal generated by the reference approach, which could be introduced by the attention block. However, listening to the samples, we observe that the additive-fusion method generates less content in the high-frequency range. As a consequence, the output signals sometimes sound less pleasant[3]. This is also reflected by the lower score on COL.-NISQAv2 - which indicates this colouration.

From results of variants 5 & 7, the FiLM layer is less effective in terms of feature fusion for the speech enhancement task. When the neural vocoder is *conditioned* by spectral information (variant 7), the performance is even worse than the baseline denoising vocoder. The performance degrades further when spectral information is taken as input (variant 5).

For the attention layer, the speech quality degrades when using the noisy spectrum information as the conditioning of the attention block, despite their similar performance in terms of the speech intelligibility. This indicates that the acoustic deatils cannot be captured by conditioning.

In terms of preserving speaker characteristics, feature fusion by addition performs best. These scores seem correlated with the STOI values.

Considering the overall performance, the proposed reference system and the additive fusion system exhibit respective advantages. Additive-fusion demands slightly lower computational resources and generates cleaner speech, whereas the attention-based system provides a speech signal with more details and fine-structure. The FiLM layers, however, degrade the neural vocoder performance. It should be noted that there is still a noticeable delta between the upper bound (variant 8) and the optimal systems across all metrics, particularly in the STOI score. Since the upper-bound system uses speech embeddings extracted from the clean speech, this supports the hypothesis that the network primarily relies on the speech embedding to obtain semantic information.

7.5 Conclusions

We proposed a method to enhance noisy and reverberant speech by a neural vocoder that incorporates both noisy spectrogram and the distorted embeddings extracted from the noisy input using pre-trained SSL models. Experiments show that intro-

[3]The audio samples can be found at https://aspireugent.github.io/EUSIPCO2024YS/.

ducing *relevant* additional information improves the quality of the speech generated by neural vocoder. Among the three fusion methods investigated, ablation studies demonstrate that addition or cross attention block conditioned on SSL embedings are effective. FiLM layers, however, degrade the neural vocoder performance no matter which feature is chosen as the condition. We also examine the reconstruction fidelity in terms of preserved speaker characteristics with the help of a speaker verification system. The score indicates that additive-fusion best captures speaker information. Additionally, by analysing the learnable weights used to combine different layers of embeddings, it becomes evident that the network extracts phonetic cues from the spectrogram and semantic information from the embeddings. Compared to the upper bound, which utilises SSL embeddings from clean signals, there is still a gap in the performance, highlighting potential for further research.

References

[1] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. *Librispeech: an asr corpus based on public domain audio books*. In Proc. IEEE Intl. Conf. on Acoust., Speech and Signal Process. (ICASSP), pages 5206–5210. IEEE, 2015.

[2] S.-W. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, , et al. *SUPERB: Speech processing universal performance benchmark*. In Proc. INTERSPEECH, pages 2127–2131, 2021.

[3] Z. Huang, S. Watanabe, S.-W. Yang, P. García, and S. Khudanpur. *Investigating self-supervised learning for speech enhancement and separation*. In Proc. IEEE Intl. Conf. on Acoust., Speech and Signal Process. (ICASSP), pages 6837–6841, 2022.

[4] B. Irvin, M. Stamenovic, M. Kegler, and L.-C. Yang. *Self-Supervised Learning for Speech Enhancement Through Synthesis*. In Proc. IEEE Intl. Conf. on Acoust., Speech and Signal Process. (ICASSP), pages 1–5, 2023.

[5] J. Kong, J. Kim, and J. Bae. *HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis*. Proc. Adv. in Neural Inf. Process. Syst., 33:17022–17033, 2020.

[6] M. Riviere, A. Joulin, P.-E. Mazaré, and E. Dupoux. *Unsupervised pretraining transfers well across languages*. In Proc. IEEE Intl. Conf. on Acoust., Speech and Signal Process. (ICASSP), pages 7414–7418. IEEE, 2020.

[7] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin. *MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement*. In Proc. Intl. Conf. on Machine Learning, pages 2031–2041. PMLR, 2019.

[8] S.-W. Fu, C. Yu, T.-A. Hsieh, P. Plantinga, M. Ravanelli, et al. *MetricGAN+: An improved version of metricgan for speech enhancement*. In Proc. INTERSPEECH, pages 201–205, 2021.

[9] A. Polyak, Y. Adi, J. Copet, E. Kharitonov, et al. *Speech resynthesis from discrete disentangled self-supervised representations*. In Proc. INTERSPEECH, pages 3615–3619, 2021.

[10] Y. Song, D. Kim, N. Madhu, and H.-G. Kang. *On the Disentanglement and Robustness of Self-Supervised Speech Representations*. In Intl. Conf. on Electron., Inf. and Commun. (ICEIC), pages 662–665. IEEE, 2024.

[11] A. T. Liu, S.-W. Li, and H.-y. Lee. *TERA: Self-supervised learning of transformer encoder representation for speech*. IEEE/ACM Trans. Audio, Speech, and Language Process., 29:2351–2366, 2021.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, et al. *Attention is all you need*. Proc. Adv. in Neural Inf. Process. Syst., 30, 2017.

[13] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville. *FiLM: Visual reasoning with a general conditioning layer*. In Proc. AAAI conf. on artificial intelligence, volume 32, 2018.

[14] C. K. Reddy, H. Dubey, V. Gopal, R. Cutler, et al. *ICASSP 2021 deep noise suppression challenge*. In Proc. IEEE Intl. Conf. on Acoust., Speech and Signal Process. (ICASSP), pages 6623–6627, 2021.

[15] T. Ko, V. Peddinti, D. Povey, et al. *A study on data augmentation of reverberant speech for robust speech recognition*. In Proc. IEEE Intl. Conf. on Acoust., Speech and Signal Process. (ICASSP), pages 5220–5224, 2017.

[16] C. Valentini-Botinhao. *Noisy speech database for training speech enhancement algorithms and TTS models*. University of Edinburgh. School of Informatics. Centre for Speech Technology Research, 2017. doi:https://doi.org/10.7488/ds/2117.

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

[18] J. Traer and J. H. McDermott. *Statistics of natural reverberation enable perceptual separation of sound and space*. Proc. of the National Academy of Sciences, 113(48):E7856–E7865, 2016.

[19] D. de Oliveira, J. Richter, J.-M. Lemercier, T. Peer, and T. Gerkmann. *On the Behavior of Intrusive and Non-intrusive Speech Enhancement Metrics in Predictive and Generative Settings*. In Speech Communication; 15th ITG Conference, pages 260–264. VDE, 2023.

[20] C. K. Reddy, V. Gopal, and R. Cutler. *DNSMOS P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors*. In Proc. IEEE Intl. Conf. on Acoust., Speech and Signal Process. (ICASSP), 2022.

[21] G. Mittag, B. Naderi, A. Chehadi, and S. Möller. *NISQA: A Deep CNN-Self-Attention Model for Multidimensional Speech Quality Prediction with*

Crowdsourced Datasets. In Proc. INTERSPEECH, pages 2127–2131, 2021. doi:10.21437/Interspeech.2021-299.

[22] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen. *A short-time objective intelligibility measure for time-frequency weighted noisy speech*. In Proc. IEEE Intl. Conf. on Acoust., Speech and Signal Process. (ICASSP), pages 4214–4217, 2010.

[23] B. Desplanques, J. Thienpondt, and K. Demuynck. *ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN based speaker verification*. In Proc. INTERSPEECH, pages 3830–3834, 2020.

8 Conclusions and Future Work

8.1 Conclusions

The main objective of this dissertation was to improve speech enhancement methods by incorporating prior knowledge, including both physiological models of speech production and perception as well as data-driven models. The journey started from classical statistical speech estimators and progressed to advanced deep neural network (DNN)-based approaches.

The investigation began with traditional speech enhancement approaches in Chapter 3, where speech magnitude spectrogram estimation was the focus of research. To emphasise the structure of speech harmonics, we introduced data-dependent parameters into the cepstral excitation manipulation (CEM) method. By refining the initial speech estimates according to the source-filter model, the proposed system yields speech estimates of better quality. The performance boost demonstrated the benefit of incorporating the source-filter model into the system even though it is a simplified approximation.

Next, we focused on enhancing the envelope component within the traditional speech enhancement framework, as detailed in A. The statistical gain function depends on both the *a priori* SNR and noise floor estimation. While the proposed approaches to improve the *a priori* SNR estimation led to some improvement in speech quality, the overall benefits remained incremental. A better noise floor estimation would be surely advantageous. However, noise distributions are significantly more diverse than speech, and the limited modelling capacity of statisti-

cal methods poses a fundamental challenge. Consequently, achieving substantial improvements in noise estimation is highly non-trivial. For this reason, the subsequent chapters shifted focus to DNN-based methods in the short-time Fourier transform (STFT) domain to better model speech distributions.

As the baseline for subsequent studies, we employed the convolutional recurrent U-net architecture for speech enhancement (CRUSE), an efficient architecture optimised for speech enhancement. To further enhance its performance, we successively examined and refined the estimation of the two components of the speech spectral coefficients, magnitude and phase. In Chapter 4, we addressed the challenge of preserving weak speech harmonics in low-signal-to-noise ratio (SNR) regions. Based on our findings in Chapter 3 about harmonic enhancement for statistical methods, we proposed to introduce the source-filter model as an additional loss function term (CEM loss) which measures harmonic reconstruction errors. Furthermore, inspired by the success of harmonics boosting in the cepstrum in Chapter 3, we applied a similar approach to emphasis the harmonics in the clean reference signal in the CEM loss. Experiments showed that this loss function guided the network to produce a more structured magnitude estimate without adding computational cost during inference. This indicates that the source-filter model was effectively integrated into DNNs, thus improving speech enhancement performance.

Next, we aimed at reconstructing a better phase spectrogram for the estimated magnitude. Phase estimation is known to be a challenge, due to the lack of structure and its sensitivity to time-shift. In Chapter 5, we explored a relatively understudied relationship in speech enhancement, the connection between log magnitudes and temporal and spectral phase derivatives of clean speech signals. DNNs could effectively learn this relationship to predict derivatives from known magnitude spectra. For phase retrieval, the final phase estimate is obtained by fusing results derived from temporal and spectral derivative estimates. When it comes to the phase reconstruction for speech enhancement, however, these derivatives must be predicted from the distorted magnitudes, which are estimated using typical speech enhancement methods such as CRUSE from the previous chapter. To leverage the initial phase estimate from speech enhancement, which is usually accurate at high SNRs, we proposed to fuse it with the other two derivative-based estimates. Experiments showed that the proposed phase reconstruction could improve enhanced speech quality when integrated with DNNs that predicted both real and complex-valued outputs.

In the next two chapters, the potential of self-supervised learning (SSL) models which automatically capture and encode relevant speech features was explored for speech enhancement transitioning from incorporating explicit, hand-crafted domain knowledge to exploiting data-driven representations.

Chapter 6 laid the foundation for utilising SSL models in speech enhancement.

Through quantitative analysis of SSL embeddings, we evaluated the encoded information and their robustness against noise and reverberation, providing a framework for SSL model selections. Our experiments showed that while several state-of-the-art SSL models can capture speech-related features, TERA is more robust to distortions caused by noise and reverberation among all tested models.

Consequently, Chapter 7 explored how TERA embeddings could be optimally utilised to synthesise high-quality clean speech from the embeddings of distorted signals. We employed a HiFi-GAN-based neural vocoder to convert embeddings back into time-domain signals. One major challenge of speech generation is to synthesis natural signals, which could be degraded when embeddings contain insufficient acoustic details. To tackle this challenge, we proposed to provide noisy spectrograms together with embeddings to the neural vocoder. Potentially, the high-level speech information summarised by embeddings could be more efficiently used when the acoustic details were directly accessible. Experiments showed that the access to noisy spectrogram enabled the neural vocoder to synthesise more natural signals.

Collectively, this dissertation demonstrates how statistical or DNN-based speech enhancement approaches can benefit from integrating prior knowledge into the systems. This prior knowledge could be domain knowledge summarised by human and abstracted as signal models, or speech-related information extracted by fully data-driven methods such as self-supervised learning approaches. With a well-designed integration, the additional knowledge improves speech enhancement performance in a complementary manner.

8.2 Future Research

While this dissertation presents advancements in incorporating prior knowledge into speech enhancement systems, several promising directions remain for further exploration and improvement. These areas could lead to more efficient, accurate, and robust speech enhancement systems.

One potential direction for future work is to explore the performance of the proposed CEM loss function within the context of more advanced network architectures, such as dual-path transformers (DPTs) or dual-path recurrent neural networks (DPRNNs). The designs of these models enable them to learn both time and frequency dependencies. It would be interesting to investigate whether the proposed CEM loss function can still offer improvements when applied to architectures that are inherently able to capture frequency relationships.

The phase reconstruction approach proposed in Chapter 5 could benefit from a more effective weighting strategy for different estimates. For example, the weight assigned to the noisy phase could be adjusted based on the input SNR, as the noisy phase is more reliable in high-SNR conditions. Moreover, experiments show

that the derivative prediction networks perform better when the magnitude distortion levels during training and interference are matched. Specifically, the networks trained to predict clean phase derivatives from estimated magnitudes are sub-optimal in high-SNR regions, where magnitude estimates are more accurate and phase distortion is minimal. Therefore, the estimated phase could be further improved if the weights are correctly conditioned on the input signal.

While the use of SSL models has been promising to introduce machine-learning-based, automatically discovered prior knowledge into speech enhancement systems, several questions and challenges remain to be addressed in future work. One ongoing challenge is to improve the *speaker similarity* between the clean reference and the enhanced signal. Our experiments show that although the generative methods can produce high-quality, natural enhanced speech, there is a lack of speaker identity consistency, as indicated by the low similarity between the speaker embedding extracted from the reference and the synthesised speech signals. Future research could investigate strategies for refining speaker representation within SSL-based systems, ensuring that the enhanced speech resembles the original speaker more closely.

Another challenge lies in an effective exploitation of the *semantic information* encapsulated by SSL representations. It has been verified that SSL models extract rich contextual information; however, this is not fully leveraged in the proposed enhancement framework, as the generative behaviour is more acoustic instead of semantic. Finding ways to use this semantic information could lead to SSL-based enhancement models that understand spoken languages and preserve—even predict when the environment is extremely challenging—the linguistic content of the target speech signals.

Finally, there is an opportunity to explore *the integration of the predictive models into the re-synthesis framework*. For example, the predictive model provide an initial speech estimate as the reference signal for the generative model to produce high-quality speech signals. Such a hybrid approach leverages the strengths of both systems—the robust noise suppression of predictive models and the high-quality speech reconstruction of generative models— achieving superior enhancement quality.

A

Investigations on the Optimal Estimation of Speech Envelopes for the Two-Stage Speech Enhancement

After examining the excitation signal in chapter 3, we investigate the enhancement of the other component—the envelope—of the source-filter model, within the two-stage speech enhancement framework. Since the speech envelope is related to the phoneme and the total number of unique phonemes is limited for a certain language, a codebook is a common tool for envelope analysis and enhancement. The intelligibility of enhanced speech should be improved when replacing the envelope of the initial speech estimate by the correct 'template' from the codebook of the language.

To optimise such a framework, we first investigate how to construct a better codebook. The codebook generation process (the necessity of manual division of speech active/inactive frames) and envelope representations (linear prediction coding (LPC) VS cepstrum) are examined for this purpose. Next, we examine the prediction of the underlying codebook template using a statistical model based on the Gaussian mixture model (GMM)-hidden Markov model (HMM) back-end envelope classifier and compare this to a convolutional recurrent neural network (CRNN). Moreover, a direct regression network that maps the distorted envelope coefficients to the clean envelope coefficients is also trained and compared with its codebook-based counterpart. With a complexity similar to the existing GMM-HMM system, the CRNNs improve the prediction accuracy and the overall perfor-

mance. Although the modifications effectively improve the quality of the enhanced signals, the overall performance of this method remains constrained by other components such as the noise estimator, with only marginal improvements in speech intelligibility observed. This reflects the inherent limitations of statistical methods in accurately modelling complex distribution, especially under challenging conditions. This observation motivates the exploration of more sophisticated neural network architectures to enhance their modelling capacity, thereby addressing the challenges posed by highly non-stationary, real-world noise.

⋆ ⋆ ⋆

Y. Song, and N. Madhu.

Published in MDPI sensors, July 2023.

Abstract Using the source-filter model of speech production, clean speech signals can be decomposed into an excitation component and an envelope component that is related to the phoneme being uttered. Therefore, restoring the envelope of degraded speech during speech enhancement can improve the intelligibility and quality of output. As the number of phonemes in spoken speech is limited, they can be adequately represented by a correspondingly limited number of envelopes. This can be exploited to improve the estimation of speech envelopes from a degraded signal in a data-driven manner. The improved envelopes are then used in a second stage to refine the final speech estimate. Envelopes are typically derived from the linear prediction coefficients (LPCs) or from the cepstral coefficients (CCoefs). The improved envelope is obtained either by mapping the degraded envelope onto pre-trained codebooks (classification approach) or by directly estimating it from the degraded envelope (regression approach). In this work, we first investigate the optimal features for envelope representation and codebook generation by a series of oracle tests. We demonstrate that CCoefs provide better envelope representation compared to using the LPCs. Further, we demonstrate that a unified speech codebook is advantageous compared to the typical codebook that manually splits speech and silence as separate entries. Next, we investigate low-complexity neural network architectures to map degraded envelopes to the optimal codebook entry in practical systems. We confirm that simple recurrent neural networks yield good performance with a low complexity and number of parameters. We also demonstrate that with a careful choice of the feature and architecture, a regression approach can further improve the performance at a lower computational cost. However, as also seen from the oracle tests, the benefit of the two-stage framework is now chiefly limited by the statistical noise floor estimate, leading to only a limited improvement in extremely adverse conditions. This highlights the need for further research on joint estimation of speech and noise for optimum enhancement.

A.1 Introduction

Speech captured by microphone in the real-world environment is prone to being corrupted by background noise. In order to reduce listener fatigue and the loss of intelligibility, speech enhancement which aims at removing the background noise and improving intelligibility has been an important and active field for many years.

The established statistical speech enhancement methods, the minimum mean-square error short-time spectral amplitude (MMSE-STSA) estimator [1], and the minimum mean-square error log-spectral amplitude (MMSE-LSA) [2], require an initial estimate of *a priori* signal-to-noise ratio (SNR) and *a posteriori* SNR derived from the spectral density power estimates of the clean speech and the background noise. On top of that, the decision-directed (DD) approach—a recursive smoothing procedure for the *a priori* SNR—is proposed in combination with these estimators to reduce the residual musical tones and improve the naturalness of the processed audio. However, this technique also introduces an estimation bias in the SNRs and leads to an annoying reverberation effect [3]. Therefore, a two-stage framework has been proposed in Reference [3] to avoid the speech distortion from this bias. In the two-stage framework, a better speech estimate is refined from the initial one so that a priori SNR can be calculated *without* recursive smoothing. Additional prior knowledge on speech can also be introduced during the procedure.

In our recent work [4], for example, we improve the speech harmonic recovery method termed cepstral excitation manipulation (CEM) [5] using the source-filter model of speech production to highlight its periodic structure. In this model, the speech signal is decomposed into an excitation and an envelope component in order to represent the excitation source and the vocal tract filter, respectively. It is proposed to amplify the quefrency related to the fundamental frequency and its harmonics in order to highlight the periodic structure of the voiced speech in the cepstral domain. To maintain the fine structure of the speech, the excitation signal in high quefrencies is smoothed with a quefrency-dependent window. In this work, we investigate the contribution of enhancing the other component of the source-filter decomposition result, the spectral envelope, for speech enhancement.

It has been shown that there is a strong correlation between the speech envelope and its intelligibility [6]. Consequently, the short-time spectral envelope of speech has been widely exploited in many areas such as automatic speech recognition (ASR) [7], artificial bandwidth extension [8], and speech intelligibility prediction [6, 7]. The well-known and widely used short-time objective intelligibility (STOI) [9] is based on the linear correlation between the envelopes of the processed noisy speech signal and those of the clean reference. One well-known problem of speech enhancement is that many methods hardly improve (and, sometimes, even degrade) speech intelligibility, although they perform well on noise reduction. Given the relationship between the speech envelope and intelligibility,

it is possible to improve speech quality as well as its intelligibility by refining the spectral envelope of the clean speech estimate.

For a given language, the possible patterns of speech spectral envelope are limited because the number of phonemes is limited, which makes the codebook technique an efficient solution to capture a priori information about speech envelopes. Thereby, a good estimate of the codebook entries integrates this prior knowledge into the following tasks. For example, [10] trained two sets of codebooks for speech and noise, respectively. Then, the gain function was estimated by the codebook-constrained Wiener filter, with optimal codebook entries being searched for in a maximum likelihood framework.

As speech has temporal dependency, a combination of the Gaussian mixture model (GMM) and hidden Markov model (HMM), GMM-HMM, was widely used in classical ASR systems as the back-end to recognise phonemes [7]. This statistical approach models the distribution of phonemes and their temporal dependency through two individual components: the GMMs, which learn the feature distribution, and the HMM, which imposes temporal dependencies on the hidden state sequences inferred from the GMMs. A similar pre-trained codebook with a GMM-HMM back-end also serves as the baseline in the speech envelope enhancement research of Reference [11] using the aforementioned two-stage framework. Therein, only the speech envelope codebook is generated, and the speech envelope is estimated from it in an minimum mean square error (MMSE) manner. This envelope is introduced to update the a priori SNR for a second-stage estimate of the clean speech. In Reference [12], noisy signals are enhanced by resynthesising the clean speech from the inferred acoustic cues (e.g., pitch and spectral envelope). The underlying clean speech envelope is, again, estimated with a codebook-aided Kalman filter, the codebook having been designed to capture not only the envelope shapes, but also the evolution of the envelopes in a given number of consecutive frames.

The classifier for such codebook-based methods can, of course, be replaced by deep learning models. In Reference [13], using two separate sets of codebooks for speech and noise, the codebook entries corresponding to the envelopes of both components are estimated by a feedforward deep neural network (DNN). These codebook entries are used to update the time-smoothed Wiener filter which performs the final speech enhancement. The work in Reference [11] also investigates the utilisation of DNN-based classifiers for codebook-based speech envelope estimation. Compared to the GMM-HMM baseline, the trained DNN classifier, with a similar computational cost, shows an advantage in both the classification accuracy and the instrumental metrics for speech enhancement. Compared to its regression counterpart, in which the network architecture is kept but the model is repurposed to predict the envelope coefficients directly from the initial estimation, it is shown that this architecture benefits from the codebook.

There are different ways to extract and represent speech envelope. Subband representation is a popular approach. For example, in STOI, the spectro-temporal envelope is calculated as the average of one-third octave decomposition results of 30 consecutive frames [9]. The STOI loss function to optimise speech enhancement DNN adopts, naturally, the same features [14]. Analogously, Reference [15] uses equivalent rectangular bandwidth to compress the spectrum and Reference [16] uses the auditory filterbank. It should be noted there is no direct inverse from these subband presentations to spectra. Applying the subband gain directly to the spectrum yields a 'rougher' signal, as the processed signal is less harmonic [15]. Therefore, the subband gain function is combined with a comb filter to restore the distorted harmonics in Reference [15].

The envelope can also be obtained from the auto-regressive (AR) filter applied, e.g., in linear prediction coding (LPC). For stability reasons, instead of directly using the AR filter coefficients, the equivalent line spectral frequencies are adopted for speech enhancement in References [12, 13]. Another equivalent representation of the AR filter coefficients is given by linear prediction coding coefficients (LPCCs). These are employed in Reference [11] to define the spectral codebook and to estimate the enhanced envelope within the two-stage framework. However, LPCCs can suffer from quantisation issues [17], which can cause degradation in codebook-based approaches. Another alternative envelope representation is based on the cepstral representation of the signal, which also implicitly describes the spectral envelope. Based on the relationship between the spectrum and the cepstrum, the first few cepstral coefficients (CCoefs) of a signal frame can be regarded as the description of its spectral shape. This is exploited in cepstral smoothing approaches [18–20] in order to remove musical noise. By preserving the first few cepstral coefficients, and strongly smoothing the rest, instantaneous temporal spectral fluctuations in the signal are suppressed while the principal structure (i.e., the spectral envelope) of the processed speech is maintained. Of the aforementioned envelope representations, we focus on the LPCCs and CCoefs in this work due to the convenience of their transformations between the domains and the easy fitting of the decomposition results to the source-filter model.

While data-driven, deep-learning-based end-to-end speech enhancement offers a powerful solution, the computational cost of such a system is still relatively high. Furthermore, a drawback of such systems is the black-box nature of the enhancement, which makes interpretability and control difficult. Data-driven envelope estimation, incorporated into classical speech enhancement, can provide us a compromise, and at a low computational complexity. This work is developed from the idea of cepstral envelope estimation (CEE) using the pre-trained codebook in [11]. We further explore its potential and investigate the achievable results of this method. Specifically, the following questions will be answered by our investigation: what is the maximum benefit of such data-driven two-stage enhancement? How (much)

does the quantisation of the envelopes affect the quality of the enhanced audio quality? What is the optimal cepstral speech envelope representation for the purpose of speech enhancement? Will the speech envelope classifier benefit from temporal modeling?

Below, we start from a series of oracle tests to investigate the potential and the limitation of the codebook method, in which different envelope representations are compared and benchmarked against each other in the two-stage framework. Then, several practical systems are trained and evaluated.

The temporal dependency of speech is usually taken into consideration in the aforementioned envelope enhancement methods via explicit temporal models or components such as Kalman filters and HMMs in the frameworks. For the DNN structure, however, including an HMM is counterproductive, as reported in Reference [11]. Therefore, we will investigate other possibilities to enable temporal modeling within the neural network. Literature shows that recurrent layers are powerful to this end. For instance, the long-short term memory (LSTM) layer is widely used in end-to-end speech enhancement systems due to its effective usage of long-term information granted by the gate mechanism. Recent research [21] shows that gated recurrent units (GRUs) can achieve comparable performance with less complexity, which is also verified in the speech enhancement tasks [22]. Therefore, in this contribution, we will investigate the performance of the GRU-based classifier for the speech envelope estimation using codebooks.

Furthermore, we will explore the usage of the more recent network architecture, convolutional recurrent neural network (CRNN), as the regression estimator. It is hypothesised in Reference [11] that the repurposed feedforward DNN is too small for the regression problem. Yet, fully connected neural networks have been gradually replaced by the convolutional layers and convolutional neural networks (CNNs) have been reported to yield high performance on many tasks and to do so with fewer parameters than feedforward DNNs. For a deep or complex network, CNNs can be easily trained in an end-to-end style. By inserting the recurrent layers into CNN, the network benefits from both the strong feature extraction ability of convolutional layers and the temporal modelling ability of the recurrent layers. Therefore, we propose to make use of the CRNN architecture for the regression problem.

The remainder of this paper is organised as follows. We provide an overview of the two-stage speech enhancement framework in Section A.2, so that the purpose and the target of CEE are clear. Section A.3 introduces the cepstral envelope estimation in a systematic manner, followed by its use in the two-stage enhancement framework. We report and discuss the evaluation results of the oracle tests and the practical systems in Section A.4 and answer the core questions raised above. The paper is summarised and concluded in Section A.5.

A.2 Speech Enhancement Framework

We consider the noisy observation $y(k)$, which consists of the target speech $s(k)$ corrupted by noise $v(k)$ in an additive way: $y(k) = s(k) + v(k)$, with k being the discrete time sample index. The microphone signal can then be transformed using the short-time Fourier transform (STFT) with an M-point windowed discrete Fourier transform (DFT). This yields $Y(m) = S(m) + V(m)$, where m is the frequency bin index and l is the frame index.

As summarised in Figure A.1, we adopt the same two-stage speech enhancement framework in Reference [11]. A preliminary denoising is performed in the first stage. The MMSE-LSA gain function is employed to obtain the initial speech estimate $\widehat{S}(m)$. As with the majority of the gain functions, this estimator $\widehat{G}_l(m)$ is expressed as a function of two crucial parameters: a priori SNR $\xi(m)$ and a posteriori SNR $\gamma(m)$. They are defined as:

$$\xi(m) = \frac{\lambda_s(m)}{\lambda_v(m)}, \tag{A.1}$$

and

$$\gamma(m) = \frac{|Y(m)|^2}{\lambda_v(m)}, \tag{A.2}$$

where $\lambda_s(m)$ and $\lambda_v(m)$ are the power spectral densities (PSDs) of the speech and noise signals, respectively. Since the true values of these PSDs cannot be obtained in practice, $\widehat{\gamma}_l(m)$ is approximated using the estimated noise PSD $\widehat{\lambda_v}(m)$ from the noise floor estimator, and $\widehat{\xi}_l(m)$ is obtained from the decision-directed (DD) approach. The clean speech amplitude estimate is then obtained by applying the gain function to the amplitude of the noisy observation:

$$|\widehat{S}(m)| = |Y(m)| \cdot \widehat{G}_l(m). \tag{A.3}$$

Then, according to the source-filter model, the enhanced signal is decomposed into the excitation signal $\widehat{R}(m)$ and the envelope $\widehat{H}(m)$, and each component can be enhanced individually. The enhancement of the speech excitation signal has been discussed in References [4, 5, 23], showing that the idealised excitation signal $\overline{R}_l(m)$ brings the benefit of recovering the weak or lost harmonics in the initial speech estimate.

While the excitation signal can be modeled by straightforward mathematical equations due to its periodic nature in the voiced frames with the largest energy [4, 5, 23], data-driven methods are more common in the estimation of the speech envelopes as in References [10–13]. If the underlying clean-speech envelope can be accurately estimated from the distorted or noisy signal envelope, it should improve the final speech estimate. One option to introduce prior knowledge of speech envelopes is to use codebooks. Thereby, the envelope estimation

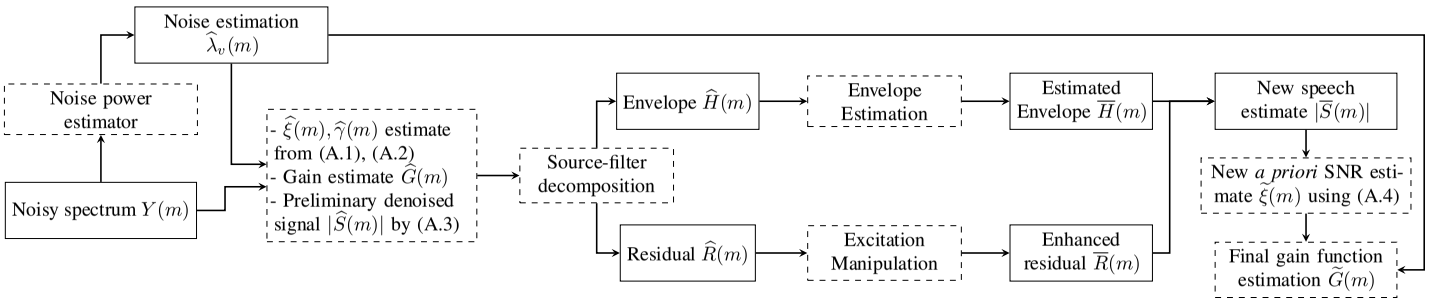


Figure A.1: Block diagram of the gain function calculation in two-stage noise reduction. Dashed boxes represent manipulation blocks whereas solid rectangular boxes indicate data contained. Please note that all terms are in the STFT domain, where the frame index l has been dropped for conciseness.

problem is converted into a classification problem. First, we create a codebook representing the different speech envelope patterns. Next, we train a suitable classifier to estimate the correct codebook entry for each frame, conditioned on the initial estimate $\widehat{H}(m)$. The other perspective is to regard the envelope estimation problem as a regression, which predicts the underlying clean-speech envelope from the noisy observation.

The improved speech envelope $\overline{H}_l(m)$ is subsequently combined with the refined speech excitation signal $\overline{R}_l(m)$, yielding an improved speech estimate $\overline{S}(m)$. However, this synthetic speech estimate sounds less natural than the initial speech estimate, as the excitation signal and the envelope are artificially imposed. Thus, instead of using $\overline{S}(m)$ to recover the speech, we use $\overline{S}(m)$ to update the a priori SNR:

$$\widetilde{\xi}_l = \frac{|\overline{S}(m)|^2}{\widehat{\lambda_v}(m)}. \tag{A.4}$$

This is then used to compute a new gain function: $\widetilde{G}_l(m) = G^{\mathrm{LSA}}(\widetilde{\xi}_l(m), \widehat{\gamma}_l(m))$. The final speech estimate $\widetilde{S}(m)$ is given by applying this new gain $\widetilde{G}_l(m)$ to the microphone signal $Y(m)$:

$$\widetilde{S}(m) = Y(m) \cdot \widetilde{G}_l(m)\,. \tag{A.5}$$

The enhanced time-domain signal can be obtained from $\widetilde{S}$ by over-lap add.

A.3 Cepstral Envelope Estimation

Since the envelope can be compactly represented in the cepstral domain, this estimation is named cepstral envelope estimation (CEE) in Reference [11]. It has been shown that the classification DNN (C-DNN) is the optimal system in their framework in comparison with the GMM-HMM baseline, DNN-HMM pipeline, and the regression DNN. Thus, we take C-DNN as the baseline of our research.

In this baseline system, the envelope is extracted by LPC analysis and represented by LPCCs. The codebook is generated in two steps. First, using the energy level criterion, the windowed frames of clean signals are divided into two categories, namely speech active frames and speech inactive frames. Then, the zero-mean speech active frames are clustered into C classes by the Linde–Buzo–Gray (LBG) algorithm [24]. To complete the codebook generation, one spectrally flat envelope is added as the template for silent frames. Once this codebook is generated, the speech active frames of the training data are labelled by assigning them to the closest template (codebook vector). A classifier can be trained using these labels and appropriate multi-condition data. During inference, the output of the classifier is interpreted as the posterior distribution of the codebook entries conditioned on the observation. The final envelope estimate $\overline{\mathbf{c}}_l$ is then obtained either

by maximum a posteriori considerations, or by a weighted sum of the different templates (i.e., the optimal estimate in the MMSE sense):

$$\overline{\mathbf{c}}_l = \sum_{i=0}^{C-1} p_l^i \cdot \mathbf{h}^i \,, \tag{A.6}$$

where $\mathbf{h}^i$ is the ith template in the codebook, and p_l^i is the posterior probability of the ith template for frame l, conditioned on the noisy observation.

In this section, we will take a closer look at each individual step of this baseline to further optimise it for speech enhancement.

A.3.1 Feature Extraction

A.3.1.1 LPCC

The LPC coefficients $(\{b1, b2, \ldots, bN\})$ of the AR model, for frame l, can be derived from the auto-correlation function [25] of the preliminary speech estimate. The coefficients are then converted to the cepstrum in the following recursive manner:

$$\begin{aligned} c_l(0) &= \ln N \\ c_l(p) &= bp + \sum_{i=1}^{p-1} \frac{i}{p} c_l(i) bp - i \,, \text{for } 1 \leq p \leq N \,. \end{aligned} \tag{A.7}$$

These coefficients $\mathbf{c}_l = \{c_l(1), c_l(2), \ldots, c_l(N)\}$ derived from LPC are taken as speech envelope representations.

A.3.1.2 Cepstral Coefficients (CCoef)

Cepstral coefficients are straightforward to calculate from the preliminary speech spectrum $\widehat{S}(m)$ by a M-point inverse discrete Fourier transform (iDFT) as:

$$d_l(q) = \text{iDFT}\{\log |\widehat{S}(m)|\} \,, \tag{A.8}$$

with q being the quefrency bin. Given the symmetric nature of the cepstrum (property of the (i) DFT on real-valued symmetric spectra), only the first half of the cepstrum (from bin 0 to bin $M/2$) needs to be preserved for further investigation. Then, the coefficients can be divided into three parts according to the source-filter model: first, the energy term $d_l(0)$; next, the initial few coefficients representing the speech envelope, namely $\mathbf{d}_l = \{d_l(1), d_l(2), \ldots, d_l(N)\}$; and, lastly, the remaining coefficients encoding the speech fine structure.

A.3.2 Codebook

It is proposed to create the codebook from speech active frames in Reference [11]. The partition is reasonable given the purpose of the codebook, but it should be noted that the energy criterion is not perfect to generate speech activity detection labels. The short-time Fourier representation is computed on overlapped, windowed frames. Thus, some frames that are classified as speech inactive actually possess very weak speech (due to the leakage of speech into the adjacent silent frames), and thus, their envelopes move away from a flat shape. This indicates potentially more variance even in the *low-energy* frames. In CEE, this error can degrade the system performance in two possible ways. First, according to this procedure, the idealised flat envelope is assigned to all low-energy frames, although some of them are closer to one of the speech templates. If the classifier is perfectly accurate, this would introduce speech distortion to the speech estimate when updating the a priori SNRs according to (A.4). In addition, this assignment error increases the difficulty of classifier training. The codebook assignment can be regarded as a quantisation of the clean speech signal envelope, and the classifier is trained to learn the mapping from the distorted coefficients to these quantised templates. This is the standard setting of the classification problem. However, the frames are assigned based on different rules, which makes the learning target ambiguous: it can be either a mapping to the most-likely speech envelope template for a speech active frame, or a replacement by a complete flat envelope for a *low-energy* frame. Moreover, the major indication of this speech/non-speech mapping, the frame energy level, is not available to the classifier. From the point of view of network training, this manually separated 'silent' codebook entry actually introduces noise into the training set. Consequently, the network training could be deteriorated by this elaborate division.

In order to investigate the influence of this assignment error, we followed the procedure to create the LPCC codebook from the speech-active frames of the TIMIT training set. After including the ideal flat envelope for silence, we reassigned all the 'non-speech' frames to entries of this codebook based on the cepstral distance. Figure A.2 depicts the codebook entry distribution for the two types of frames, where it is seen that, in fact, a large portion of the 'non-speech' frames have envelopes similar to templates corresponding to speech-active frames and, therefore, should *not* be quantised as a single idealised entry with a flat envelope. Another interesting observation from this figure is that the envelopes of these low-energy frames concentrate on a few entries.

In order to take a closer look at those wrongly assigned frames, we plot all the envelope templates in Figure A.3 and arrange them in a descending order of the posterior distribution of the codebook entries on non-speech frames ($p(i|H_0)$). In other words, the low-energy frames are more likely to take the envelopes on the left side of the figure. Two points are now obvious: (1) despite manually removing

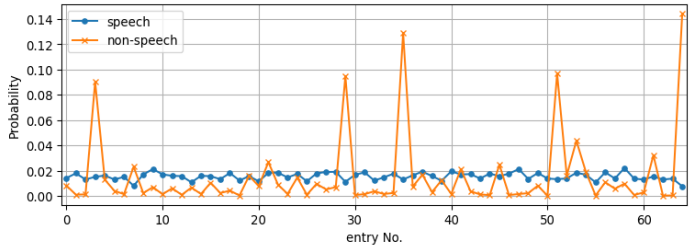


Figure A.2: Distribution of speech/non-speech frames where the speech codebook entries are obtained from the speech active frames and the non-speech frames are idealised by a single entry with a flat envelope. We term this the *separate* codebook. Note, however, that when assigning non-speech frames on this codebook using the cepstral distance, they often correspond to codebook entries of speech-active frames.

the low-energy ('silence') frames, the speech codebook can still contain spectrally flat templates; (2) low-energy frames have, more often than not, non-flat spectral envelope shapes.

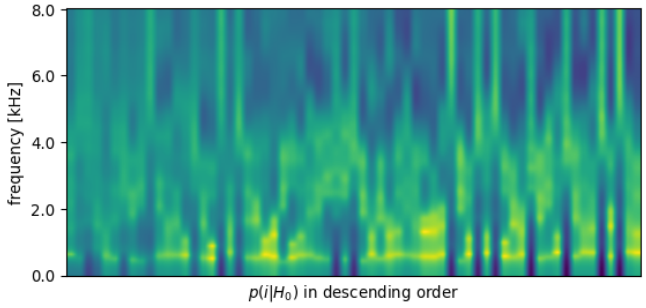


Figure A.3: Envelope templates in the separate codebook, arranged by the posterior distribution of the codebook entries on non-speech frames ($p(i|H_0)$). The envelope with the highest possibility of being non-speech is on the left.

With this observation on the existing codebook generating method, we propose to create the codebook from the envelopes of *all* frames in the clean-speech data. Figure A.2 has shown that the clustering generates clearly distinct templates. Thus, separating the frames in advance by the energy criterion is not necessary. Using zero-mean LPCCs of all frames, now, we create a new *unified* codebook for the same speech corpus and plot the entry distribution in Figure A.4. It can now be observed that, in contrast to using the separate codebook, the two types of frames are relatively mutually exclusive with regard to their distribution among the unified codebook entries. Therefore, the unified codebook could be a better choice for

the speech envelope enhancement task. It should be noted that although we only demonstrate the comparison between the separate and the unified codebook for LPCC, an identical trend can also be observed in the CCoef-based codebooks.

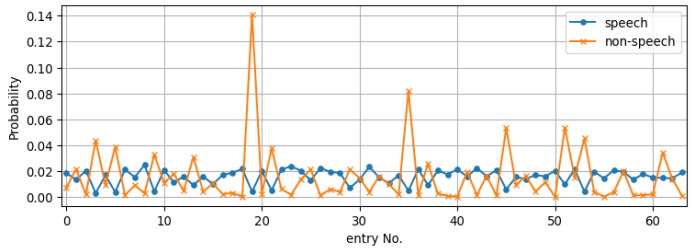


Figure A.4: Entry distribution of speech/non-speech frames on the unified codebook. Here, the distributions of the speech and non-speech frames across the codebook entries seem relatively mutually exclusive—speech-active frames rarely correspond to codebook entries where non-speech frames show a high probability of association.

A.3.3 Envelope Estimator

A.3.3.1 Feedforward DNN Classifier

In Reference [11], it was shown that using a feedforward DNN for the classifier outperformed the GMM-HMM approach. This network has a stack of fully connected layers as hidden layers, and the output is normalised by the softmax layer for the classification. The investigation showed that when the size of the network is fixed, the number of hidden layers and the choice of activation functions both have very limited influence on the network classification accuracy. We take the network composed of four hidden layers as our baseline, because this network achieves the highest accuracy on both the development set and the test set. We choose the activation functions of the network as follows: Leaky ReLU for the input and hidden layers, and the sigmoid function for the output layer, followed by a softmax layer to normalise the output.

A.3.3.2 Recurrent Neural Network-Based Classifier

Although the aforementioned DNN shows superiority to the GMM-HMM backend baseline in terms of classification accuracy, the temporal modeling ability provided by HMM is missing in this DNN architecture. As a remedy, the DNN was complemented by an HMM. However, the HMM results in a performance bottleneck [11]. In this regard, recurrent neural networks could be a more suitable comparison to the GMM-HMM baseline. In order to examine the function of the

recurrent layers in the envelope estimation, we investigate the GRU-based classifier in this work. We adopt the simplest architecture here: one or several GRU layers in cascade with one fully connected (FC) layer to compress the output feature dimension. The final output is normalised by the softmax function, allowing for its interpretation as the *a posteriori* probability distribution across the codebook entries.

A.3.3.3 CRNN-Based Envelope Estimation by Regression

The envelope estimation problem can also be formulated as a regression problem from the distorted envelope coefficients to the clean ones. Yet, according to Reference [11], the performance of a regression feedforward DNN is imbalanced at different SNRs. We propose to use an alternative architecture, the convolutional neural network, for the regression problem. CNN is popular because of its self-learning feature extraction ability. In order to fully exploit this feature, we reformulate the original coefficient-to-coefficient envelope estimation problem into a regression from the noisy spectrum with the initial gain function estimate to the clean speech envelope coefficients. The two available features, the logarithm of the zero-mean noisy magnitude spectrum and the logarithm of the MMSE-LSA gain function, are taken as two separate channels of the network input. After several convolutional layers with the leaky ReLU activation function, the feature map is flattened and combined by the FC layer. One GRU layer is employed as the final output layer of this regression network in order to combine the past states with the current prediction and obtain a final estimate of the envelope coefficients.

A.4 Evaluations

We will now evaluate the proposed speech envelope enhancement method in two settings: (a) the oracle tests that assume the classifier is perfectly accurate—this provides us with the basis for feature selection and demonstrates the full potential of the current framework—and (b) the practical system tests that evaluate and compare the different envelope estimation approaches in realistic settings.

A.4.1 Experimental Setup

All of the networks were trained and tested with the same synthesised data set; 90% of the TIMIT training set and 21 files from ETSI noise set were mixed at 6 SNRs: $\{-5$ dB, 0 dB, 5 dB, 10 dB, 15 dB, 20 dB$\}$. The remaining 10% of the TIMIT training set was reserved for the validation set. For the evaluation, the test set was created from the TIMIT test set and the unseen noise signals from the ETSI database (Car, Traffic) and QUT database (Cafe, Kitchen, and City) at the same 6 SNR levels. All of the speech and noise signals were down-sampled to 16 kHz

and high-pass filtered by a second-order Butterworth filter with a cutoff frequency at 100 Hz before mixing. The SNRs for the mixing were calculated according to Reference [26] in which, for speech, the speech active level is used, and the noise level is computed using the long-term root-mean-square.

For all tests, the input noisy signal was first processed as follows in the preliminary denoising stage. A pre-emphasis filter with a coefficient of -0.97 was first applied. The signal was then segmented with 50% overlap and windowed by the square-root von Hann window of $M = 512$ prior to computing its spectrum. For the LSA gain function, the smoothing factor of the DD approach was $\alpha = 0.97$. Further, the a priori SNR and a posteriori SNR were bounded between -40 and 40 dB in order to avoid numerical issues. The gain function was, finally, lower-bounded to -15 dB. The noise floor was estimated by the speech presence probability minimum mean-square error (SPP-MMSE) approach with fixed priors [27]. Without prior knowledge, we assumed an equal a priori probability for speech presence and absence as suggested, and the optimal a priori SPP was set to 15 dB.

We kept the length of both feature vectors (LPCCs and CCoefs) set to $N = 20$ for a fair comparison. The sizes of the networks and the computational costs are indicated by the total number of their parameters and multiply-accumulate operations (MACs) per frame. It should, however, be noted that there is a small difference of these values when using the unified or the separate codebooks: in the latter case, there is one additional entry, so the output layer of the classification networks needs to be modified accordingly. However, the difference caused by the choice of the codebook is negligible compared to the total size and complexity of the networks. Therefore, we report here the network size and MACs taking the unified codebook as example. As with the baseline system summarised in Table A.1, the C-DNN with four FC hidden layers with 73 units in each layer has 29,820 parameters and performs 27,448 MACs per frame.

Table A.1: Parameters of the Feedforward DNN classifier.

Input Size	**#Hidden Layer**s	**#Unit in the Hidden Layer**
$N = 20$	4	73
Default activation functions		Leaky ReLU, slope$= 0.03$
Activation functions of the last layer		Sigmoid
Output normalisation		Softmax
Number of parameters		29,820
MACs per frame		27,448

For the GRU classifier, we took one single layer of GRU with 62 nodes, which yielded a network with 19,656 parameters and 19,220 MACs per frame, as shown in Table A.2.

Table A.2: Parameters of the GRU classifier.

Input size	#GRU layers	#Unit in the GRU layer
$N = 20$	1	62
Activation functions		Sigmoid
Output normalisation		Softmax
Number of parameters		19,656
MACs per frame		19,220

These two classification networks were trained on the negative log-likelihood (NLL) loss function. Since the training set was imbalanced, the NLL loss was further weighted by the inverse of the normalised distribution of the codebook entries on the training set. The learning rate was 0.001 for all networks. It was shown in Reference [11] that envelopes estimated in an MMSE manner have an advantage over their maximum a posteriori (MAP) counterparts. Thus, we adopted the MMSE approach for all classifiers.

Detailed parameters of the CRNN that predicts envelope coefficients from noisy logarithmic power spectrum (LPS) and the LSA gain are listed in Table A.3. The regression network was trained by the MSE loss between the network prediction and the clean reference envelope coefficients.

Table A.3: Parameters of the CRNN regression net

Channels	4,8,8,1
Kernel size (Time=1, Frequency)	3,3,3,1
Stride (Time=1, Frequency)	2,2,1,1
Default activation functions	Leaky ReLU, slope= 0.03
Number of parameters	4101
MACs per frame	11,044

A.4.2 Quality Measures

We evaluated the quality of the processed signal through four metrics from different perspectives. The speech quality was measured based on the wide-band perceptual evaluation of speech quality (WB-PESQ) [28]. We used the mean opinion score - listening quality objective (MOS-LQO) scores whose range fell between 1.04 and 4.64 for the evaluation. In the following text, we denote WB-PESQ MOS-LQO as PESQ in shorthand. The second metric was STOI [9]. STOI indicates the speech intelligibility as a value between 0 (incomprehensible) and 1 (perfect intelligibility). A higher score is preferred on both metrics.

Apart from these two widely used metrics, we also employed the white-box approach [5] to separately benchmark noise suppression and signal distortion. To this end, the final gain function estimation $\widetilde{G}_l(l)$ was applied to the speech and noise component of the noisy input in order to obtain the filtered components, respectively:

$$s' = \text{iSTFT}\{\widetilde{G}_l(l) \cdot S(l)\}\,, \tag{A.9}$$

and

$$v' = \text{iSTFT}\{\widetilde{G}_l(l) \cdot V(l)\}\,. \tag{A.10}$$

then, noise attenuation (NA)—the metric that measures the noise reduction ability—was given by:

$$\text{NA} = 10\log_{10}\big[\frac{1}{L}\sum_{l=0}^{L-1}\frac{\sum_{k=0}^{T-1} v(k+lT)^2}{\sum_{k=0}^{T-1} v'(k+lT)^2}\big]\,, \tag{A.11}$$

where T is the frame length.

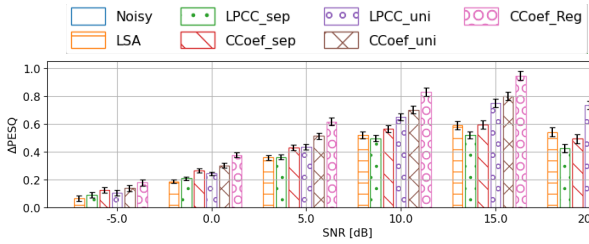
Similarly, the introduced signal distortion was measured by the segmental speech-to-speech-distortion ratio (SSDR) as:

$$\text{SSDR} = \frac{1}{||L_1||}\sum_{l\in L_1} 10\log_{10}\big\{\frac{\sum_{k=0}^{T-1} s(k+lT)^2}{\sum_{k=0}^{T-1}[s(k+lT)-s'(k+lT)]^2}\big\}\,, \tag{A.12}$$

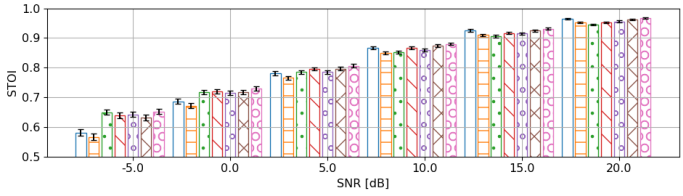
where L_1 is the set of speech active frames, and $||\cdot||$ is the cardinality of the set.

A.4.3 Oracle Test Results

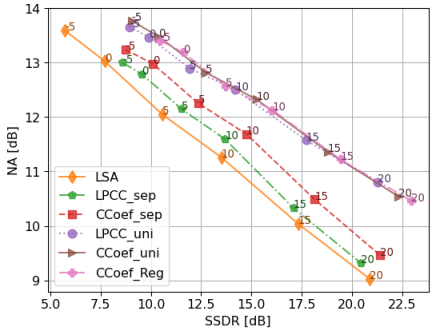
First of all, we investigated the optimal speech envelope representations by using the oracle tests. In these tests, the ground truth of the envelope codebook entries are available while the two-stage framework is maintained. In other words, the oracle tests demonstrate the upper bound of the envelope enhancement method in the given two-stage framework. The performance difference among these systems depends, then, purely on the adequacy of representations and the codebook generation methods in this task. Therefore, we can choose the optimal feature based on the oracle test results. The features were examined in the following aspects: (a) the original codebook whose corpus was manually separated into two categories (dubbed 'separate codebook') vs. the proposed codebook that was created from all available materials ('unified codebook'); (b) LPCC vs. CCoef as the speech envelope representation. Apart from the preliminary denoised results (LSA), one extra baseline used was the oracle regression method, which utilizes the unquantised clean envelope coefficients in the two-stage framework. Comparison of this oracle regression and the oracle codebook systems measures the distortion introduced by envelope quantisation. The evaluation results of these oracle tests are shown in Figure A.5.



(a) ΔPESQ compared to the noisy signals



(b) STOI metrics for the oracle tests



(c) NA−SSDR

Figure A.5: The oracle system evaluated by ΔPESQ, STOI, and NA−SSDR, and grouped by input SNRs. For the bar−plots, the 95% confidence interval is given by the error bars. LSA: the preliminary denoising output; _sep: using separate codebook; _uni: using unified codebook; _Reg: the regression method.

A.4.3.1 Codebook Optimisation: Findings from Oracle Tests

In terms of the speech quality, the unified codebook has a clear advantage, and the gap between two codebooks increases with the input SNR. When the SNR exceeds 5 dB, the separate codebook quantisation even degrades compared to the output from preliminary denoising. The separate codebook has a marginal advantage on STOI when the input SNR is low ($\leq$0 dB). Yet, it is still questionable how much of this advantage can be replicated by a trained classifier under such low SNRs and in realistic settings. As the input SNR increases, the unified codebook begins to gain a small advantage. However, no method shows

a significant improvement on STOI compared to the noisy input signal when SNR $> 5\,\mathrm{dB}$. This is somewhat expected, as the intelligibility of the noisy input signal also increases at higher SNR and, especially above $5\,\mathrm{dB}$, the intelligibility is quite high (close to 0.9, signifying almost perfect intelligibility). With the white-box decomposition of different methods on the noise and speech component, it is clear that the unified codebook mainly improves the noise reduction ability of the system. Given the fact that the major difference between the two codebooks lies in the envelopes of low-energy frames, it is understandable that its influence on the speech distortion metric is small.

In general, the evaluation results of the oracle tests indicate that the unified codebook is more suitable for envelope enhancement than the baseline separate codebook. This improvement from the baseline is in line with our analysis of the distribution of the codebook entries of clean speech: a direct clustering of all frames is enough to create the codebook, because the envelope patterns of speech-inactive frames show low overlap with their speech-active counterparts. It is also beneficial to have a finer quantisation of the speech-inactive frames, even though their energy levels are low.

A.4.3.2 Feature Representation

No matter which codebook is chosen (unified or separate), CCoef has a consistent advantage over LPCC in terms of both metrics most of the time. LPCCs only provide a marginally higher boost to STOI when the input SNR is $-5\,\mathrm{dB}$. It is clear that CCoef is more appropriate than LPCC for this task. When the input SNR is high, the choice of codebook actually plays a more important role in the final speech quality. The NA-SSDR decomposition indicates that both features perform similarly in terms of noise reduction when the input SNR is low. CCoefs show a clear advantage in both components over LPCCs when using the separate codebook. When the codebook is generated in a unified way and the input SNR is higher than $0\,\mathrm{dB}$, there is a trade-off between noise reduction and signal distortion: CCoefs introduce less speech distortion while LPCCs suppress more noise. This divergence grows as the input SNR increases. Yet, it should be noted that the difference in noise reduction is smaller than the difference in signal distortion.

Based on the observations on the oracle tests, we can conclude that cepstral coefficients quantised by the unified codebook demonstrate the greatest potential for application in two-stage speech enhancement. Consequently, this is the feature set we shall use in the subsequent evaluations.

A.4.3.3 Quantisation Error

Comparing the best oracle classification system (the CCoef-based unified codebook) with the oracle regression system, the major difference comes up at high

SNRs on PESQ. It is interesting to note that a better envelope restoration provides more benefit if the input signal itself is of higher quality, which presumably comes from a clearer excitation signal. When the excitation signal is poorly structured, applying an improved envelope to it introduces vocoding noise. On the contrary, if the excitation signal is of good quality, a good envelope estimation can restore the underlying speech to a better degree. The quantisation makes basically no difference on STOI scores. From the PESQ and STOI scores, it is clear that even with perfect envelope estimation, there is still room for improvement in the two-stage framework with oracle envelope information, especially under low SNR conditions. Generally speaking, envelope enhancement only provides us with a substantial improvement in speech quality when the input SNR is high. This indicates that our two-stage framework is limited by other components in the system, e.g., the noise floor estimator and the quality of the initial estimate.

A.4.4 Practical System Evaluation Results

Next, we evaluate the trained speech envelope estimators on the same test set. Apart from the optimal feature set decided in Section A.4.3—the unified codebook based on cepstral coefficients (CCoefs)—we also evaluate the classifiers using CCoefs quantised by the separate codebook to verify our conclusions from the oracle test results.

A.4.4.1 Impact of Separate Codebook

First, we investigated the impact of the codebook entry assignment errors on a practical system when using the separate codebook. Figure A.6 compares the GRU classifier based on the separate and the unified codebook against three baselines: (i) the results from the preliminary denoising; (ii) oracle assignment of codebook entry based on the separate codebook; and (iii) oracle assignment of codebook entry based on the *unified* codebook.

Using envelope enhancement (with separate or unified codebooks) yielded a consistent improvement over the preliminary denoised signals. However, an interesting observation is that, when using the separate codebook, the GRU performs *better* than the corresponding oracle system when SNR is higher than 5 dB. This is true for both metrics and is more obvious with PESQ. However, this seems somewhat unsettling given that the oracle assignment should represent the upper bound. This peculiar phenomenon only makes sense when taking our previous analysis on the separate codebook into account: the energy-based partition of the clean speech signals blurs the differences among the low-energy frames and excessively simplifies many frames into a flat envelope—both of which lead to degradation of the subsequent enhancement stage. Our statement can be circumstantially verified by this extra improvement from the GRU. Compared with the LSA baseline,

the improvement by the GRU classifier on both metrics indicates that the network successfully learns to map the distorted envelopes to the pre-determined templates, although there is the noise of assignment errors in the training data due to the extra energy criterion. When the input SNR is high enough, the network is able to 'correct' the silent 'ground truth' labels to a better estimate from the speech envelope templates. It is this correcting that makes the trained classifier a better estimator than the oracle entry assignment for the separate codebook, which would force a flat spectrum on such frames. Note, however, that this GRU estimation is never as good as the oracle system using the *unified* codebook, which assigns all of the frames solely by their envelope coefficients. Further, the unified-codebook-based GRU classifier is never better than this oracle system, either. This further provides empirical validation of our analysis regarding the benefit of the unified codebook vs. the separate codebook, and the selection of the oracle assignment as the upper bound.

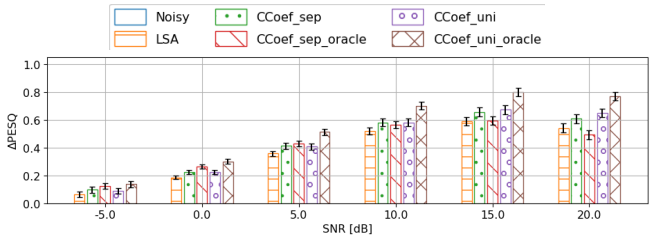
In order to further verify our hypothesis, we plot the envelope estimation cepstral MSE error of the three systems in Figure A.6c. When using the separate codebook, the GRU-classifier gains an advantage over the oracle system from 5 dB onward, which is in line with other objective metrics. This, again, provides us with a reason to choose the unified codebook over the separate one in the practical system.

A.4.4.2 Comparative Benchmark against DNN Baseline

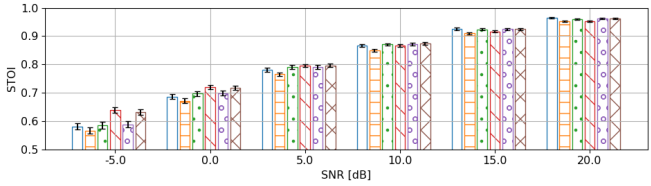
Having established the superiority of the CCoef-based feature representation and the use of a unified codebook, we now focus on benchmarking the performance of this envelope estimator against the DNN classifier baseline [11] and the CRNN-based regression approach described in Section A.3.3.3 which directly predicts the envelope coefficients. The evaluation results are shown in Figure A.7, where the results of the preliminary denoising are also included in order to better interpret the additional benefit provided by enhancing the envelope in the two-stage framework.

Generally speaking, all three methods show consistent improvement from the preliminary denoised signals. In terms of the speech quality, two classifier networks (DNN and GRU) perform similarly. GRU has a minor advantage when the input SNR is low, while DNN scores slightly higher when SNR is high. The CRNN regression network, however, shows a consistent improvement over the two classifiers at even lower computational cost.

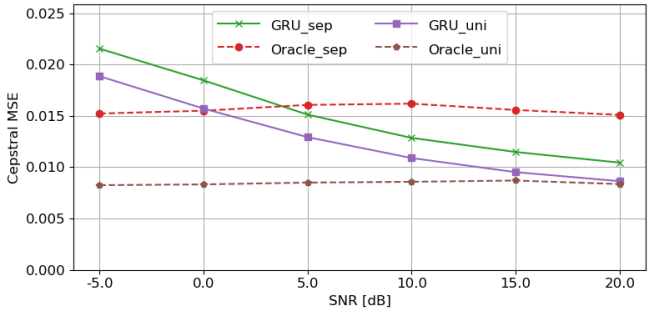
If we take a closer look at the NA-SSDR decomposition, it can be observed that, at low SNRs, the GRU and the CRNN architectures better preserve the target signal (SSDR is 1–2 dB more) compared to the DNN architecture, whereas the DNN architecture has better noise attenuation (up to $\sim$1.5 dB more) here. The better signal preservation could be due to the temporal modeling capability introduced by the recurrent layers, which is absent in the DNN architecture. In most cases,



(a) ΔPESQ compared to the noisy signals



(b) STOI



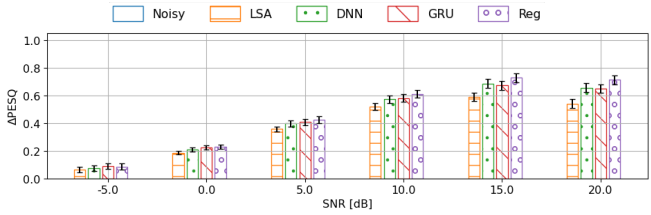
(c) Cepstral MSE

Figure A.6: Comparison of the oracle systems and the GRU−classifier based on PESQ, STOI, and cepstral coefficient prediction MSE, grouped by input SNRs.

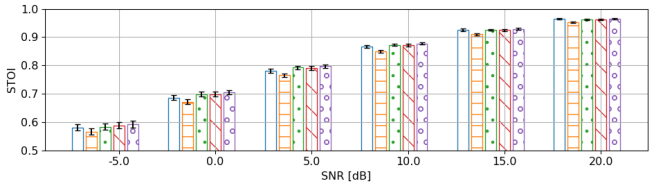
the regression network has better signal preservation than the codebook methods, the reason for which could be the inherent limitation due to quantisation in the latter methods. The choice between the GRU codebook approach and the CRNN regression approach is a trade-off between noise reduction and speech distortion.

All of the trained networks demonstrate only minor differences on STOI compared to the noisy signals. Nevertheless, compared with the preliminary denoising results, the boost to STOI is consistent. When SNR is at 0 dB or 5 dB, the networks marginally benefit the speech intelligibility.

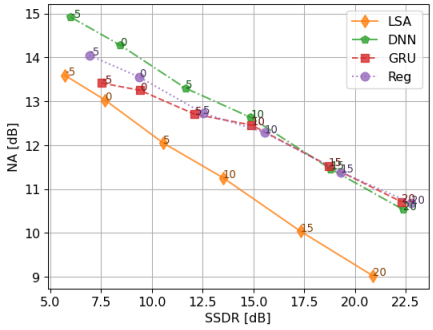
In Figure A.8, we provide two sets of samples to illustrate the performance of the proposed envelope estimators. Compared to the clean reference envelopes, it is clear that the CRNN better preserves the speech details, which gives us an intuitive impression of the performance difference of the two networks: when the speech is



(a) ΔPESQ compared to the noisy signals



(b) STOI

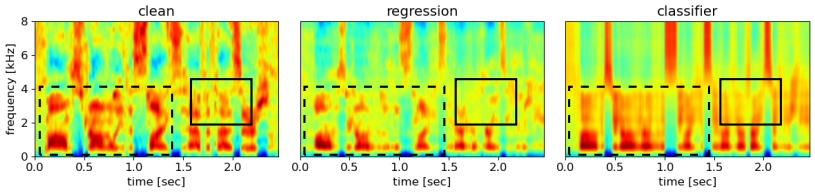


(c) NA-SSDR

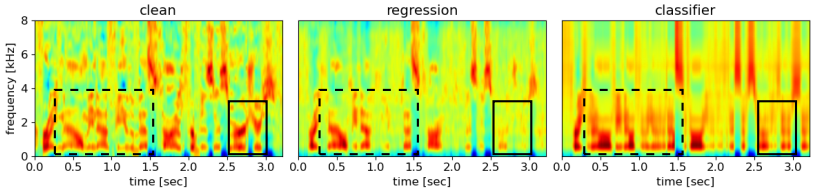
Figure A.7: Comprehensive benchmark of the proposed systems against the DNN baseline in realistic settings. Performance is evaluated by ΔPESQ, STOI and NA−SSDR, grouped by input SNRs. The 95% confidence interval is given by the error bars in the bar−plots.

better estimated in the initial stage, the regression network provides a more detailed structure of the envelope, whereas the classifier constrained by the codebook seems more beneficial when the speech is unclear. This trade-off can also be observed from the audio samples: https://aspireugent.github.io/speech-envelope-estimation/ (accessed on: 16 April 2023).

Envelope estimation can affect the excitation signal generated by CEM. Therefore, the optimal combination of two methods requires further investigation.



(a) Speech envelope estimate for speech with cafe noise at SNR $= 0$ dB



(b) Speech envelope estimate for speech with street noise at SNR $= 0$ dB

Figure A.8: Comparison of the speech envelope estimate by different estimators. We highlight the regions where the CRNN-regression network estimates a more refined structure with dashed rectangles and the regions where the GRU-classifier shows an advantage with solid rectangles.

A.5 Conclusions

In this paper, we investigated and optimised the cepstral envelope estimation for speech enhancement using the two-stage framework. Through oracle tests, we conclusively demonstrated that cepstral coefficients provide a better envelope representation compared to linear prediction cepstral coefficients. Furthermore, the manual division of the speech/non-speech frames for codebook creation was shown to be unnecessary and even *detrimental* to the system performance. Using the optimal envelope feature representation, the GRU-based classifier achieved better performance than the baseline feedforward DNN-based classifier. This performance improvement was, additionally, obtained with fewer parameters and lower computational cost. Envelope estimation could be further improved by performing a regression onto the envelope coefficients instead of utilising a codebook-based template. The CRNN network designed for the regression took the noisy input spectrum and initial gain function estimate as input and performed better with a lower computational cost in comparison with the codebook-based estimator. Compared to the initial speech estimate (preliminary denoising), all of the evaluated methods brought benefits to the quality of the enhanced signal without reducing the intelligibility.

More importantly, the oracle tests revealed that the fundamental shortcoming of the two-stage framework lay not in the envelope estimation, but in limitations

resulting from other components, such as the noise floor estimate and the statistical-model-based gain function, which performed poorly in very dynamic noise conditions.

Given a better initial estimate of the underlying speech signal, the proposed envelope estimators could be integrated into the signal processing pipeline in post-processing or as a second neural network focusing on the envelope estimation.

In summary, if the goal is to have improved single-microphone noise suppression within an interpretable, controllable, low-cost framework, then the work presented in this paper may be a good option. On the other hand, end-to-end enhancement can yield better noise suppression and speech quality, but at the cost of higher computational expense, poorer interpretability, and lack of control possibilities.

References

[1] Y. Ephraim and D. Malah. *Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator*. IEEE Transactions on acoustics, speech, and signal processing, 32(6):1109–1121, 1984.

[2] Y. Ephraim and D. Malah. *Speech enhancement using a minimum mean-square error log-spectral amplitude estimator*. IEEE transactions on acoustics, speech, and signal processing, 33(2):443–445, 1985.

[3] C. Plapous, C. Marro, and P. Scalart. *Improved signal-to-noise ratio estimation for speech enhancement*. IEEE transactions on audio, speech, and language processing, 14(6):2098–2108, 2006.

[4] Y. Song and N. Madhu. *Improved CEM for speech harmonic enhancement in single channel noise suppression*. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30:2492–2503, 2022.

[5] S. Elshamy, N. Madhu, W. Tirry, and T. Fingscheidt. *Instantaneous a priori SNR estimation by cepstral excitation manipulation*. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(8):1592–1605, 2017.

[6] A. M. C. Martinez, C. Spille, J. Roßbach, B. Kollmeier, and B. T. Meyer. *Prediction of speech intelligibility with DNN-based performance measures*. Computer Speech & Language, 74:101329, 2022.

[7] K. Arai, S. Araki, A. Ogawa, K. Kinoshita, T. Nakatani, K. Yamamoto, and T. Irino. *Predicting Speech Intelligibility of Enhanced Speech Using Phone Accuracy of DNN-Based ASR System.* In Interspeech, pages 4275–4279, 2019.

[8] C. Yağlı, M. T. Turan, and E. Erzin. *Artificial bandwidth extension of spectral envelope along a Viterbi path*. Speech communication, 55(1):111–118, 2013.

[9] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen. *A short-time objective intelligibility measure for time-frequency weighted noisy speech*. In 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, pages 4214–4217. IEEE, 2010.

[10] S. Srinivasan, J. Samuelsson, and W. B. Kleijn. *Codebook driven short-term predictor parameter estimation for speech enhancement*. IEEE Transactions on Audio, Speech, and Language Processing, 14(1):163–176, 2005.

[11] S. Elshamy, N. Madhu, W. Tirry, and T. Fingscheidt. *DNN-supported speech enhancement with cepstral estimation of both excitation and envelope*. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(12):2460–2474, 2018.

[12] R. Chen, C.-F. Chan, and H. C. So. *Model-based speech enhancement with improved spectral envelope estimation via dynamics tracking*. IEEE Transactions on Audio, Speech, and Language Processing, 20(4):1324–1336, 2011.

[13] Y. Li and S. Kang. *Deep neural network-based linear predictive parameter estimations for speech enhancement*. IET Signal Processing, 11(4):469–476, 2017.

[14] M. Kolbæk, Z.-H. Tan, and J. Jensen. *Monaural speech enhancement using deep neural networks by maximizing a short-time objective intelligibility measure*. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5059–5063. IEEE, 2018.

[15] J.-M. Valin, U. Isik, N. Phansalkar, R. Giri, K. Helwani, and A. Krishnaswamy. *A perceptually-motivated approach for low-complexity, real-time enhancement of fullband speech*. arXiv preprint arXiv:2008.04259, 2020.

[16] J. L. Carmona, J. Barker, A. M. Gomez, and N. Ma. *Speech spectral envelope enhancement by HMM-based analysis/resynthesis*. IEEE Signal Processing Letters, 20(6):563–566, 2013.

[17] J. Makhoul. *Linear prediction: A tutorial review*. Proceedings of the IEEE, 63(4):561–580, 1975.

[18] C. Breithaupt, T. Gerkmann, and R. Martin. *Cepstral smoothing of spectral filter gains for speech enhancement without musical noise*. IEEE Signal processing letters, 14(12):1036–1039, 2007.

[19] C. Breithaupt, T. Gerkmann, and R. Martin. *A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing*. In 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, pages 4897–4900. IEEE, 2008.

[20] R. M. Nickel, R. F. Astudillo, D. Kolossa, and R. Martin. *Corpus-based speech enhancement with uncertainty modeling and cepstral smoothing*. IEEE Transactions on Audio, Speech, and Language Processing, 21(5):983–997, 2013.

[21] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. *Empirical evaluation of gated recurrent neural networks on sequence modeling*. arXiv preprint arXiv:1412.3555, 2014.

[22] S. Braun, H. Gamper, C. K. Reddy, and I. Tashev. *Towards efficient models for real-time deep noise suppression*. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 656–660. IEEE, 2021.

[23] Y. Song and N. Madhu. *Aiding speech harmonic recovery in DNN-based single channel noise reduction using cepstral excitation manipulation (CEM) components (in press)*. In 2023 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2023.

[24] Y. Linde, A. Buzo, and R. Gray. *An algorithm for vector quantizer design*. IEEE Transactions on communications, 28(1):84–95, 1980.

[25] P. Vary and R. Martin. *Digital speech transmission: Enhancement, coding and error concealment*. John Wiley & Sons, 2006.

[26] *Objective Measurement of Active Speech Level*. Number Rec. P.56. International Telecommunication Union-Telecommunication Standardisation Sector (ITU), 2011.

[27] T. Gerkmann and R. C. Hendriks. *Unbiased MMSE-based noise power estimation with low complexity and low tracking delay*. IEEE Transactions on Audio, Speech, and Language Processing, 20(4):1383–1393, 2011.

[28] *Corrigendum 1, Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs*. Number Rec. P.862.2. International Telecommunication Union-Telecommunication Standardisation Sector (ITU), 2017.

B

List of Acronyms

AR	auto-regressive
ASR	automatic speech recognition
B-LSTM	bidirectional long short-term memory
BCE	binary cross-entropy
CC	cepstral convolution
CCoef	cepstral coefficient

CEE	cepstral envelope estimation
CEM	cepstral excitation manipulation
CNN	convolutional neural network
CPC	contrastive predictive coding
CRNN	convolutional recurrent neural network
CRUSE	convolutional recurrent U-net architecture for speech enhancement
DCT	discrete cosine transform
DCT-II	discrete cosine transformation of type II
DD	decision-directed
DFT	discrete Fourier transform
DNN	deep neural network
DNSMOS	deep noise suppression mean opinion score
F0	fundamental frequency
FC	fully connected
GAN	generative adversarial network
GLA	Griffin-Lim algorithm
GMM	Gaussian mixture model
GRU	gated recurrent unit
HMM	hidden Markov model
HuBERT	hidden unit bidirectional encoder representations from transformers
iDFT	inverse discrete Fourier transform
IRM	ideal ratio mask
iSTFT	inverse short-time Fourier transform
LBG	Linde–Buzo–Gray

LPC	linear prediction coding
LPCC	linear prediction coding coefficient
LSTM	long-short term memory
MAC	multiply-accumulate operation
MAP	maximum a posteriori
MMSE-LSA	minimum mean-square error log-spectral amplitude
MMSE-SPP	minimum mean square error-speech presence probability
MMSE-STSA	minimum mean-square error short-time spectral amplitude
MOS-LQO	mean opinion score - listening quality objective
MPD	multi-period discriminator
MS	minimum statistics
MSD	multi-scale discriminator
MMSE	minimum mean square error
MSE	mean-square error
NA	noise attenuation
NISQA	non-intrusive objective speech quality assessment
NLL	negative log-likelihood
PESQ	perceptual evaluation of speech quality
PGHI	phase gradient heap integration
POLQA	perceptual objective listening quality analysis
PSD	power spectral density
PSM	phase sensitive mask
RAE	residual amplitude estimation
RIR	room impulse response

RMS	root-mean-square
RNN	recurrent neural network
SE	speech enhancement
SI-SDR	scale-invariant signal-to-distortion ratio
SNR	signal-to-noise ratio
SS	spectral subtraction
SSDR	speech-to-speech-distortion ratio
SSL	self-supervised learning
STFT	short-time Fourier transform
STOI	short-time objective intelligibility
SUPERB	speech processing universal performance benchmark
TERA	transformer encoder representations from alteration
TF	time-frequency
TSNR	two-step noise reduction
TTS	text-to-speech
VAD	voice activity detector
WB-PESQ	wide-band perceptual evaluation of speech quality