Université du Québec
Institut national de la recherche scientifique
Centre Énergie Matériaux Télécommunications

APPLICATIONS OF DEEP LEARNING TO SPEECH ENHANCEMENT

By

João Felipe Santos

A thesis submitted in fulfillment of the requirements for the degree of
Doctorate of Sciences, Ph.D
in Telecommunications

Evaluation Committee

Internal evaluator and committee president:	Prof. Douglas O'Shaughnessy
External evaluator 1:	Prof. Timo Gerkmann Universität Hamburg
External evaluator 2:	Prof. Emmanuel Vincent INRIA
Research advisor:	Prof. Tiago H. Falk INRS-EMT

Acknowledgements

This work would not have been possible without the support of so many colleagues and collaborators. Starting at MuSAE lab, I would like to acknowledge the support and supervision of professor Tiago Falk, and colleagues and collaborators Milton Sarria, Diana Tobón-Vallejo, Stefany Bedoya, Rachelle Bouserhal, Mohammed Sennoussaoui, Anderson Ávila, Raymundo Cassani, João Monteiro, and Isabela Albuquerque.

During my time as a visiting researcher and applied AI scientist at Mila, I have profited a lot from conversations and collaborations with professors Yoshua Bengio, Aaron Courville, Roland Memisevic, Christopher Pal, and students Kyle Kastner, José Sotelo, Soroush Mehri, Kundan Kumar, Jessica Thompson, Chiheb Trabelsi, Dmitriy Serdyuk, and Loren Lugosch.

I would also like to acknowledge my external collaborators Kostantinos Drossos (TUT), Sebastian Braun (Microsoft Research), Bob Sturm (KTH), Iryna Korshunova (Ghent University), Benjamin Cauchi, and Stylianos Mimilakis (both at Fraunhofer IMDT).

This work has been financially supported by the Fonds de Recherche du Québec - Nature et Technologies (FRQNT) and by the Natural Sciences and Engineering Research Council of Canada (NSERC). Many of the experiments were performed on hardware donated by NVIDIA through their GPU Grant Program.

Abstract

Deep neural networks (DNNs) have been successfully employed in a broad range of applications, having achieved state-of-the-art results in tasks such as acoustic modeling for automatic speech recognition and image classification. However, their application to speech enhancement problems such as denoising, dereverberation, and source separation is more recent work and therefore in more preliminary stages. In this work, we explore DNN-based speech enhancement from three different and complementary points of view.

First, we propose a model to perform speech dereverberation by estimating its spectral magnitude from the reverberant counterpart. Our models are capable of extracting features that take into account both short and long-term dependencies in the signal through a convolutional encoder and a recurrent neural network for extracting long-term information. Our model outperforms a recently proposed model that uses different context information depending on the reverberation time, without requiring any sort of additional input, yielding improvements of up to 0.4 on PESQ, 0.3 on STOI, and 1.0 on POLQA relative to reverberant speech. We also show our model is able to generalize to real room impulse responses even when only trained with simulated room impulse responses, different speakers, and high reverberation times. Lastly, listening tests show the proposed method outperforming benchmark models in reduction of perceived reverberation.

We also study the role of residual and highway connections in deep neural networks for speech enhancement, and verify whether they function in the same way that their digital signal processing counterparts. We visualize the outputs of such connections, projected back to the spectral domain, in models trained for speech denoising, and show that while skip connections do not necessarily improve performance with regards to the number of parameters, they make speech enhancement models more interpretable. We also discover, through visualization of the hidden units of the context-aware model we proposed, that many of the neurons are in a state that we call "stuck" (i.e. their outputs is a constant value different from zero). We propose a method to prune those neurons away from the model without having an impact in performance, and compare this method to other methods in the literature. The proposed method can be applied post hoc to any pretrained models that use sigmoid or hyperbolic tangent activations. It also leads to dense models, therefore not requiring special software/hardware to take advantage of the compressed model.

Finally, in order to investigate how useful current objective speech quality metrics are for DNN-based speech enhancement, we performed online listening tests using the outputs of three different DNN-based speech enhancement models for both denoising and dereverberation. When assessing the predictive power of several objective metrics, we found that existing non-intrusive methods fail at monitoring signal quality. To overcome this limitation, we propose a new metric based on a combination of a handful of relevant acoustic features. Results inline with those obtained with intrusive measures are then attained. In a leave-one-model-out test, the proposed non-intrusive metric is also shown to outperform two non-intrusive benchmarks for all three DNN enhancement methods, showing the proposed method is capable of generalizing to unseen models. We then take the first steps in incorporating such knowledge into the training of DNN-based speech enhancement models by designing a quality-aware cost function. While our approach increases PESQ scores for a model with a vocoder output, participants of a small-scale preference test considered it less natural than the same model trained with a conventional MSE loss function, which signals more work is needed in the development of quality-aware models.

Keywords: speech enhancement, dereverberation, denoising, deep neural networks, recurrent neural networks, interpretability, pruning, speech quality

Contents

List of Figures

List of Tables

List of Abbreviations

ANN	artificial neural network
ASR	automatic speech recognition
BSS	blind source separation
CD	cepstral distance
CNN	convolutional neural network
CRM	complex ratio masking
DNN	deep neural network
F0	fundamental frequency
FWSegSNR	frequency-weighted segmental signal-to-noise ratio
GPU	graphics processing unit
GRU	gated recurrent unit
IBM	ideal binary mask
IRM	ideal ratio mask
LLR	log-likelihood ratio
LOO	leave-one-out
LP	linear prediction
LPC	linear predictive coding
LSTM	long short-term memory
MGC	mel-generalized cepstrum
MMSE	minimum mean-squared error
MOS	mean opinion score
MSE	mean-squared error
PESQ	Perceptual Evaluation of Speech Quality

POLQA	Perceptual Objective Listening Quality Assessment
ReLU	rectified linear unit
RIR	room impulse response
RNN	recurrent neural network
segSNR	segmental signal-to-noise ratio
SGD	stochastic gradient descent
SNR	signal-to-noise ratio
SRMR	speech-to-reverberation modulation energy ratio
STFT	short-time Fourier transform
STOI	short-time objective intelligibility
STSA	short-time spectral amplitude
SVM	support vector machine
T60	reverberation time
WER	word error rate

Synopsis

0.1 Introduction

Ces dernières annés, les technologies de la parole sont devenues omniprésentes. Grâce aux progrès de la reconnaissance automatique de la parole (*automatic speech recognition* en anglais, ou ASR — nous utiliserons les acronymes anglais afin d'éviter toute confusion), la parole est devenue un moyen d'interagir avec de nombreux appareils et logiciels. Les applications vocales vont au-delà de la simple transcription et permettent désormais une interaction avec nos télévisions, téléphones et voitures. Dans ces applications, les appareils sont passés d'environnements acoustiques contrôlés, tels que des salles calmes avec des microphones de proximité, à des environnements acoustiques complexes, où des microphones en champ lointain sont utilisés en présence de différents types de bruit, de réverbération et d'autres locuteurs. Les distorsions ajoutées par ce type d'environnement acoustique entraînent une dégradation importante des performances de l'ASR, ainsi qu'une réduction de l'intelligibilité et de la qualité de la parole dans les systèmes de communication vocale, en particulier pour les malentendants.

Dans la plupart des applications réelles, le signal de parole clair est déformé par une combinaison de différentes catégories de distorsions avant d'être capturé par un appareil. Dans les environnements clos, par exemple, les distorsions sont généralement modélisées comme une combinaison de bruit de fond additif et de réverbération. Pour traiter les distorsions dans de tels environnements, plusieurs types de systèmes d'amélioration de la parole ont été proposés, allant de systèmes monocanaux basés sur une simple soustraction spectrale aux systèmes à plusieurs étages qui exploitent les signaux de multiples microphones.

Récemment, les réseaux de neurones profonds (*deep neural network* en anglais, ou DNN) ont été utilisés avec succès dans un large éventail d'applications. Ils ont permis d'obtenir des résultats de pointe dans des tâches telles que la modélisation acoustique pour l'ASR et la classification d'images. Deux phénomènes expliquent ces progrès : premièrement, l'évolution des systèmes informatiques et des accélérateurs matériels tels que les unités de traitement graphique (*graphics processing unit* en anglais, ou GPU) a entraîné une augmentation significative de la puissance de calcul et des capacités de stockage, deux conditions essentielles à l'étape d'apprentissage d'un réseau de neurones à plusieurs couches. Deuxièmement, les progrès des algorithmes d'optimisation et des architectures de réseau tels que les réseaux de neurones récurrents et convolutionnels ont amélioré la capacité de tels systèmes à apprendre les dépendances spatiales et temporelles des données.

Les DNN jouent un rôle important dans les systèmes ASR à grande échelle utilisés présentement, tant dans les modèles acoustiques que pour le prétraitement améliorant la robustesse en présence de réverbération. Cependant, leur application aux problèmes d'amélioration de la parole tels que le débruitage, la déréverbération et la séparation des sources est un travail plus récent et, par conséquent, plus préliminaire.

Contrairement aux algorithmes traditionnels d'amélioration de la parole basés sur des estimateurs d'erreur quadratique moyenne, un système d'amélioration de la parole basé sur un DNN ne nécessite pas de solution analytique au problème de l'estimation et ne se limite pas à des modèles statistiques simples (tels que des variables indépendantes et identiquement distribuées). Il ne nécessite qu'une fonction de coût différentiable et est formé à l'aide d'algorithmes d'optimisation stochastiques, généralement à partir de grandes quantités de données. Cela permet, par exemple, de concevoir un jeu de données d'apprentissage contenant plusieurs types de scénarios de distorsion. La formation de tels réseaux demande beaucoup de temps et de puissance de calcul, mais les réseaux ainsi formés, eux, nécessitent peu de ressources.

Malgré la flexibilité des DNN, la plupart des solutions existantes reposent sur des architectures à propagation avant avec une fonction de coût d'erreur quadratique moyenne, qui est simple à optimiser, mais qui n'est pas appropriée pour optimiser la qualité et à l'intelligibilité de la parole.

0.1.1 Débruitage et déréverbération de la parole sur un seul canal

La réverbération et le bruit ont des effets complémentaires sur un signal clair de parole. L'équation suivante modélise de façon simple un signal de parole déformé par une combinaison de bruit et de réverbération:

$$y(n) = h(n) * x(n) + \nu(n)$$

où la réverbération est représentée par la convolution de la réponse impulsionnelle de la pièce (*room impulse response* en anglais, ou RIR) au signal clair et forme un bruit additif.

La réverbération entraîne deux effets de perception importants sur un signal clair de parole: les réflexions initiales provoquent une coloration spectrale du signal, tandis que la réverbération ultérieure provoque un étalement temporel. Ses effets sont particulièrement importants pour les auditeurs malentendants, pour qui l'intelligibilité de la parole est réduite même en présence de temps de réverbération relativement faibles. Les distorsions additives de bruit affectent l'intelligibilité de la parole différemment: les consonnes faibles subissent plus de masquage que les voyelles de plus forte intensité, et cet effet ne dépend pas de l'énergie des segments précédents (comme c'est le cas pour la réverbération).

0.1.2 Défis

Les approches existantes pour l'amélioration de la parole ont plusieurs limites. Les modèles de parole et de bruit sont souvent des distributions simplifiées des coefficients d'amplitude (gaussien, super gaussien, etc.). Cela est nécessaire afin de trouver des solutions analytiques au problème d'estimation.

La plupart des algorithmes actuels d'amélioration de la parole ont une fonction objectif simplifiée qui ne correspond pas aux caractéristiques de perception. Une mesure de distorsion commune utilisée est l'erreur quadratique moyenne des coefficients spectraux, qui ne tient pas compte du masquage auditif ni des distorsions dues à la phase. Cela conduit à des effets indésirables, tels que la génération d'artefacts de perception.

0.1.3 Contributions de la thèse

Le travail présenté dans cette thèse explore l'utilisation des DNN pour l'amélioration de la parole de trois manières différentes et complémentaires. Premièrement, nous proposons un modèle afin d'effectuer la déréverbération de la parole en estimant sa magnitude spectrale à partir de la contrepartie réverbérante. Notre modèle est capable d'extraire des caractéristiques qui tiennent compte des dépendances à court et à long terme du signal via un codeur convolutionnel et un réseau de neurones récurrent afin d'extraire des informations à long terme. Notre modèle surpasse un modèle proposé récemment qui utilise différentes informations de contexte en fonction de la durée de réverbération, sans nécessiter d'apport supplémentaire, générant des améliorations allant jusqu'à 0.4 sur la mesure *Perceptual Evaluation of Speech Quality* (PESQ), 0.3 sur *Short-Time Objective Intelligibility* (STOI) et 1.0 sur *Perceptual Objective Listening Quality Analysis* (POLQA) par rapport au discours réverbérant. Nous montrons également que notre modèle fonctionne avec des réponses impulsionnelles de salles réelles bien s'il ne soit entraîné qu'avec des simulations de réponses impulsionnelles de salle, des locuteurs différents et des temps de réverbération élevés. Enfin, des tests d'écoute montrent que la méthode proposée surpasse les modèles de référence en termes de réduction de la réverbération perçue.

Nous étudions également le rôle des connexions résiduelles et des "*highway connections*" dans les DNN pour l'amélioration de la parole et vérifions si elles fonctionnent de la même manière que leurs homologues en traitement de signal numérique. Nous visualisons les sorties de telles connexions, reprojetées dans le domaine spectral, dans des modèles formés au débruitage de la parole, et montrons que, si les connexions sautées n'améliorent pas nécessairement les performances en ce qui concerne le nombre de paramètres, elles rendent les modèles d'amélioration de la parole plus facilement interprétables. Nous découvrons également, par la visualisation des unités cachées du modèle sensible au contexte que nous avons proposé, que de nombreux neurones sont dans un état que nous appelons "bloqué" (c'est-à-dire que leurs sorties sont une valeur constante différente de zéro). Nous proposons une méthode pour éliminer ces neurones du modèle qui n'a pas d'impact sur les performances la comparons à d'autres méthodes publiées. La méthode proposée peut être appliquée post-hoc à tout modèle pré-entraîné utilisant des activations sigmoïdiennes ou à tangentes hyperboliques. Cela conduit également à des modèles denses, ne nécessitant donc pas de logiciel ni de matériel spécial afin de tirer parti du modèle compressé.

En vue de l'amélioration de modèles DNN d'amélioration de la parole, nous avons évalué l'utilité des mesures objectives actuellement utilisées à l'aide de tests d'écoute en ligne en utilisant les sorties de trois modèles pour le débruitage et la déréverbération. Lors de l'évaluation du pouvoir prédictif de plusieurs mesures objectives, nous avons constaté que les méthodes non intrusives existantes ne permettent pas d'évaluer la qualité du signal. Pour surmonter cette limitation, nous proposons une nouvelle mesure basée sur une combinaison de quelques caractéristiques acoustiques pertinentes. Des résultats conformes à ceux obtenus avec des mesures intrusives sont alors atteints. Dans un test ne laissant qu'un modèle (*leave-one-model-out test*), il est également démontré que la mesure non intrusive proposée surpasse deux normes de référence non intrusives pour les trois modèles DNN, montrant que la méthode proposée est généralisable. Nous procédons ensuite à l'incorporation de ces connaissances à la formation des modèles DNN d'amélioration de la parole en concevant une fonction de coût tenant compte de la qualité. Bien que notre approche augmente les scores PESQ pour un modèle avec une sortie de vocodeur, les participants à un test de préférence à petite échelle l'ont considéré moins naturel que le même modèle formé avec une fonction de perte MSE conventionnelle, ce qui indique que davantage de travail est nécessaire afin de développer de modèles d'aussi bonne qualité que les modèles conventionnels.

0.1.4 Organization de la thèse

Cette thèse est organisée comme suit. Le chapitre 1 présente une brève motivation pour l'amélioration de la parole basée sur DNN, ses défis, et résume les contributions. Le chapitre 2 contient un résumé rapide des systèmes d'amélioration de la parole, des réseaux de neurones profonds et de l'état de la technique en matière d'amélioration de la parole basée sur DNN. Le chapitre 3 présente notre modèle de déréverbération de la parole sensible au contexte. Le chapitre 4 comprend nos travaux sur l'interprétation des connexions sautées dans les modèles d'amélioration de la parole et sur la méthode d'élagage des neurones que nous avons proposée en nous basant sur nos observations d'activations internes. Au chapitre 5, nous présentons les résultats des tests d'écoute que nous avons effectués afin de mieux comprendre les performances des modèles d'amélioration de la parole basés sur le DNN et leur corrélation avec les métriques de qualité objective, et les premières étapes vers des modèles d'amélioration de la parole basées sur le DNN et informées par la qualité. Enfin, le chapitre 6 contient des considérations sur les travaux présentés dans cette thèse et quelques idées pour des travaux futurs.

0.2 Résumé

0.2.1 Chapitre 2: Contexte

Le chapitre 2 contient une discussion plus approfondie du problème de l'amélioration de la parole, ainsi qu'une description de plusieurs approches différentes en matière de débruitage et de déréverbération de la parole. Nous discutons brièvement de certaines approches qui exploitent des informations sur l'environnement acoustique, la qualité de la parole et l'intelligibilité. Nous discutons également des principes de base des réseaux de neurones profonds et des deux principales composantes utilisées plus tard dans la thèse : les réseaux de neurones convolutionnels et récurrents. Enfin, nous présentons un bref aperçu de l'état de l'art de la recherche en matière d'amélioration de la parole basée sur les DNN, en abordant les approches à la fois du domaine fréquentiel et temporel.

Sommaire des méthodes d'amélioration de la parole

Plusieurs algorithmes d'amélioration de la parole ont été proposés pour traiter le bruit et la réverbération. Parmi les types d'algorithmes d'amélioration de la parole les plus largement utilisés, on trouve les méthodes d'amélioration de l'amplitude spectrale à court terme ("*short-time spectral amplitude*", STSA) à canal unique, telles que la soustraction spectrale et les estimateurs basés sur la minimisation d'une métrique de distance dans le domaine de l'amplitude spectrale à court terme, sous forme d'erreur quadratique moyenne ou, plus récemment, de distances inspirées de façon perceptuelle. Bien que les dispositifs multicanaux soient couramment utilisés et permettent des approches de filtrage spatial telles que la formation de faisceau, l'amélioration spectrale sur un seul canal est généralement réalisée comme une étape de post-filtrage.

Une hypothèse courante dans les méthodes d'estimation spectrale fondées sur l'erreur minimale moyenne (MMSE) est que les signaux de parole et de bruit ont des amplitudes spectrales à court terme dont les coefficients peuvent être modélisés comme des variables aléatoires gaussiennes statistiquement indépendantes (pour le modèle de parole, des distributions Laplace, Gamma, ou super-gaussiennes ont également été proposées). Une telle hypothèse implique que les coefficients de Fourier de la parole ne sont pas corrélés; Cependant, il est connu que les signaux de parole ont une structure forte, à la fois dans le temps et dans la fréquence, en raison de la structure des formants et des harmoniques. Bien que l'hypothèse simplifie le calcul des solutions analytiques au problème (comme dans l'approche MMSE), elle n'est valable que dans un sens asymptotique et pour de très grandes tailles de trame, ce qui ne s'applique pas dans les cas pratiques car la taille des trames dans de tels systèmes est généralement réduite. de l'ordre de 20-40 ms).

D'autres solutions comme l'approche subspatiale, amélioration sur le domaine de la prédiction linéaire, filtrage sur le domaine du spectre de modulation, et des approches de déconvolution aveugle ont été aussi explorées.

Approches informés par l'environement, la qualité et l'intelligibilité

La plupart des systèmes d'amélioration de la parole dans la littérature utilisent uniquement des informations minimales sur l'environnement. Les méthodes de réduction du bruit, par exemple, nécessitent souvent une estimation du rapport signal sur bruit ("*signal-to-noise ratio*", SNR) a priori et une estimation du spectre du bruit. Certains systèmes de déréverbération, comme mentionné dans la session précédente, effectuent un filtrage inverse, ce qui nécessite une estimation de la réponse impulsionnelle de la pièce, tandis que d'autres utilisent uniquement une estimation du temps de réverbération pour calculer une estimation des effets de la pièce.

Réseaux de neurones profondes

Les réseaux de neurones artificiels ("*artificial neural networks*", ANN) constituent une catégorie de modèles d'apprentissage statistique inspirés par le fonctionnement des réseaux de neurones biologiques. L'élément principal d'un ANN est constitué de couches d'unités neuronales, qui sont des applications non linéaires d'un vecteur à un autre vecteur (pas nécessairement de même longueur) et peuvent être représentées par la relation suivante:

$$h = g(Wx + b)$$

où h est la sortie de la couche, g est une fonction non linéaire telle qu'un sigmoïde, une tangente hyperbolique ou une ReLU (max(ů, 0)), x est l'entrée et W et b sont les paramètres de poids et de biais de la couche, respectivement, et sont celles à apprendre d'un jeu de données. Toute couche non connectée à l'entrée ou à la sortie est appelée couche masquée. Un réseau neuronal profond ("*deep neural network*", DNN) est une composition de plusieurs couches d'unités neuronales, qui effectue des projections séquentielles non linéaires de l'entrée de la couche précédente. Cette composition de plusieurs couches permet à un DNN d'apprendre plusieurs couches d'abstractions qui représentent une entrée donnée à partir de données brutes, par opposition à d'autres algorithmes d'apprentissage automatique qui nécessitent souvent une ingénierie de fonctionnalités de l'utilisateur.

Les DNN ont récemment été utilisés dans une large gamme de tâches, telles que la vision par ordinateur et la reconnaissance vocale, pour obtenir des résultats impressionnants. Les applications les plus récentes des DNN aux signaux multimédias utilisent des variantes de l'architecture DNN illustrée ici, à savoir les réseaux neuronaux récurrents et convolutifs, qui conviennent mieux au travail avec des données séquentielles et des données composées de plusieurs matrices, respectivement.

Les réseaux de neurones récurrents ("*recurrent neural networks*", RNN) ont été appliqués avec succès au traitement de données séquentiel dans de nombreux domaines, tels que les modèles acoustiques dans l'ASR et la traduction automatique, ainsi que dans l'amélioration de la parole. La principale différence entre un réseau de neurones récurrent et les réseaux de neurones à rétroaction traditionnels réside dans le fait que leurs états masqués ne sont pas seulement fonction des entrées de couche, mais également de l'état de couche masqué actuel. Cette connexion récurrente permet au réseau de «mémoriser» la représentation passée de ses entrées et de les prendre en compte, ainsi que de nouvelles entrées, ce qui s'est révélé utile pour modéliser la dynamique de données séquentielles telles que les signaux de parole. En pratique, la plupart des travaux en cours sur RNN utilisent des unités gated, telles que la mémoire longue à court terme (LSTM) ou les unités récurrentes gated (GRU), pour traiter des problèmes de gradients disparus et explosés sur l'entraînement. Ces unités utilisent un mécanisme de déclenchement pour contrôler le flux d'informations entrant et sortant de l'unité.

Les réseaux de neurones convolutifs ("*convolutional neural networks*", CNN) sont une variante des réseaux de neurones à anticipation inspirés par le fonctionnement des champs récepteurs dans le cortex visuel. La transformation affine dans chaque couche avant l'opération de non-linéarité est remplacée par une opération de convolution (généralement une convolution 2D pour des entrées telles que des images ou des spectrogrammes) avec un noyau plus petit que l'entrée. Ce faisant, les CNN identifient des interactions moins importantes que les réseaux à anticipation, car leur connectivité est locale. En raison de l'opération de convolution, chaque noyau (ou filtre) est répliqué sur toute l'entrée, ce qui permet au réseau d'apprendre à détecter des entités indépendamment de leur «positionnement spatial» dans l'entrée. Les CNN peuvent également fonctionner sur des entrées multicanaux et les sorties de plusieurs noyaux forment une carte de caractéristiques dans chaque couche cachée. Les CNN sont populaires dans le traitement des images, mais peuvent également être utilisés dans des applications de parole, soit en effectuant des convolutions 1D, soit en considérant le spectrogramme (ou toute autre représentation temps-fréquence) d'un signal de parole comme une entrée 2D. L'idée de traiter les entrées multicanaux peut également être explorée dans les applications vocales à plusieurs microphones.

Etat de l'art dans l'amélioration de la parole avec des modèles DNN

Récemment, plusieurs travaux ont exploré les réseaux de neurones profonds pour l'amélioration de la parole selon deux approches principales: l'estimation spectrale et le masquage spectral. Dans le premier, l'objectif du réseau de neurones est de prédire directement le spectre de magnitude d'un signal de parole amélioré, tandis que le dernier vise à prédire une forme de masque idéal (masque binaire ou masque de rapport) à appliquer à l'entrée distordue. La cible de masque binaire idéale transforme le problème d'amélioration de la parole en problème de classification, l'objectif du modèle étant de prédire quelles cellules temps-fréquence de l'entrée doivent être masquées et il a été démontré que l'amélioration de l'intelligibilité était améliorée avec ces modèles. Le masque binaire idéal est défini quantitativement en fonction d'un seuil de critère local pour le rapport signal sur bruit. À savoir, si le SNR d'une cellule temps-fréquence donnée est inférieur au seuil, cette cellule temps-fréquence est masquée. En variante, le masque de rapport idéal est étroitement lié au filtre de Wiener dans le domaine fréquentiel avec parole et bruit non corrélés. Il s'agit d'une technique de masquage souple dans laquelle la valeur du masque correspond au rapport local entre les énergies du signal et du bruit plus le bruit pour chaque cellule temps-fréquence. Des masques complexes aux ratios idéaux ont également été récemment proposés, dans lesquels le masque est appliqué aux composants réels et imaginaires du STFT au lieu de la magnitude.

La plupart des travaux publiés sur le sujet utilisent un réseau de neurones multicouche "feed-forward" relativement large, avec trois couches cachées contenant plusieurs centaines d'unités, ayant en entrée une fenêtre de contexte contenant un nombre arbitraire de trames du spectre de magnitude logarithmique et ciblant la trame centrale à sec de cette fenêtre. La plupart des modèles mentionnés dans la littérature n'explorent qu'un des deux contextes possibles à partir du signal corrompu. Les modèles avec une fenêtre fixe d'un nombre arbitraire de trames passées et futures ne prennent en compte que le contexte local et sont incapables de représenter la structure à long terme du signal. De plus, les modèles "feed-forward" n'ayant pas d'état interne conservé entre les trames, ils ne prennent pas en compte les trames prédites précédemment, ce qui peut entraîner des artefacts dus à des discontinuités spectrales. Les architectures basées sur RNN, d'autre part, peuvent apprendre la structure à court et à long terme. Cependant, l'apprentissage de l'une ou l'autre de ces structures n'est pas imposé par l'algorithme d'apprentissage ni par l'architecture. Il est donc impossible de contrôler si l'état interne représentera un contexte à court terme, à long terme, ou les deux.

Enfin, depuis l'avènement de WaveNet, un modèle de synthèse vocale opérant dans le domaine temporel, plusieurs études se sont concentrées sur l'amélioration directement dans le domaine temporel au lieu de convertir le signal dans le domaine fréquentiel en tant qu'étape intermédiaire. Ces approches évitent la réutilisation de la phase bruyante pour la resynthèse du signal, car la sortie de ces modèles est un signal temporel. Il convient de noter que les modèles WaveNet et dérivés sont très coûteux du point de vue informatique, en raison de leur taille et du nombre d'échantillons en cours de traitement, bien que des travaux soient en cours pour l'optimiser en termes de performances. D'autres architectures alternatives pour les signaux temporels ont été proposées, la plupart d'entre elles étant basées sur des couches convolutives 1D avec une structure UNet. Les réseaux accusatoires génératifs ("*generative adversarial networks*", GAN) ont également été explorés en tant que fonctions de coût alternatives afin que les modèles génèrent de la parole plus naturel.

0.2.2 Chapitre 3: Déréverbération de la parole avec réseaux de neurones récurrents sensibles au contexte

Dans ce chapitre, nous proposons une nouvelle architecture de déréverbération de la parole qui exploite les informations de contexte à court et à long terme. Tout d'abord, les informations de contexte local fixes sont générées directement à partir de la séquence d'entrée par un codeur de contexte convolutionnel. Nous formons le réseau à apprendre à utiliser les informations de contexte à long terme en utilisant des couches récurrentes et en le formant à améliorer des phrases entières à la fois, au lieu d'une trame à la fois. De plus, nous exploitons les connexions résiduelles de l'entrée aux couches masquées et entre les couches masquées. Nous montrons que la combinaison de contextes à court et à long terme, ainsi que l'inclusion de telles connexions résiduelles, améliore considérablement les performances de la déréverbération sur quatre métriques de qualité et d'intelligibilité de la parole différentes (PESQ, SRMR, STOI et POLQA) et réduit également la quantité de réverbération perçue selon des tests subjectifs.

Montage expérimental

Afin d'évaluer les avantages de l'architecture proposée pour la déréverbération de la parole, nous avons effectué une série d'expériences avec le modèle proposé et deux autres modèles comme référence: le modèle prenant en compte les informations T60 proposé dans [Wu2016] et un modèle similaire ne contenant pas d'informations T60 qui utilise un chevauchement fixe de 16 ms et un contexte fixe de 11 trames (5 trames antérieures et 5 futures). Pour le modèle tenant en compte les informations T60, nous avons extrait les valeurs T60 directement à partir des RIR à l'aide d'une méthode similaire à celle utilisée pour l'ensemble de données ACE Challenge [AceChallenge]. Les modèles ont été testés avec un seul jeu de données de locuteur simple (jeu de données IEEE prononcé par un homme) et un jeu de données de plusieurs locuteurs (à l'aide du corpus TIMIT).

Les énonciations réverbérantes ont été générées en convoluant de manière aléatoire des sous-ensembles d'énoncés dans l'ensemble d'apprentissage avec 740 RIR générés à l'aide d'une implémentation rapide du procédé source-image, le T60 variant de 0.2s à 2.0s par incréments de 0.05s. Vingt RIR différents (avec une géométrie de pièce, un positionnement source-microphone et des caractéristiques d'absorption différents) ont été générés pour chaque valeur T60. Un sous-ensemble aléatoire de 5% des fichiers a été sélectionné en tant qu'ensemble de validation et utilisé pour la sélection du modèle. L'ensemble de tests a été généré de manière similaire, mais en utilisant un ensemble différent de 740 RIR simulés et des énoncés différents de ceux utilisés pour l'ensemble d'apprentissage (dans le cas de TIMIT, l'ensemble de tests par défaut a été utilisé). En outre, afin d'explorer les performances des modèles proposés et des modèles de référence dans un cadre réaliste, nous avons testé les mêmes modèles de locuteur unique décrits ci-dessus avec des phrases convoluées avec de vrais RIR de l'ensemble de données ACE Challenge [ACEChallenge].

Nous avons comparé les performances des modèles en utilisant quatre métriques objectives différentes (PESQ, POLQA, STOI et SRMR) et également à l'aide d'un test d'écoute subjectif à petite échelle utilisant le protocole MUSHRAR [MUSHRAR], afin d'évaluer l'efficacité des différentes méthodes quant à la réduction de la quantité de réverbération perçue. Les résultats de quatre expériences principales ont été rapportés: 1. l'effet de la taille du contexte dans le modèle proposé, 2. une comparaison entre l'architecture proposée et les valeurs de référence sur des conditions appariées (un locuteur unique et des RIR simulés pour la formation et les tests), 3. la performance de toutes les architectures sur de nouveaux locuteurs (entraînement et tests effectués sur un jeu de données avec de multiples locuteurs) 4. la performances de toutes les architectures dans des conditions de test plus réalistes (modèles formés à l'aide de RIR simulés mais testés avec le jeu de données RIR réel).

Conclusions

Les résultats ont montré que cette architecture surpasse les architectures de pointe et qu'elle généralise à différentes géométries de salles et T60 (y compris les RIR réels), ainsi qu'à différents locuteurs. Notre architecture extrait des caractéristiques à la fois dans un contexte local (c'est-à-dire quelques trames du passé / futur de de la frame estimée) et dans un contexte à long terme. Dans de futurs travaux, nous avons l'intention d'explorer des fonctions de coût améliorées (par exemple, incorporer des matrices de plus faible densité (*sparsity*) dans les sorties) ainsi que d'appliquer l'architecture à des signaux déformés avec à la fois du bruit additif et de la réverbération. Nous avons également l'intention de proposer une extension à plusieurs canaux de l'architecture. Enfin, nous avons l'intention d'explorer un certain nombre de solutions au problème de la reconstruction du signal à l'aide de la phase réverbérante.

0.2.3 Chapitre 4: Interprétation et optimisation des modèles DNN d'amélioration de la parole

Ce chapitre présente quelques idées pour rendre les modèles d'amélioration de la parole plus interprétables et pour mieux comprendre la représentation interne du modèle présenté au chapitre 3. Ces idées nous ont amené à proposer une nouvelle méthode d'élagage de modèle permettant de le compresser significativement sans perte de performance.

0.2.4 Conception de modèles plus interpretables grâce à l'utilisation de connections de saut

Dans la première partie, nous avons exploré l'hypothèse selon laquelle l'utilisation de connexions de saut (*skip connections*) pour construire la sortie du modèle rend celui-ci plus interprétable, car les sorties de chaque couche se trouvent maintenant dans le même domaine de représentation. Nos expériences ont été réalisées avec trois modèles différents: l'un basé sur les réseaux résiduels, un autre sur les réseaux d'autoroute (*highway networks*) puis un modèle ne comportant qu'une simple modification sur les réseaux résiduels, inspirés des réseaux d'autoroute (*highway networks*), qui effectue le masquage. Nous établissons des parallèles entre ces architectures et les approches classiques d'amélioration de la parole basées sur le traitement de signal numérique (*digital signal processing* en anglais, ou DSP): les modèles résiduels fonctionnent de manière similaire aux méthodes de soustraction spectrale, les réseaux de masquage de manière similaires au masquage spectral et les méthodes d'autoroute (*highway methods*) comme combinaison des deux.

En utilisant la couche de sortie pour transformer les connexions de saut (*skip connections*) de chaque couche séparément, nous sommes en mesure de comprendre la contribution de chaque partie du modèle à la sortie finale. Deux versions de chaque modèle ont été testées : l'une avec une seule connexion de saut (*skip connection*) de l'entrée à la sortie et trois couches de type *gated recurrent units* (GRU) entre les deux, et une autre où chaque couche de GRU est entourée d'une connexion de saut (*skip connection*). Les modèles ont été testés sur une tâche de débruitage et de déréverbération. Ils ont été bâtis à partir d'un jeu de données de locuteur simple corrompu soit par le bruit du corpus de bruit DEMAND, soit par les RIR simulés des expériences du chapitre 3. Tous les modèles utilisaient une représentation de type *short-time Fourier transform* (STFT) pour l'entrée et la sortie, utilisant des trames de 32ms avec un chevauchement de 50%.

Globalement, le modèle résiduel était le plus facile à interpréter. Nous avons observé que plus une couche de réseau est profonde, plus elle est sélective à la fréquence. La première couche isole les composantes les plus fortes des basses fréquences, la seconde les fréquences moyennes et la dernière les fréquences les plus élevées, ainsi que le bruit diffus. Le modèle d'autoroute (*highway model*) était plus difficile à interpréter, mais en observant sa fonction de synchronisation (*gating function*), nous avons pu comprendre qu'il fonctionnait comme une mesure de la confiance des résultats du modèle par rapport aux intrants. Enfin, le modèle de masquage semble fonctionner de manière similaire au modèle résiduel, la résolution de fréquence étant améliorée de couche en couche. Nous avons également montré que par rapport aux modèles avec peu ou pas de connexions de saut (*skip connections*), ces dernières n'affectent pas la performance du modèle (observée via des métriques objectives de qualité de la parole et d'intelligibilité) mais qu'elle rendent son fonctionnement interne plus lisible.

Élagage des modèles LSTM et GRU en eliminant des neurones "bloquées" et hautement corrélées

Nous avons ensuite proposé des outils de visualisation interactifs pour quatre modes de visualisation / ablation, afin de nous aider à mieux comprendre les modèles d'amélioration de la parole et leurs représentations internes. Le premier mode est une simple visualisation des activations cachées dans le modèle, tandis que les trois autres sont des expériences d'ablation dans lesquelles une seule caractéristique ou une seule couche du modèle est perturbée de l'une des trois manières suivantes: en injectant du bruit dans toutes les activations d'une couche donnée, en fixant une entité à zéro et en fixant une entité à sa médiane.

En utilisant ces visualisations, ainsi que des matrices de corrélation pour observer si les neurones d'une même couche génèrent des sorties similaires à d'autres, nous avons observé que de nombreuses neurones étaient saturées la plupart du temps (nous appelons ces neurones "bloqués"), et semblaient donc redondantes en regard de la tâche. Nous avons ensuite appliqué plusieurs méthodes de régularisation et / ou d'élagage de modèle, telles que le décrochage (*dropout*), l'élagage pondéral (*weight pruning*) et la réduction de la taille du modèle, mais ces méthodes n'ont pas permis de réduire significativement le nombre de neurones bloqués. C'est pourquoi, dans ce chapitre, nous avons également proposé une nouvelle méthode d'élagage de neurones dans les DNN, complémentaire à l'élagage de neurones basé sur la magnitude, mais destinée aux modèles utilisant des fonctions saturant à des valeurs non nulles (tels que les modèles utilisant des couches LSTM et GRU).

La méthode consiste à éliminer les neurones qui sont toujours bloqués et à ajuster les biais de toutes les couches connectées à ce neurone pour compenser la perte de cette entrée. Nous utilisons une heuristique simple pour détecter quels neurones peuvent être élagués une fois le modèle formé. Nous montrons que la méthode proposée est efficace pour élaguer de tels neurones sans modification significative des performances du modèle, ce qui nous permet de compresser une couche de GRU à

48% de sa taille initiale dans le modèle évalué. La méthode proposée ne nécessite pas non plus de matériel particulier pour tirer parti de la compression, contrairement aux méthodes basées sur le clairsemage (*sparsifying*) de matrices de poids (*weight matrix*), et elle est post-hoc, ce qui la rend utile même pour les modèles déjà entraînés.

Conclusions

Le travail présenté dans ce chapitre montre les résultats initiaux de notre enquête sur le rôle des connexions de saut (*skip connections*) dans les modèles d'amélioration de la parole. Nos expériences préliminaires montrent que, même si elles n'ont pas d'impact significatif sur les performances des modèles, de telles connexions peuvent aider à rendre les modèles plus interprétables, puisqu'elles permettent d'identifier la contribution de chaque couche à la tâche. Dans nos prochains travaux, nous avons l'intention d'enquêter sur des modèles plus complexes, tels que des modèles basés sur l'architecture UNet, ainsi que des modèles utilisant une fenêtre de contexte temporel en entrée plutôt qu'un seul cadre, car ceux-ci sont davantage utilisés dans les modèles de pointe.

Nous avons également proposé une méthode d'élagage post-hoc pour les modèles récurrents basée sur les connaissances acquises à partir d'observations de modèles d'activation dans de telles couches. À elle seule, notre méthode réalise plus de 50% de compression d'une couche récurrente dans le modèle utilisé lors de notre expérimentation et une compression totale d'environ 3% de manière totalement post hoc. Cela permet à cette méthode d'être utilisée dans tous les modèles de pré-entraînement utilisant des unités GRU et LSTM afin de réduire les besoins en calcul et en stockage. De plus, la méthode proposée est complémentaire aux méthodes basées sur les fonctions d'importance de magnitude pour les neurones et peut être facilement combinée avec celles-ci.

Dans de futurs travaux, nous avons l'intention d'évaluer les méthodes de visualisation proposées ici avec d'autres modèles liés à la parole, tels que des modèles de reconnaissance de la parole, d'identification / vérification du locuteur et de synthèse de la parole. Nous aimerions également évaluer l'utilité de la stratégie d'élagage proposée dans ce chapitre pour d'autres modèles utilisant des RNN, comme pour les tâches liées à la parole mentionnées plus haut, ainsi que pour le traitement du langage naturel.

0.2.5 Chapitre 5: Vers une amélioration de la parole basée sur les DNN et informé par la qualité de la parole

Analyse de métriques de la qualité de la parole pour des méthodes d'amélioration de la parole basée sur DNN

La qualité de la parole est multidimensionnelle et prend en compte plusieurs aspects du signal. Le mode d'évaluation de la qualité d'un signal dépendra généralement du type de distorsion et de l'amélioration attendue. De telles mesures reflètent généralement également l'effort d'écoute d'un signal de parole, tandis que l'intelligibilité ne concerne que la question de savoir si un mot ou une phrase est intelligible ou non. Les métriques de qualité objective sont souvent développées en tenant compte de certains types de distorsion (par exemple, la métrique interne de robotisation de la norme ITU-T P.563); Cependant, l'amélioration de la parole à l'aide de méthodes basées sur le DNN, en particulier les méthodes qui effectuent une estimation spectrale du signal au lieu de masquer, peut présenter des types d'artefacts inattendus qui n'ont pas été envisagés lors du développement de telles mesures. De plus, de nombreuses métriques ne sont pas sensibles à la phase, mais les auditeurs humains considèrent la distorsion de phase comme un élément de la qualité de la parole.

La majorité de l'amélioration de la parole basée sur le DNN est entraînée par apprentissage supervisé, nécessitant à la fois un signal de référence propre comme cible et un signal déformé comme entrée du modèle. Bien qu'il soit relativement facile de simuler des distorsions en ajoutant du bruit enregistré et des signaux de convolution aux réponses impulsionnelles de la pièce pour les rendre réverbérants, cela ne reflète pas nécessairement la manière dont les signaux naturels sont créés. Avoir accès à des modèles précis et non intrusifs de la qualité et de l'intelligibilité de la parole pourrait permettre d'effectuer de l'apprentissage semi-supervisée (nécessitant des signaux de référence pour un sous-ensemble du jeu de données utilisé pour la formation) ou non supervisée (ne nécessitant aucun signal de référence). Il existe des travaux dans la littérature qui utilisent l'apprentissage semi-supervisé pour améliorer / séparer la parole en utilisant une factorisation matricielle non négative, où le modèle de bruit est appris de manière non supervisée, mais nous n'avons trouvé aucune recherche sur l'apprentissage semi-supervisé ou non supervisé appliqué à l'amélioration de la parole basée sur le DNN.

Dans le travail décrit dans ce chapitre, nous avons examiné les résultats de tests d'écoute en ligne pour évaluer la qualité de la parole de plusieurs systèmes récemment proposés de débruitage et de déréverbération basés sur les DNN, et avons comparé les évaluations subjectives de ces expériences avec différents indicateurs objectifs de qualité et d'intelligibilité de la parole. Nous examinons si de tels indicateurs sont adaptés à l'évaluation de la qualité de la parole pour les modèles DNN d'amélioration de la parole, et proposons des indicateurs combinés alternatifs basés sur des indicateurs internes non intrusifs du P.563. Compte tenu des connaissances tirées de nos tests d'écoute, nous avons conçu certaines expériences pour inclure certains des indicateurs de qualité alternatifs que nous avons trouvés en tant que fonction de coût pour la formation de modèles DNN d'amélioration de la parole.

Montage expérimental

Nous avons formé trois modèles différents pour effectuer le débruitage et la déréverbération. Nous avons utilisé le corpus TIMIT comme source pour les stimuli de parole. L'ensemble d'apprentissage par défaut (sans les énoncés "SA", ces enregistrements ayant été enregistrés par tous les locuteurs) a été utilisé pour générer les ensembles d'apprentissage et de validation, et l'ensemble d'essai (avec les énoncés "SA" supprimés) a été utilisé pour générer l'ensemble d'essai. Les énonciations réverbérantes ont été générées en convoluant de manière aléatoire des sous-ensembles d'énoncés dans l'ensemble d'apprentissage avec 740 RIR générés à l'aide d'une implémentation rapide du procédé source-image, le T60 variant de 0.2s à 2.0s par incréments de 0.05 s. Vingt RIR différents (avec une géométrie de pièce, un positionnement source-microphone et des caractéristiques d'absorption différents) ont été générés pour chaque valeur T60.

Pour le jeu de données de débruitage, nous avons mélangé les phrases d'apprentissage et de validation aux bruits du jeu de données DEMAND à des rapports signal sur bruit (*signal to noise ratio* en anglais, ou SNR) de 12, 6, 3, 0, -3 et -6 dB. Les phrases de test ont été mélangées à deux bruits (brouhaha et bruit d'usine) du jeu de données NOISEX, à des SNR de 13, 7, 4, 1, -2 et -5 dB. Pour chaque phrase, un segment de bruit aléatoire de même longueur que la phrase fut sélectionné.

Tests d'écoute

Nous avons ensuite effectué deux tests d'écoute en ligne différents, l'un pour la déréverbération et l'autre pour la réduction du bruit. Les deux tests étaient des tests de type MUSHRA pour évaluer la qualité de la parole en la comparant à une référence propre. Lors des tests, les participants ont été présentés avec les résultats de tous les modèles (déréverbération ou réduction du bruit) pour un seul stimuli de parole simultanément, ainsi qu'une référence cachée, le signal corrompu amélioré par les modèles et une ancre, et étaient invités à évaluer la qualité du signal à l'aide de curseurs dont les positions quantifiées en nombres entiers étaient compris entre 0 et 100. Dans le cas de conditions de réduction du bruit, l'ancre était le même stimuli corrompu avec le même type de bruit mais avec un SNR inférieur de 5 dB. Dans le cas de conditions de déréverbération, l'ancre était un signal convolué avec un RIR d'un T60 de 2s. Dans les deux cas, la référence était le signal anéchoïque net. Un total de 10 stimuli ont été utilisés pour chaque condition, chaque stimulus étant constitué de deux phrases concaténées du jeu de données TIMIT prononcées par différents locuteurs avec un intervalle de 2 secondes entre les phrases. La durée minimale totale des stimuli était de 8 secondes.

Nous avons calculé des corrélations pour plusieurs indicateurs objectifs lors des tests d'écoute de débruitage et de déréverbération. Les résultats ont montré que des indicateurs plus simples basées sur le signal, telles que SegSNR et FWSegSNR, offrent des performances similaires à des indicateurs plus complexes telles que PESQ et POLQA. Les performances des indicateurs relatifs à la parole déréverbérée sont légèrement inférieurs que celles relatives au débruitage, ce qui peut s'expliquer par la petite plage couverte par la parole traitée dans l'axe d'évaluation, puisque tous les modèles d'amélioration ont des performances relativement modestes par rapport à la tâche de débruitage.

Recherche de métriques non intrusives

Afin de découvrir des caractéristiques pouvant être utiles pour estimer la qualité de la parole pour une amélioration basée sur les DNN de manière non intrusive, nous avons examiné les corrélations des caractéristiques internes de l'indicateur P.563 avec les scores d'opinion moyens obtenus lors de nos tests d'écoute. Nous avons montré que les indicateurs de l'analyse LPC et des modules de paramètres spécifiques à la distorsion montraient les corrélations les plus élevées. Le kurtosis et l'asymétrie des paramètres LPC étaient fortement corrélés à la qualité de la parole pour le débruitage (0.870 et 0.862) et la déréverbération (0.837 et 0.846). Le paramètre de bruit de fond local et les caractéristiques liées au bruit multiplicatif étaient également fortement corrélés à la qualité de la parole.

Nous avons proposé un indicateur combiné en formant des modèles de régression linéaire distincts pour les tâches de débruitage et de déréverbération, et nous avons évalué les performances de ce modèle par validation croisée, en laissant les données d'un modèle hors de l'ensemble d'entraînement. L'indicateur proposé a des corrélations en ligne avec les indicateurs intrusifs et surpasse les autres indicateurs non intrusifs telles que SRMR et l'indicateur P.563 d'origine.

Amélioration de la parole basée sur DNN et informé par la qualité de la parole

Nous avons ensuite entraîné quelques modèles basés sur l'architecture proposée au chapitre 3, en utilisant les connaissances tirées de ces tests d'écoute en incluant un terme de qualité à la fonction de coût via un multiplicateur lagrangien. En raison des difficultés numériques liées à la rétroprojection par estimation de paramètres de vocodeur, nous avons expérimenté avec des modèles d'entraînement afin de générer directement des représentations de vocodeur au niveau de leurs sorties, en faisant correspondre le domaine de de certaines des caractéristiques rencontrées hautement corrélées à la qualité de la parole des réseaux DNN. L'un des modèles utilisait la représentation des coefficients de réflexion pour le codage par prédiction linéaire et l'autre utilisait un vocodeur plus récent, à savoir le vocodeur WORLD, qui utilise des coefficients *mel-generalized cepstrum* (MGC) au lieu de LPC pour représenter l'enveloppe spectrale. Nous avons formé les deux modèles en utilisant plusieurs constantes différentes pour le multiplicateur lagrangien (qui pondère la pertinence du terme sensible à la qualité par rapport au terme MSE d'origine). Les deux modèles ont été formés avec le jeu de données de débruitage de locuteur unique décrit dans les chapitres précédents.

D'un point de vue d'indicateur objectif, le modèle proposé avec sortie de vocodeur de type WORLD a entraîné une amélioration du PESQ, tandis que le modèle basé sur le LPC avec la perte proposée a sous-performé le modèle de base formé avec MSE uniquement dans toute la plage d'hyperparamètres que nous avons testée. Un test d'écoute en ligne à petite échelle a montré que, malgré l'amélioration du PESQ, plus de 50% des participants ont estimé que les phrases améliorées par le modèle de référence étaient plus naturelles que celles utilisant la fonction de coût proposée. Nous comprenons que l'étude a une portée limitée car la sortie du modèle dans l'espace vocodeur ne correspond pas aux expériences que nous avons effectuées pour les indicateurs de qualité. En outre, la prévision de la hauteur tonale dans des conditions bruyantes est un problème épineux. De plus, aucune recherche d'hyperparamètres autre que le multiplicateur de Lagrange n'a été effectuée, même si nous avons complètement modifié le domaine de sortie du modèle.

Conclusions

Le travail présenté dans ce chapitre portait sur les performances de plusieurs indicateurs objectifs de qualité de parole pour l'amélioration de la parole basée sur DNN et ont présenté les premières étapes vers un indicateur non intrusif pour de tels systèmes. L'indicateur proposé, basé sur des indicateurs internes de la norme P.563, fonctionne en parallèle avec des indicateurs intrusifs pour le débruitage de la parole, mais pas aussi bien pour la déréverbération de la parole. Bien que l'indicateur proposé soit un travail en cours, les informations fournies par les indicateurs présentant des corrélations élevées avec la qualité de la parole peuvent être pris en compte pour concevoir de meilleures fonctions de coût ou des contraintes pour la formation supervisée de modèles d'amélioration de la parole.

De manière connexe, nous avons également présenté des expériences de validation de principe sur un modèle DNN d'amélioration de la parole soucieux de la qualité. En raison des difficultés numériques liées à la rétroprojection par estimation de paramètres de vocodeur, nous avons expérimenté avec des modèles d'entraînement afin de générer directement des représentations de vocodeur au niveau de leurs sorties, en faisant correspondre le domaine de certaines des caractéristiques rencontrées hautement corrélées à la qualité de la parole qui fut améliorée par DNN. Bien que notre approche avec le vocodeur WORLD augmente les scores PESQ, les participants à un test à petite échelle ont montré une préférence pour les signaux qui n'utilisaient pas cette approche. La portée de ces expériences est limitée car nous n'avions pas testé initialement les indicateurs de qualité avec des vocodeurs.

En ce qui concerne les travaux futurs, nous envisageons un test d'écoute similaire pour l'intelligibilité de la parole, qui pourrait déboucher sur une étude similaire qui examinerait les aspects du signal liés à l'intelligibilité de la parole améliorée par DNN. Les études futures pourraient également prendre en compte les modèles récemment proposés pour l'amélioration basée sur DNN qui traitent les signaux dans le domaine temporel, tels que [...], car ils généreront probablement un ensemble d'artefacts différent de celui du traitement du spectre de magnitude. Pour les modèles sensibles à la qualité, les prochaines étapes consisteraient à améliorer les performances du modèle en matière de prévision de la hauteur tonale, probablement avec des branches distinctes pour les coefficients MGC, log-F0 et d'apériodicité. D'autres représentations / mesures de qualité évitant les problèmes numériques / computationnels rencontrés avec le LP kurtosis pourraient également être explorées. Enfin, une version étendue des tests d'écoute utilisant des modèles avec des sorties dans des domaines de vocodeur pourrait être utile pour identifier d'autres indicateurs potentiellement utilisables comme fonctions de perte complémentaires.

Chapter 1

Introduction

In the last few years, speech technologies have become ubiquitous. Thanks to advances in automatic speech recognition (ASR), speech has become a way of interacting with devices and applications. Voice-driven applications went beyond simple transcription and now allow interaction with our TVs, phones, and cars. In these applications, devices were moved from more controlled acoustic environments, like quiet rooms with close-talking microphones, to complex acoustic environments, where far-field microphones are used in the presence of different kinds of noise and reverberation, and interfering speakers. The distortions added by this kind of acoustic environment lead to severe performance degradation in ASR performance [1], as well as a reduction in speech intelligibility and quality in speech communication systems, especially for the hearing impaired [2].

In most real-world applications, the clean speech signal is distorted by a combination of different categories of distortions before being captured by a device. In enclosed environments, for example, distortions are usually modeled as a mix of additive background noise and reverberation. To deal with the distortions in such environments, several types of speech enhancement systems have been proposed, ranging from single-channel systems based on simple spectral subtraction [3] to multistage systems that leverage signals from multiple microphones [4].

Recently, deep neural networks (DNNs) [5] have been successfully employed in a broad range of applications, having achieved state-of-the-art results in tasks such as acoustic modeling for ASR [6] [7] and image classification [8]. The reason for such advancements is two-fold. First, the evolution of computing systems and hardware accelerators such as graphics processing units (GPUs) has led to a significant increase in compute power and storage capabilities, which are two important requirements for successfully training a neural network with several layers. Second, advances in optimization algorithms and network architectures such as recurrent and convolutional neural networks have improved the ability of such systems to learn spatial and temporal dependencies in data.

DNNs already play an important role in current large-scale ASR systems in production, being used especially in acoustic models [7], as well as a feature preprocessing stage to improve robustness to reverberation [9]. However, their application to speech enhancement problems such as denoising [10], dereverberation [11], and source separation [12] is more recent work and therefore in more preliminary stages, although some promising results have been achieved.

As opposed to traditional speech enhancement algorithms based on minimum mean-squared error estimators, a DNN-based speech enhancement system does not require an analytical solution to the estimation problem nor is limited to simple statistical models (such as the i.i.d. Gaussian assumption in [13] and derivative works). DNN-based speech enhancement systems only require a differentiable cost function and are trained using stochastic optimization algorithms, usually on vast amounts of data. This allows, for example, the design of a training dataset containing multiple types of distortion scenarios. While training of such networks is time-consuming and requires a large amount of computational power, this does not apply to using trained networks in practice.

Despite the flexibility in DNNs, most existing solutions rely on a simple mean-squared error cost function, which is simple to optimize but does not necessarily correlate well with speech quality and intelligibility.

1.1 Single channel speech denoising and dereverberation

Reverberation and noise have complementary effects on a clean speech signal. A simple signal model for a speech signal distorted by a combination of noise and reverberation could be given by the following equation:

$$y(n) = h(n) * x(n) + \nu(n) \tag{1.1}$$

where reverberation is represented by the convolution of the room impulse response (RIR) $h(n)$ to the clean signal $x(n)$ and $\nu(n)$ represents additive noise.

Reverberation causes two significant perceptual effects to a clean speech signal: early reflections (i.e., reflections that happen with a delay of up to 30-50 ms) cause spectral coloration of the signal, whereas late reverberation causes temporal smearing [14]. Spectral coloration, per se, is considered benefitial as it reinforces the speech signal and makes it sound more natural, while the late reflections deteriorate the signal as they are less correlated. Its effects are especially significant for hearing-impaired listeners, which experience reduced speech intelligibility even under relatively low reverberation times [15, 16]. Additive noise distortions, on the other hand, affect speech intelligibility differently: weak consonants suffer more masking than higher intensity vowels, and this effect is not dependent on the energy of preceding segments (which is the case for reverberation).

1.2 Challenges

Existing approaches for speech enhancement have several limitations. Speech and noise models are often simplified distributions of the amplitude coefficients (Gaussian, super-Gaussian, etc.). This is necessary in order to achieve analytic solutions for the estimation problem.

Most current speech enhancement algorithms have a simplified objective function that does not correspond to perceptual characteristics. One common distortion metric used is the mean squared error of spectral coefficients, which disregards auditory masking and distortions due to phase. This leads to adverse effects, such as the generation of perceptual artifacts.

1.3 Thesis contributions

This work presented in this thesis explores DNN-based speech enhancement in three different and complementary ways. First, we propose a model to perform speech dereverberation by estimating its spectral magnitude from the reverberant counterpart. Our models are capable of extracting features that take into account both short and long-term dependencies in the signal through a convolutional encoder and a recurrent neural network for extracting long-term information. Our model outperforms a recently proposed model that uses different context information depending on the reverberation time, without requiring any sort of additional input, yielding improvements of up to 0.4 on PESQ, 0.3 on STOI, and 1.0 on POLQA relative to reverberant speech. We also show our model is able to generalize to real room impulse responses even when only trained with simulated room impulse responses, different speakers, and high reverberation times. Lastly, listening tests show the proposed method outperforming benchmark models in reduction of perceived reverberation.

We also study the role of residual and highway connections in deep neural networks for speech enhancement, and verify whether they function in the same way that their digital signal processing counterparts do. We visualize the outputs of such connections, projected back to the spectral domain, in models trained for speech denoising, and show that while skip connections do not necessarily improve performance with regards to the number of parameters, they make speech enhancement models more interpretable. We also discover, through visualization of the hidden units of the context-aware model we proposed, that many of the neurons are in a state that we call "stuck" (i.e. their outputs is a constant value different from zero). We propose a method to prune those neurons away from the model without having an impact in performance, and compare this method to other methods in the literature. The proposed method can be applied post hoc to any pretrained models that use sigmoid or hyperbolic tangent activations. It also leads to dense models, therefore not requiring special software/hardware to take advantage of the compressed model.

Finally, in order to investigate how useful current objective speech quality metrics are for DNN-based speech enhancement, we performed online listening tests using the outputs of three different DNN-based speech enhancement models for both denoising and dereverberation. When assessing the predictive power of several objective metrics, we found that existing non-intrusive methods fail at monitoring signal quality. To overcome this limitation, we propose a new metric based on a combination of a handful of relevant acoustic features. Results in line with those obtained with intrusive measures are then attained. In a leave-one-model-out test, the proposed non-intrusive metric is also shown to outperform two non-intrusive benchmarks for all three DNN enhancement methods, showing the proposed method is capable of generalizing to unseen models. We then take the first steps in incorporating such knowledge into the training of DNN-based speech enhancement models by designing a quality-aware cost function. While our approach increases PESQ scores for a model with a vocoder output, participants of a small-scale preference test considered it less natural than the same model trained with a conventional MSE loss function, which signals more work is needed in the development of quality-aware models.

1.4 Publications derived from the thesis

- João Felipe Santos and Tiago H. Falk. Speech dereverberation with context-aware recurrent neural networks. IEEE Transactions on Audio, Speech, and Language Processing, July 2018.
- João Felipe Santos and Tiago H Falk. Investigating the effect of residual and highway connections in speech enhancement models. In NeurIPS Workshop on Interpretability and Robustness in Audio, Speech, and Language, December 2018.
- João Felipe Santos and Tiago H. Falk. Pruning LSTM and GRU-based in speech processing models through eliminating "stuck" features. Submitted to IEEE Transactions on Audio, Speech, and Language Processing.
- João Felipe Santos and Tiago H. Falk. Towards the development of a non-intrusive objective quality measure for DNN-enhanced speech. 11th International Conference on Quality of Multimedia Experience (QoMEX), June 2019.
- João Felipe Santos and Tiago H. Falk. Towards quality-aware DNN-based speech enhancement. Submitted to IEEE Signal Processing Letters.

1.4.1 Related work

Source separation

- Stylianos Ioannis Mimilakis, Konstantinos Drossos, João Felipe Santos, Gerald Schuller, Tuomas Virtanen, and Yoshua Bengio. Monaural singing voice separation with skip-filtering connections and recurrent inference of time-frequency mask. In International Conference on Acoustics, Speech, and Signal Processing (ICASSP). 2018.

Speech synthesis

- Kyle Kastner, João Felipe Santos, Yoshua Bengio, and Aaron Courville. Representation Mixing for TTS Synthesis. In International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2019.

- Jose Sotelo, Soroush Mehri, Kundan Kumar, João Felipe Santos, Kyle Kastner, Aaron Courville, and Yoshua Bengio. Char2Wav: End-to-end speech synthesis. In International Conference on Learning Representations (Workshop Track). April 2017.

Deep neural networks and generative models

- Chiheb Trabelsi, Olexa Bilaniuk, Dmitriy Serdyuk, Sandeep Subramanian, João Felipe Santos, Soroush Mehri, Negar Rostamzadeh, Yoshua Bengio, and Christopher J. Pal. Deep complex networks. In International Conference on Learning Representations (ICLR). April 2018.
- Bob L. Sturm, João Felipe Santos, Oded Ben-Tal, and Iryna Korshunova. Music transcription modelling and composition using deep learning. In 1st Conference on Computer Simulation of Musical Creativity. June 2016.
- Bob L. Sturm, João Felipe Santos, and Iryna Korshunova. Folk music style modelling by recurrent neural networks with long short term memory units. In International Society for Music Information Retrieval (ISMIR) conference. October 2015. (late-breaking demo).

Objective metrics for speech signals

- Anderson Avila, Zahid Akhtar Momin, João Felipe Santos, Douglas O'Shaghnessy, and Tiago H. Falk. Feature pooling of modulation spectrum features for improved speech emotion recognition in the wild. IEEE Transactions on Affective Computing, July 2018.
- Sebastian Braun, João Felipe Santos, Emanuel Habets, and Tiago H. Falk. Dual-channel modulation energy metric for direct-to-reverberation ratio estimation. In International Conference on Acoustics, Speech, and Signal Processing (ICASSP). April 2018.
- Mohammed Senoussaoui, João Felipe Santos, and Tiago H. Falk. Speech temporal dynamics fusion approaches for noise-robust reverberation time estimation. In International Conference on Acoustics, Speech, and Signal Processing (ICASSP). March 2017.
- João Felipe Santos and Tiago H. Falk. Blind room acoustics characterization using recurrent neural networks and modulation spectrum dynamics. In AES 60th International Conference. February 2016.

- Benjamin Cauchi, João Felipe Santos, Kai Siedenburg, Tiago H. Falk, Patrick Naylor, Simon Doclo, and Stefan Goetze. Predicting the quality of processed speech by combining modulation based features and model trees. In ITG Conference on Speech Communication. 2016.

- João Felipe Santos, Rachel Bouserhal, Jérémie Voix, and Tiago H. Falk. Objective quality estimation of in-ear microphone speech. In 5th ISCA/DEGA Workshop on Perceptual Quality of Systems. September 2016.

- Mohammed Senoussaoui, João Felipe Santos, and Tiago H. Falk. SRMR variants for improved blind room acoustics characterization. In ACE Challenge Workshop. October 2015.

- Tiago H. Falk, Vijay Parsa, João Felipe Santos, Kathryn Arehart, Oldooz Hazrati, Rainer Huber, James Kates, and Susan Scollie. Objective quality and intelligibility prediction for users of assistive listening devices. IEEE Signal Processsing Magazine, March 2015.

- João Felipe Santos, Anderson Avila, Rachel Bouserhal, and Tiago H. Falk. Improving blind reverberation time estimation on a two-microphone portable device by using speech source distance information. In Speech in Noise Workshop. January 2015.

1.5 Thesis organization

This thesis is organized as follows. This chapter presented a brief motivation for DNN-based speech enhancement, its challenges, and summarized the contributions. Chapter 2 contains a quick summary of speech enhancement systems, deep neural networks, and the state of the art in DNN-based speech enhancement. Chapter 3 presents our context-aware speech dereverberation model. Chapter 4 includes our work on interpreting skip connections in speech enhancement models and on the neuron pruning method we proposed based on our observations of internal activations. In Chapter 5, we present the results of the listening tests we performed to better understand the performance of DNN-based speech enhancement models and how they correlated to objective quality metrics, propose a new metric based on the insights gained from such tests, and the first steps towards quality-aware DNN-based speech enhancement models. Finally, Chapter 6 contains some considerations about the works presented in this thesis and some ideas for future work.

Chapter 2

Background

2.1 Review of speech enhancement methods

2.1.1 Spectral amplitude estimation

Several different speech enhancement algorithms have been proposed to deal with noise and reverbertion. The most broadly used speech enhancement algorithm types are based on single-channel short-time spectral amplitude (STSA) enhancement methods, such as spectral subtraction [3] and estimators based on the minimization of a distance metric in the short-term spectral amplitude domain, such as mean-squared error [13] or, more recently, perceptually-inspired distances [17]. Even though multichannel devices are currently common and allow spatial filtering approaches such as beamforming, single-channel spectral enhancement is usually performed as a postfiltering stage [4].

An illustration of the basic flow diagram of such methods is shown in Figure 2.1: first, the corrupted speech signal y is converted to the short-time Fourier transform domain and separated into its magnitude and phase components ($|Y|$ and $\angle Y$), respectively). From the magnitude spectrum, the noise power at each frequency band is estimated (usually by identifying silent frames with a voice activity detector) and subtracted from $|Y|$ to yield an estimate of the clean magnitude spectrum $|\hat{X}|$, which is combined to the noisy phase in the synthesis stage to generate a time-domain signal x.

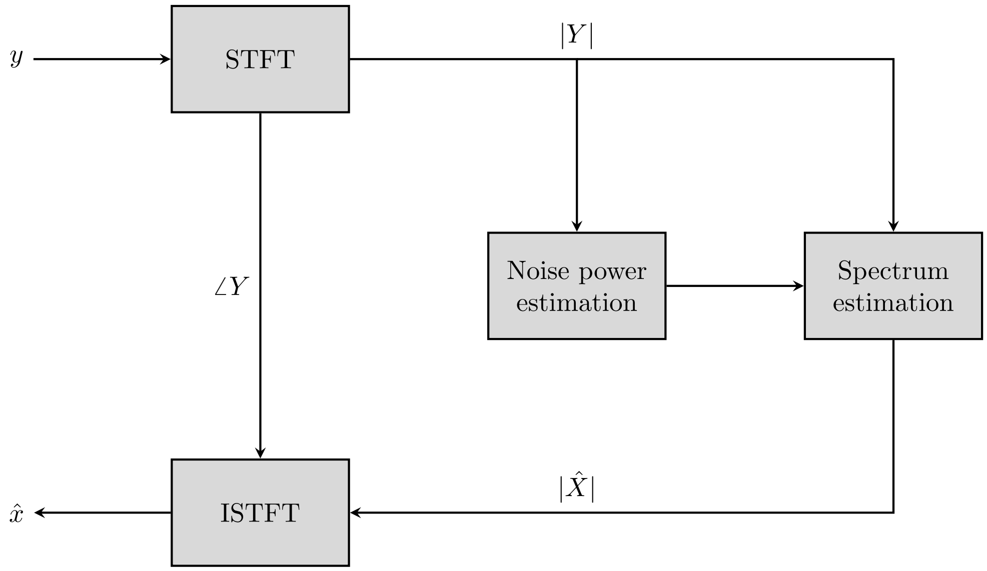


Figure 2.1 – Block diagram of a spectral estimation system

One common assumption in minimum mean-squared error (MMSE) based spectral estimation methods is that both the speech and noise signals have short-time spectral representations whose complex coefficients can be modeled as statistically independent Gaussian random variables [13] (for the speech model, Laplace, Gamma, or super-Gaussian [18, 19] distributions were also proposed in derivative works). Such an assumption implies the Fourier coefficients of speech are uncorrelated; however, it is known that speech signals have a strong structure both in time and frequency due to formant structure and harmonics [20]. While the assumption simplifies the computation of analytical solutions to the problem (as done in the MMSE approach), it only holds in an asymptotic sense and for very large frame sizes, which does not apply in practical cases as frame sizes in such systems are usually on the order of 20-40 ms).

2.1.2 Other approaches

The subspace approach is based on a decomposition of the vector space of a noisy signal into a signal-plus-noise and a noise subspace [21]. This approach has been combined with perceptual cues [22] to take into account auditory masking in the filtering performed on the subspace domain. Eigenvalue decomposition has also been applied to the multichannel problem in blind source separation (BSS) [23]. Other well-studied BSS approaches based on the assumption that speech and potential distortions can be factored into different subspaces include non-negative matrix factorization [24, 25, 26] and independent component analysis [27].

The authors of [28] and [29] use a linear prediction (LP) speech model for noise reduction and dereverberation. The LP residual obtained from a short-window (2 ms) LP analysis is weighted selectively based on the predicted *a priori* SNR for noise reduction. A similar approach is employed for dereverberation: since the LP residual of clean speech has higher kurtosis than that of reverberant speech, the LP residual is weighted differently, depending on the speech-to-reverberant ratio. In [30], the authors propose a two-stage approach to dereverberation by combining LP residual kurtosis minimization to spectral subtraction.

Modulation spectrum filtering is based on psychoacoustic and physiological evidences of the limited bandwidth of the temporal envelopes of the acoustic magnitude spectrum. Results show that low frequency modulations (between 1 and 16 Hz) carry the majority of the information contained in a speech signal. However, distortions outside of this modulation range also affect intelligibility. In [31], the authors explore this by performing bandpass filtering on the modulation spectrum of speech signals corrupted by additive noise, and a modulation domain spectral subtraction method is proposed in [32].

Several dereverberation techniques are based on blind deconvolution of the reverberant signal. Computation of the inverse filter in a blind way, however, is a hard problem, especially in the single channel case, which yields an under-constrained problem [33]. Channel equalization can be performed in a straightforward way assuming the input is white, and some approaches exploit this fact by estimating a source whitening filter for the input, and then applying multi-channel linear prediction (e.g., Linear-predictive Multi-input Equalization [34]). In a multichannel system, the multiple-input/output inverse theorem (MINT) can be applied to invert the effect of a room by finding multiple finite impulse response filters such that the sum of all filters convolved with the room response results in the inversion of the room response [35]. The method called harmonicity-based dereverberation (HERB) [36] uses the fact that speech signals have a harmonic structure to design an inverse filter. By using an adaptive harmonic filter, an estimate of the harmonic components of the direct signal is obtained, and a inverse filter is computed in the frequency domain as a weighted average of the ratio between the STFTs of the estimated harmonic components and the corresponding reverberant signal.

A recently-proposed single-channel approach for joint denoising and dereverberation is based on estimating spectral gains for the distorted signal from an estimate of reverberation power through an autoregressive model and a hidden Markov model for clean speech production [37]. The signal is first decomposed into mel-filterbank energies, which are then used to compute gains by a Bayesian filtering formulation of the problem based on the previously mentioned models. The method was shown to yield state-of-the-art performance, reducing both the effects of reverberation and noise, as well as improving speech quality and intelligibility.

Beamforming approaches exploit information from multiple microphones to reinforce the desired signal while blocking undesired interferences, ranging from simple delay-and-sum and filter-and-sum methods to adaptive beamformers such as the generalized sidelobe canceller (GSC) and minimum variance distortionless response beamformers [38] [39]. Recently, such systems have been combined with single-channel spectral enhancement schemes for joint reduction of noise and reverberation effects [4]. Another effective multichannel approach based based on linear prediction is the weighted prediction error algorithm [40], for which a recent formulation has been used to yield very high performance for a commercial ASR system by Google [41].

2.1.3 Environment, quality, and intelligility-aware approaches

Most speech enhancement systems in the literature only use minimal information about the environment. Noise reduction methods, for example, often require an estimate of the *a priori* signal-to-noise ratio (SNR) and an estimate of the noise spectrum. Some dereverberation systems, as mentioned in the previous session, perform inverse filtering, which requires an estimate of the room impulse response, while others only use an estimate of the reverberation time to compute an estimate of the room effects; the work in [42], for example, uses a pitch-based reverberation time measure to estimate and remove echo components via a frame-by-frame iterative spectral subtraction method.

Data-driven approaches based on the *a priori* and *a posteriori* SNR are proposed in [43] and [44]. In these approaches, parameters for individual subband weighting rules are stored in a lookup table, which is indexed by the *a priori* and *a posteriori* SNRs. In [43], the authors propose storing the gains for each quantized *a priori* and *a posteriori* pair directly in the lookup table. In [44], the authors store parameters for a weighting rule, which also takes into account the speech absence probability for the frame being processed. In both papers, the authors test different distortion measures (weighted- and log-Euclidean distance, and some of the perceptually-inspired measures proposed in [17]) as objective functions for finding the optimal parameters to be stored in the lookup table. In the same lines, a more recent work [45] proposes a set of model-based suppression rules based on a dataset of *a priori* and *a posteriori* SNRs and the desired gain for perfect suppression at each time-frequency bin. The authors tested different objective functions, which combined the mean-squared error (MSE) and log-MSE cost with Perceptual Evaluation of Speech Quality (PESQ) scores obtained from processing the data in the training set with each given set of parameters. Reported performance gains, however, are small: the authors show the proposed system improves PESQ scores in the range of 0.1 to 0.2 on a dataset with SNRs ranging from -10 to 50 dB.

SNR [46] and speech-to-reverberation (SRR) dependent methods [47] have also been proposed as part of channel selection strategies in cochlear implant speech coders to improve intelligibility under noise and reverberation. In the same area, an intelligibility-aware channel selection strategy based on the short-time objective intelligibility (STOI) metric was recently proposed [48], and is based on maximizing the short-time intelligibility estimate by finding the best combination of channels via an iterative (matching pursuit) approach.

2.2 Deep neural networks

Artificial neural networks (ANNs) are a category of statistical learning models inspired by how biological neural networks operate. The main elements of an ANN are layers of neuron units, which are non-linear mappings of a vector to another vector (not necessarily with the same length) and can be represented by the following relationship:

$$\mathbf{h} = g(W\mathbf{x} + \mathbf{b}) \tag{2.1}$$

where $\mathbf{h}$ is the output of the layer, g is a non-linear function such as a sigmoid, hyperbolic tangent, or rectifier ($\max(\cdot, 0)$), $\mathbf{x}$ is the input and W and $\mathbf{b}$ are the weight and bias parameters of the layer, respectively, and are the ones to be learned from a dataset[1]. Any layer not connected to the input or the output is called a hidden layer. A deep neural network [5, 49] is a composition of multiple layers of neuron units, which performs sequential non-linear projections of the input on the previous layer,

[1] Note that in this section, we use capital letters to represent matrices and lowercase letters to represent vectors

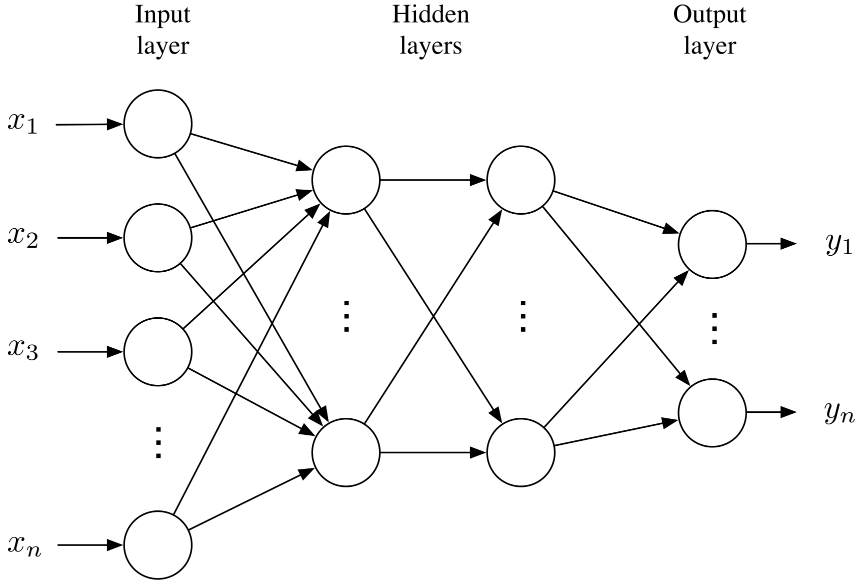


Figure 2.2 – Diagram of a deep neural network with two hidden layers

as illustrated in Figure 2.2. This composition of multiple layers allows a DNN to learn multiple layers of abstractions, which represent a given input from raw data, as opposed to other machine learning algorithms that often require feature engineering by the user.

The parameters of a deep neural network are learned in a supervised fashion by using an algorithm called backpropagation, which consists of iteratively adjusting the parameters of the network to minimize a cost function on its input and desired output. The most commonly used optimization procedure for training DNNs is stochastic gradient descent (SGD) or variants of this method; it is based on successively updating the parameters of a network according to the direction pointed by the gradient of the cost function with respect to a layer input. Unsupervised methods, which do not require target variables (e.g., labels in the case of a classification task, or a true value in case of a regression task), have also been applied; in this case, the objective of the DNN is to learn inherent characteristics of the input data, such as its probability distribution (as in restricted Boltzmann machines) or how to reconstruct corrupted data (as in the various types of autoencoders) [49]. This allows leveraging unlabeled data to train a model, which in the case of multimedia signals such as images and speech are available in larger quantities than labeled data. Unsupervised learning is often used as a way of initializing a DNN (i.e., help the model to learn a representation for the input data), and is followed by supervised learning on a smaller, labeled dataset (to map from the learned representation to the output variables of interest).

As mentioned earlier, DNNs have recently been employed in a broad range of tasks achieving impressive results, setting the state-of-the-art in tasks which had required large efforts in feature engineering, such as computer vision [8] and speech recognition [6] [7]. Most recent applications of DNNs to multimedia signals make use of variations of the DNN architecture shown here, namely recurrent and convolutional neural networks, which are better suited to working with sequential data and data composed by multiple arrays, respectively. These types of networks are described in the following sections.

2.2.1 Recurrent neural networks

Recurrent neural networks (RNNs) have been successfully applied to sequential data processing in many domains as, for example, acoustic models in ASR [7] and machine translation [50], as well as in speech enhancement [12], [51]. The main difference between a recurrent neural network and traditional feed-forward neural networks is that their hidden states are not only a function of the layer inputs, but also of the current hidden layer state. This recurrent connection allows the network to "remember" the past representation of its input and take it into account as the model sees new inputs, which has been shown as useful for modeling dynamics in sequential data such as speech signals. Given an input sequence, a recurrent neural network in its standard formulation computes a hidden vector sequence h and output vector sequence y by iterating over the sequence and computing the following [7]:

$$h_t = g(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \tag{2.2}$$

$$y_t = W_{hy}h_t + b_y \tag{2.3}$$

where the W_{nm} terms are the weight matrices of the connections between n and m (where n and m correspond either to the input x, hidden h, or output y), b are the biases and $\mathcal{H}$ is the activation function used for the hidden layer. Figure 2.3 shows a diagram of a recurrent neural network with the recurrent connection represented by the vertex marked as W_{hh} and the same network unfolded in time for better visualization of how the recurrence works.

This formulation, however, has some issues with training related to exploding and vanishing gradients during the training procedure [52], which causes training to take longer or even diverge, especially when capturing long-term dependencies is required by the problem. In practice, most

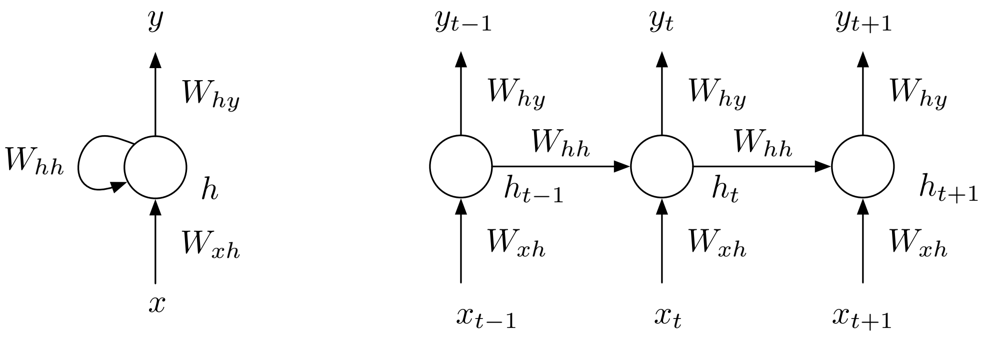


Figure 2.3 – Left: recurrent neural network; right: recurrent neural network unfolded in time.

current works with RNN use gated units such as long short-term memory (LSTM) [53] or gated recurrent units (GRUs) [54] to deal with the vanishing gradients issue. Such units use a gating mechanism to control the flow of information into and out of the unit, and are discussed in more detail in chapter 4.

2.2.2 Convolutional neural networks

Convolutional neural networks (CNNs) are a variant of feedforward neural networks inspired by how receptive fields in the visual cortex works [55]. The affine transformation in each layer prior to the non-linearity operation is replaced by a convolution operation (commonly a 2D convolution for inputs such as images or spectrograms) with a kernel smaller than the input. By doing this, CNNs identify sparser interactions than feedforward networks, as their connectivity is local. Due to the convolution operation, each kernel (or filter) is replicated across the entire input, which allows the network to learn how to detect features regardless of their “spatial positioning” in the input. CNNs can also operate on multichannel inputs, and the outputs of multiple kernels form a feature map in each hidden layer (see Figure 2.4).

In most applications of CNNs, a non-linear down-sampling operation called pooling is used. This operation has the goal of merging semantically similar features into one [5]. A typical pooling operation is max-pooling, where the maximum value of a group of neighbour features in a feature map is taken and propagated as a feature to the next layer of the network. The pooling operation reduces the dimensionality of the feature maps for the next layers in a network, and also allows the network to learn position-invariant feature detectors.

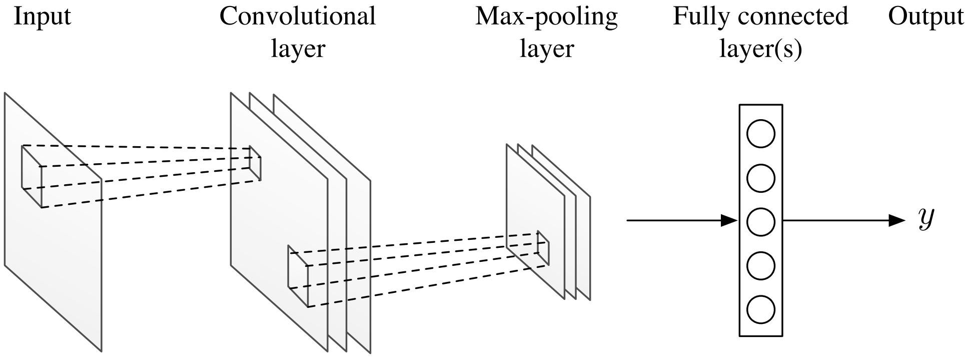


Figure 2.4 – Diagram illustrating the different layer types in a CNN

CNNs are popular in image processing, but have also potential to be used in speech applications, either by performing 1D convolutions or considering the spectrogram (or other time-frequency representation) of a speech signal as a 2D input [55, 56]. The idea of processing multichannel inputs can also be explored in speech applications with multiple microphones.

2.3 State of the art in DNN-based speech enhancement

2.3.1 Frequency-domain approaches

In our literature review, although we could find some work inspired by neural networks for source separation in the context of the computational auditory scene analysis research [57, 58], the first publications considering use of deep neural networks for speech enhancement since the renewed interest in such models coming from its success in acoustic modeling is the one by Wang et al. [59], where the authors combine a DNN with a linear support vector machine (SVM) to estimate binary masks for source separation as a classification problem. Although the DNN was trained as a classifier, the authors used a linear SVM based on the features learned by the DNN but do not report performance for the DNN model alone. The advantage of using a DNN together with a linear SVM in this context is that kernel machines do not scale well to large-scale training as the gradient descent-based approaches used in DNN. However, in a later paper by the same research group [60], a similar binary mask estimation task (for binaural speech signals) was solved as a classification task using DNNs only. These systems use the outputs of auditory filterbanks such as mel-frequency cepstral coefficients (MFCCs) and relative spectrum (RASTA) in [59] and gammatone filterbank plus binaural features in [60]) as inputs.

More recently, different targets were considered for speech enhancement with DNNs. In [61], the authors discuss the choice of training targets for speech separation by supervised learning algorithms such as DNNs. They compare results using ideal binary masks (IBM), ratio masks such as the ideal ratio mask (IRM), and FFT-mask, or directly predicting spectral magnitudes (either in the FFT or a filterbank domain). In the case of ratio masks or spectral magnitude prediction, the optimization problem assumes the form of a regression problem instead of an optimization problem as in previous works. Although [61] recommends the use of masking targets as opposed to direct prediction of the magnitudes to avoid issues with nonlinear compression, this seems to be dependent on the specific model and training conditions, as more recent papers reported better results when predicting spectral magnitudes instead of masks.

The system proposed by Xu et al. [10] predicts the spectral magnitudes and performs slightly better than the IRM and FFT-mask targets according to PESQ scores. The architecture with the best performance had 3 hidden layers with 2048 hidden units each, and was trained on 100 hours of speech sampled at 8 kHz, converted to the short-time Fourier transform (STFT) domain with windows of 32 ms and frame shift of 16 ms. The network used a context window of 11 frames as input. The speech data was corrupted by 104 different types of noise. The authors introduce the so-called noise-aware training procedure, where an estimate of the noise spectrum is obtained from the first few frames of each training sample as a feature, and show it leads to improved performance when compared to the approach using only the speech context information as input. Based on a short preference test (10 subjects and 32 real-world noisy utterances), subjects preferred the output of the DNN-based system over the output of a system based on the MMSE-STSA approach.

In [11], a network with a similar architecture (3 layers, 1600 hidden units per layer, and log-amplitude as target) is evaluated for enhancing reverberant and noisy speech. The authors propose an iterative procedure to minimize the incoherence between the noisy phase and estimated clean magnitude as a post-processing step, which consists of consecutively performing an inverse STFT followed by a STFT operation, keeping the estimated magnitude constant at each step but using the resulting phase of the previous step instead of the noisy phase at each new step. The network was trained with artificially corrupted speech (under noisy and reverberant conditions) and tested under several out-of-set conditions (different reverberation times and additive noise), and shown to outperform other enhancement algorithms and ideal binary masks according to the frequency-weighted SNR, PESQ, and STOI metrics. Using this approach as a preprocessing stage for a standard ASR system has also led to improved word-error rates for speech under noise (with relative improvements close to 30% in SNRs of -6 and -3 dB).

In a more recent paper [62], a DNN trained with the time-domain MSE as a cost function is proposed. Although the task of the network is still the estimation of a mask for the noisy magnitude spectrum, an inverse STFT using the noisy phase and the masked magnitude is performed prior to comparison to the target. This allows the system to take into account distortions due to incoherence between the estimated magnitude and the noisy phase. This system was shown to outperform previously DNN-based architectures that predicted IBMs and IRMs, as well as a recent source separation algorithm based on non-negative matrix factorization. Similar results are reported in [63], where a DNN is used to enhance the output of a non-negative matrix factorization stage for source separation, and [64], where the sources are estimated directly from the mix. In [65], a multichannel approach is proposed, where interchannel features and pre-enhanced speech (using soft masks estimated from the multichannel information) are included in the input. The neural network using only monaural information showed improved cepstral distortion and segmental SNR scores over the multichannel softmask approach, and including the pre-processed signal and interchannel features led to additional but smaller gains.

Fewer studies have explored architectures other than feedforward neural networks. The work by Weninger et al. [12, 51] proposes using RNNs based on LSTMs and bidirectional LSTMs (which process entire sentences simultaneously in the positive and negative direction in the time axis) trained on the Mel domain for noise reduction and source separation, and shows improved performance compared to networks using feedforward layers only.

Although the majority of the abovementioned methods operates solely on the magnitude spectrum of the STFT, there is some work that focuses on the complex STFT spectrum, such as [66]. The model takes as input a combination of complementary perceptually-motivated speech features, and tries to predict an ideal complex ratio mask (i.e., a ratio mask that operates both on the real and imaginary parts of the STFT spectrum). The authors reported significant improvements in objective quality and intelligibility metrics, compared to a model using the same input features but operating in the magnitude domain only.

2.3.2 Time-domain approaches

Since the advent of WaveNet [67], a speech synthesis model that operates on the time domain, several studies have focused on performing enhancement directly on the time domain instead of converting the signal to the frequency domain as an intermediate step. These approaches avoid the issue of reusing the noisy phase for resynthesizing the signal, as the output of these models is a time domain signal.

The main components of WaveNet are dilated causal convolutions. By using causal convolutions, the model uses strictly past samples to generate new samples. Dilated convolutions are convolutions that are able to cover an area larger than their length by skipping input values with a certain step (it is equivalent to filtering a downsampled version of the signal, however there are no antialiasing filters in dilated convolutions). Stacked dilated convolutions allow the model to use a large amount of past samples to generate future samples. A stack of residual blocks, composed of one dilated convolution and one 1×1 convolution (which produces a single output vector from a multi-channel input), is then connected through skip connections to the output layer, which predicts output samples as μ-law encoded 8 bit samples. WaveNet has been used as the audio front-end for speech synthesis systems [68] and also for speech denoising [69]. In the latter, the authors use non-causal convolutions and predict real-valued target fields instead of a single discrete sample per iteration (with the best results achieved by predicting 10 ms of speech per step). The model is trained conditioned on the identity of the speaker (as a binary-encoded scalar), but the authors also use an auxiliary code that represents an unknown speaker. While the model achieves higher performance (both in objective metrics and mean opinion scores) than an approach based on Wiener filtering, the authors state that it still has limitations, such as not being able to deal with sudden interferences. It should be noted that WaveNet and derived models are very costly from a computational point of view, due to their size and number of samples being processed, although there are works on optimizing it for performance.

There are other works that perform time-domain enhancement without relying on the WaveNet architecture. The Time-domain Audio Separation Network (TasNet) [70] uses input windows as short as 5 ms, so it can be implemented in real-time with low latency (although its performance can be enhanced by making it noncausal in case real-time processing is not a requirement). The model uses an encoder-decoder approach, where the encoder uses a gated convolution block, similar to the ones used in WaveNet, but the decoding is done by a stack of LSTMs followed by a fully-connected output layer. The model operates by predicting non-negative mixture weights for each segment and a masking matrix for each source, which is then multiplied by the mixture weights to recover the source weight matrices, that are finally multiplied by a matrix of basis signals to recover the time-domain segment corresponding to each source.

In [71], the authors incorporate an adversarial term to the cost function, inspired by work on generative adversarial networks [72]. Besides minimizing the L1 distance between the estimated samples and the target samples, the model is also trained to fool a classifier that tries to discriminate actual clean samples from enhanced samples. The model architecture uses 1D convolutional layers and follows an encoder-decoder approach, with skip connections between layers with similar dimensionality (as originally proposed in the UNet model [73]). Both the inputs and outputs are approximately 1 second long speech segments, which makes the model not useable for realtime and quasi-realtime applications (unless such a long delay can be tolerated). Another approach that relies on a generative adversarial framework can be found in [74], where the authors show that a DNN trained with a Wasserstein GAN formulation [75] outperforms a number of other models, including a non-negative matrix factorization approach, in terms of source-to-distortion ratio.

Chapter 3

Speech dereverberation with context-aware recurrent neural networks

3.1 Preamble

This chapter is compiled from material extracted from the manuscript published in the IEEE/ACM Transactions on Audio, Speech, and Language Processing [76].

3.2 Introduction

Reverberation plays an important role in the perceived quality of a sound signal produced in an enclosed environment. In highly reverberant environments, perceptual artifacts such as coloration and echoes are added to the direct sound signal, thus drastically reducing speech signal intelligibility, particularly for the hearing impaired [2]. Automatic speech recognition performance is also severely affected, especially when reverberation is combined with additive noise [1]. To deal with the distortions in such environments, several types of speech enhancement systems have been proposed, ranging from single-channel systems based on simple spectral subtraction [3] to multistage systems which leverage signals from multiple microphones [4].

Deep neural networks (DNNs) are currently part of many large-scale ASR systems, both as separate acoustic and language models as well as in end-to-end systems. To make these systems robust to reverberation, different strategies have been explored, such as feature enhancement during a preprocessing stage [9] and DNN-based beamforming on raw multichannel speech signals for end-to-end solutions [77]. On the other hand, the application of DNNs to more general speech enhancement problems such as denoising [78], dereverberation [11], and source separation [12] is comparatively in early stages.

Recently, several works have explored deep neural networks for speech enhancement through two main approaches: spectral estimation and spectral masking. In the first, the goal of the neural network is to predict the magnitude spectrum of enhanced speech signal directly, while the latter aims at predicting some form of an ideal mask (either a binary or a ratio mask) to be applied to the distorted input signal.

Most of the published work in the area uses an architecture similar to the one first presented in [11]: a relatively large feedforward neural network with three hidden layers containing several hundred units (1600 in [11]), having as input a context window containing an arbitrary number of frames (11 in [11]) of the log-magnitude spectrum and as target the dry/clean center frame of that window. The output layer uses a sigmoid activation function, which is bounded between 0 and 1, and normalizes targets between 0 and 1 using the minimum and maximum energies in the data. Moreover, an iterative signal reconstruction scheme inspired by [79] is used to reduce the effect of using the reverberant phase for reconstruction of the enhanced signal.

In [80], the authors performed a study on target feature activation and normalization and their impacts on the performance of DNN-based speech dereverberation systems. The authors compared the target activation/normalization scheme in [11] with a linear (unbounded) activation function and output normalized by its mean and variance. Their experiments showed the latter activation/normalization scheme leads to higher PESQ and frequency-weighted segmental SNR (fwSegSNR) scores than the sigmoid/min-max scheme.

In a follow-up study [81], the same authors proposed a reverberation-time-aware model for dereverberation that leverages knowledge of the fullband reverberation time (T60) in two different ways. First, the step size of the STFT is adjusted depending on T60, varying from 2 ms up to 8 ms (the window size is fixed at 32 ms). Second, the frame context used at the input of the network is also adjusted, from 1 frame (no context) up to 11 frames (5 future and 5 past frames). Since the input of the network has a fixed size, the context length is adjusted by zeroing the unused frames. The model was trained using speech from the TIMIT dataset convolved with 10 room impulse responses generated at a room with fixed geometry (6 by 4 by 3 meters), with T60 ranging from 0.1 to 1.0 s. The full training data had about 40 hours of reverberant speech, but the authors also presented results on a smaller subset with only 4 hours. The model proposed was a feedforward neural network with 3 hidden layers with 2048 hidden units each, trained using all the different step sizes and frame context configurations in order to find the best configuration for each T60 value. The authors considered the true T60 value to be known at test time (oracle T60), as well as estimated using the T60 estimator proposed by Keshavarz et al. [82].

The other well-known approach for speech enhancement using deep neural networks is to predict arbitrary ideal masks instead of the magnitude spectrum [61]. The ideal binary mask target transforms the speech enhancement problem into a classification problem, where the goal of the model is to predict which time-frequency cells from the input should be masked, and has been shown to improve intelligibility substantially. The ideal binary mask is defined quantitatively based on a local criterion threshold for the signal-to-noise ratio. Namely, if the SNR of a given time-frequency cell is lower than the threshold, that time-frequency cell is masked (set to zero). Alternatively, the ideal ratio mask is closely related to the frequency-domain Wiener filter with uncorrelated speech and noise. It is a soft masking technique where the mask value corresponds to the local ratio between the signal and the signal-plus-noise energies for each time-frequency cell. Recently, [66] has proposed the complex ideal ratio mask, which is applied to the real and imaginary components of the STFT instead of just the magnitude. Models based on mask prediction usually include several different features at the input (such as amplitude modulation spectrograms, RASTA-PLP, MFCC, and gammatone filterbank energies), instead of using just the magnitude spectrum from the STFT representation like the works previously described here. Masks are also often predicted in the gammatone filterbank domain.

In [83], the authors present a model that predicts log-magnitude spectrum and their delta/delta-delta, then performs enhancement by solving a least-squares problem with the predicted features, which aims at improving the smoothness of the enhanced magnitude spectrum. The method was shown to improve cepstral distance (CD), SNR, and log-likelihood ratio (LLR), but caused slight degradation of the speech-to-reverberation modulation energy ratio (SRMR). They also reported that the DNN mapping causes distortion for high T60.

Very few studies use architectures other than feed-forward for dereverberation. In [84], the authors propose an architecture based on long short-term memory for dereverberation. However, they only report mean-squared error and word-error rates for a baseline ASR system (the REVERB Challenge evaluation system) and do not report its effect on objective metrics for speech quality and intelligibility. The method described in [12] also uses recurrent neural networks based on LSTMs for speech separation in the mel-filterbank energies domain. Although their system predicts soft masks, similar to the ideal ratio mask, they use a signal-approximation objective instead of predicting arbitrary masks (i.e., the target are the mel-filterbank features of the clean signal, not an arbitrarily-designed mask).

Most current models reported in the literature only explore one of two possible contexts from the reverberant signal. Feed-forward models with a fixed window of an arbitrary number of past and future frames only take into account the local context (e.g. [81]) and are unable to represent the long-term structure of the signal. Also, since feed-forward models do not have an internal state that is kept between frames, the model is not aware of the frames it has predicted previously, which can lead to artifacts due to spectral discontinuities. LSTM-based architectures, on the other hand, are able to learn both short- and long-term structure. However, learning either of these structures is not enforced by the training algorithm or the architecture, so one cannot control whether the internal state will represent short-term, long-term context, or both.

In this chapter, we propose a novel architecture for speech dereverberation that leverages both short- and long-term context information. First, fixed local context information is generated directly from the input sequence by a convolutional context encoder. We train the network to learn how to use long-term context information by using recurrent layers and training it to enhance entire sentences at once, instead of a single frame at a time. Additionally, we leverage residual connections from the input to hidden layers and between hidden layers. We show that combining short and long-term contexts, as well as including such residual connections, substantially improves the dereverberation performance across four different objective speech quality and intelligibility metrics (PESQ, SRMR, STOI, and POLQA), and also reduces the amount of perceived reverberation according to subjective tests.

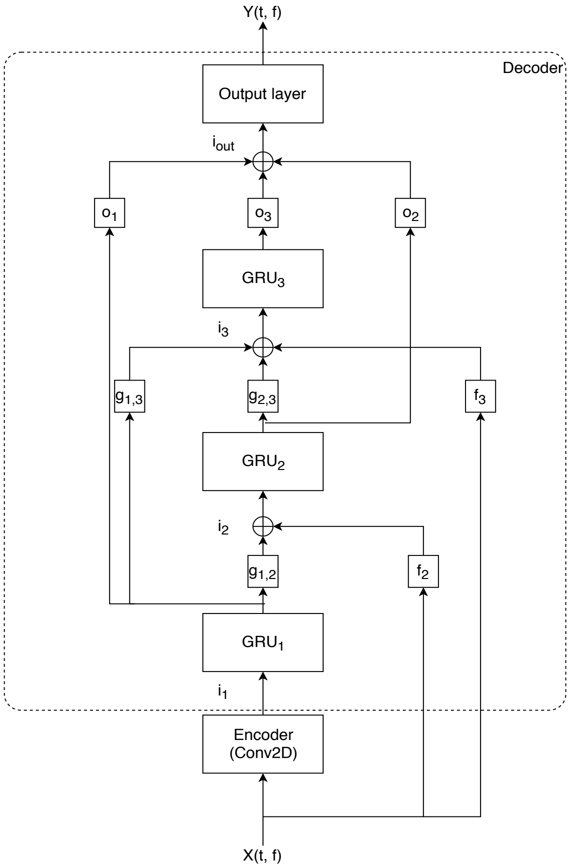


Figure 3.1 – Architecture of the proposed model

3.3 Proposed model

The architecture of the proposed model can be seen in Fig. 3.1. As discussed previously, our model combines both short- and long-term context by using a convolutional context encoder to create a representation of the short-term structure of the signal, and a recurrent decoder that is able to learn long-term structure from that representation. The decoder also benefits from residual connections, which allow each of its recurrent stages to have access both to a representation of the input signal and the state of the previous recurrent layer. Each of the blocks is further detailed in the sections to follow.

3.3.1 Context encoder

As shown in other studies, incorporating past and future frames can help on the task of estimating the current frame for dereverberation. Most works, however, use a fixed context window as the input to a fully-connected layer [11, 80]. In this work, we decided to extract local context features using 2D convolutional layers instead. By using a 2D convolutional layer, these features encode local context both in the frequency and the time axis. Our context encoder is composed by a single 2D convolutional layer with 64 filters with kernel sizes of $(21, C)$, where 21 corresponds to the number of frequency bins covered by the kernel and C to the number of frames covered by the kernel in the time axis. The model input are the log-magnitudes of the STFT representation, using a 32 ms window with 50% overlap obtained from signals sampled at 16 kHz. In our implementation, C is always an odd number as we use an equal number of past and future frames in the context window (e.g., $C = 11$ means the current frame plus 5 past and 5 future frames). We report the performance of the model for different values of C in section IV-A. The convolution has a stride of 2 in the frequency axis and is not strided in the time axis. The kernel size corresponds to a 176 ms window centered at the sample being estimated, with the frequency axis spanning 656.5 Hz.

3.3.2 Decoder

Following the encoder, we have a stack of three gated recurrent unit (GRU) layers [85] with 256 units each. The input to the first layer is the output of the convolutional context encoder with all of the channels concatenated to yield an input of shape (F, T), where

$$F = 64 \times \left\lfloor \frac{B - N_{context} + 1}{2} + 1 \right\rfloor,$$

64 is the number of filters in the convolutional layer, T is the number of STFT frames in a given sentence, B is the number of FFT bins (257 in our experiments), and $\lfloor . \rfloor$ is the *floor* operation (rounds its argument down to an integer).

The outputs to the remaining GRU layers are a combination of affine projections of the input and the states of the previous GRU layers. Consider $x_{enc}(t)$ to be the encoded input at timestep t, and $h_1(t), h_2(t), h_3(t)$ as the hidden state at timestep t for the first, second, and third GRU layers, respectively. Then, the inputs $i_1(t), i_2(t), i_3(t)$ are as follows:

$$\begin{aligned}
i_1(t) &= x_{enc}(t),\\
i_2(t) &= f_2(x(t)) + g_{1,2}(h_1(t)),\\
i_3(t) &= f_3(x(t)) + g_{1,3}(h_1(t)) + g_{2,3}(h_2(t)),
\end{aligned}$$

where $f_i, g_{i,j}$ are affine projections from the input or previous hidden states. The parameters of those projections are learned during training, and all projections have an output dimension of 256 in order to match the input dimension of the GRUs when added together.

Similarly, the output layer following the stacked GRUs has as its input the sum of affine projections o_i of h_1, h_2, h_3:

$$i_{\text{out}}(t) = o_1(h_1(t)) + o_2(h_2(t)) + o_3(h_3(t)).$$

The parameters of those projections are also learned during training. All of these projections have an output dimension of 256.

3.4 Experimental setup

3.4.1 Datasets

In order to assess the benefits of the proposed architecture for speech dereverberation, we ran a series of experiments with both the proposed model and two other models as baselines: the T60-aware model proposed in [81] and a similar model without T60 information that uses a fixed overlap of 16 ms and a fixed context of 11 frames (5 past and 5 future frames). For the T60 aware model, we extracted T60 values directly from the RIRs (oracle T60s) using a method similar to the one used for the ACE Challenge dataset [86, 87]. The fullband T60 we used was computed as the average of the estimates for the bands with center frequencies of 400 Hz, 500 Hz, 630 Hz, 800 Hz, 1000 Hz, and 1250 Hz.

For the single speaker experiments, a recording of the IEEE dataset uttered by a single male speaker was used [17]. The dataset consists of 72 lists with 10 sentences each recorded under anechoic and noise-free conditions. We used the first 67 lists for the training set and the remaining 5 lists for testing. Reverberant utterances were generated by convolving randomly selected subsets of the utterances in the training set with 740 RIRs generated using a fast implementation of the image-source method [88], with T60 ranging from 0.2 s to 2.0 s in 0.05 s steps. Twenty different RIRs (with different room geometry, source-microphone positioning and absorption characteristics) were generated for each T60 value. Fifty random utterances from the training set were convolved with each of these 740 RIRs, resulting in 37,000 files. A random subset of 5% of these files was selected as a validation set and used for model selection and the remaining 35,150 files were used to train the models. The test set was generated in a similar way, but using a different set of 740 simulated RIRs and 5 utterances (randomly selected from the test lists) were convolved with each RIR.

For the multi-speaker experiments, we performed a similar procedure but using the TIMIT dataset [89] instead of the IEEE dataset. The default training set (without the "SA" utterances, since these utterances were recorded by all speakers) was used for generating the training and validation sets, and the test set (with the SA utterances removed as well) was used for generating the test set. The training and test sets had a total of 462 and 168 speakers, respectively. The utterances were convolved with the same RIRs used for the single speaker experiments. A total of 3696 clean utterances were used for the training and validation set, and 1336 for the test set. As with the single speaker dataset, 50 sentences were chosen at random from the training and validation sets (which include all of the utterances from all speakers) and convolved with each of the 740 simulated RIRs to generate the training/validation sets. The test set was generated following the same procedure as used for the single speaker dataset.

Additionally, in order to explore the performance of the proposed and baseline models on realistic settings, we tested the same single speaker models described above with sentences convolved with real RIRs from the ACE Challenge dataset [86]. We used the RIRs corresponding to channels 1 and 5 of the cruciform microphone array for all of the seven rooms and two microphone positions, leading to a total of 28 RIRs. The test sentences were the same as for the experiments with simulated RIRs. The interested reader can refer to [86] for more details about the ACE Challenge RIRs.

3.4.2 Model Training

Both the proposed model and the baselines were implemented using the PyTorch library [90] (revision *e1278d4*) and trained using the Adam optimizer [91] with a learning rate of 0.001, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. The loss function is the mean-squared error between the estimated log-magnitudes and the log-magnitudes for the original clean signals. The models were trained for 100 epochs and the parameters corresponding to the epoch with the lowest validation error were used for evaluation. Since sequences in the dataset have different lengths, we padded sequences in each minibatch and used masking to compute the MSE loss only for valid timesteps. The models used for evaluation were the ones with the lowest validation loss amongst the 100 epochs.

3.4.3 Objective evaluation metrics

We compare the performance of the models using four different objective metrics. The Perceptual Evaluation of Speech Quality [92] and Perceptual Objective Listening Quality Assessment (POLQA) [93] are both ITU-T standards for intrusive speech quality measurement. The PESQ standard was designed for a very limited test scenario (automated assessment of band-limited speech quality by a user of a telephony system), and was superseded by POLQA. The two metrics work in a similar way, by computing and accumulating distortions, and then mapping them into a five-point mean opinion score (MOS) scale. Even though PESQ is not recommended as a metric for enhanced or reverberant speech, several works report PESQ scores for these types of processing and we report it here for the sake of completeness. POLQA is a more complete model and allows measurement in a broader set of conditions (e.g., super-wideband speech, combined additive noise and reverberation). The Short-Time Objective Intelligibility [94] metric is an intrusive speech intelligibility metric based on the correlation of normalized filterbank envelopes in short-time (400 ms) frames of speech. Finally, the speech-to-reverberation modulation energy ratio (SRMR) is a non-intrusive speech quality and intelligibility metric based on the modulation spectrum characteristics of clean and distorted speech, which has been shown to perform well for reverberant and dereverberated speech [95, 96].

3.4.4 Subjective listening tests

To further gauge the benefits of the proposed architecture for speech enhancement, we performed a small-scale subjective listening test to assess how effective different methods are in reducing the amount of perceived reverberation. In particular, comparisons with the baseline model Wu2016 were performed as it resulted in improved performance with unmatched speakers relative to the benchmark Wu2017 model. The test protocol we used was the recently proposed MUSHRAR test [97].

For the reverberation perception tests, five random samples of the TIMIT test dataset were chosen for each of simulated RIRs with T60s of 0.6, 0.9, 1.2, and 1.5 s and participants were asked to compare the outputs of all models and the reverberant signal to a reference signal. In addition to the model outputs and the reverberant signal, a hidden reference and anchor were used for both experiments: the reference signal was the anechoic signal convolved with an RIR with T60 of 0.2 s and the anchor was the anechoic signal convolved with a RIR with a T60 of 2.0 s. Users were asked to rate the amount of perceived reverberation on a 0-100 scale, with higher values corresponding to higher perceived reverberation. A total of nine participants took part in the test. Tests were performed using a web interface, which is freely available online[1].

3.5 Experimental Results

In this Section, we present the results for four different experiments, namely:

A. Evaluation of the effect of the context size (using the single-speaker, simulated RIR dataset)

B. Comparison between the proposed architecture and baselines on matched conditions (same speaker, simulated RIRs for training and testing)

C. Comparison between the proposed architecture and baselines on mismatched speakers (multispeaker dataset with simulated RIRs)

[1] Software available at `https://github.com/jfsantos/mushra-ruby`

Table 3.1 – Number of parameters in each model

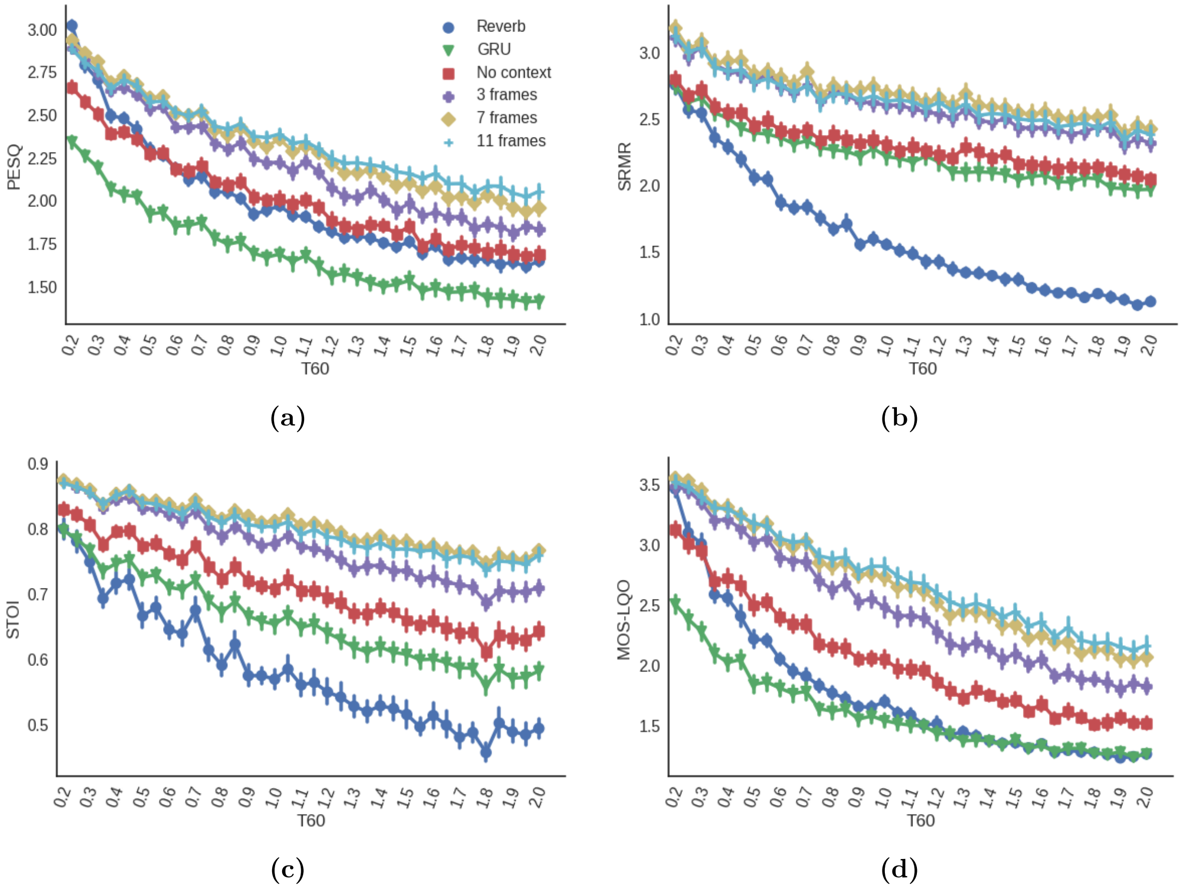
Model	# of parameters
GRU	4,458,753
Wu2016 and Wu2017	14,711,041
Proposed without context	1,838,593
Proposed, context = 3 frames	7,429,121
Proposed, context = 7 frames	7,434,497
Proposed, context = 11 frames	7,439,873

D. Comparison between the proposed architecture and baselines under realistic reverberation conditions (models trained on single-speaker, simulated RIR dataset, and tested on single-speaker, real RIR dataset).

Since some of the models have a large difference in the number of learnable parameters, we list the number of parameters for each of the models that was used in our experiments in Table 3.1.

3.5.1 Effect of context size

We first analyze the effect of the context size in the proposed model by comparing models with context sizes of 3, 7, and 11. We also included a baseline model based on GRUs without residual connections and a model without the context encoder and residual connections in these tests. All the models were trained and tested with single-speaker data convolved with simulated RIRs. The results can be seen in Figure 3.2. For this and all of the subsequent experiments in this section, we have subplots for the four objective metrics listed in Section 3.4.3: (a) PESQ, (b) SRMR, (c) STOI and (d) POLQA in this order. The GRU-only model (without the context encoder and residual connections) underperforms in all metrics, and in the cases of PESQ and POLQA, even leads to lower scores than the reverberant utterances. Adding only the residual connections and no context at all (which allows the model to perform at real-time) significantly increases the performance and leads to improvements in all metrics, except for reverberation times under 400 ms (noticeable in both PESQ and POLQA). Adding the context encoder, even with a short context of one previous and one future frame (shown as "3 frames" in the plots), leads to additional improvements. Adding more frames ("7 frames" and "11 frames") improves the performance even further. However, we see diminishing returns around 7-11 frames, as these two models have similar performances for SRMR and STOI and only a slight difference in PESQ and POLQA. Since the model with a context of 11



(a) **(b)**

(c) **(d)**

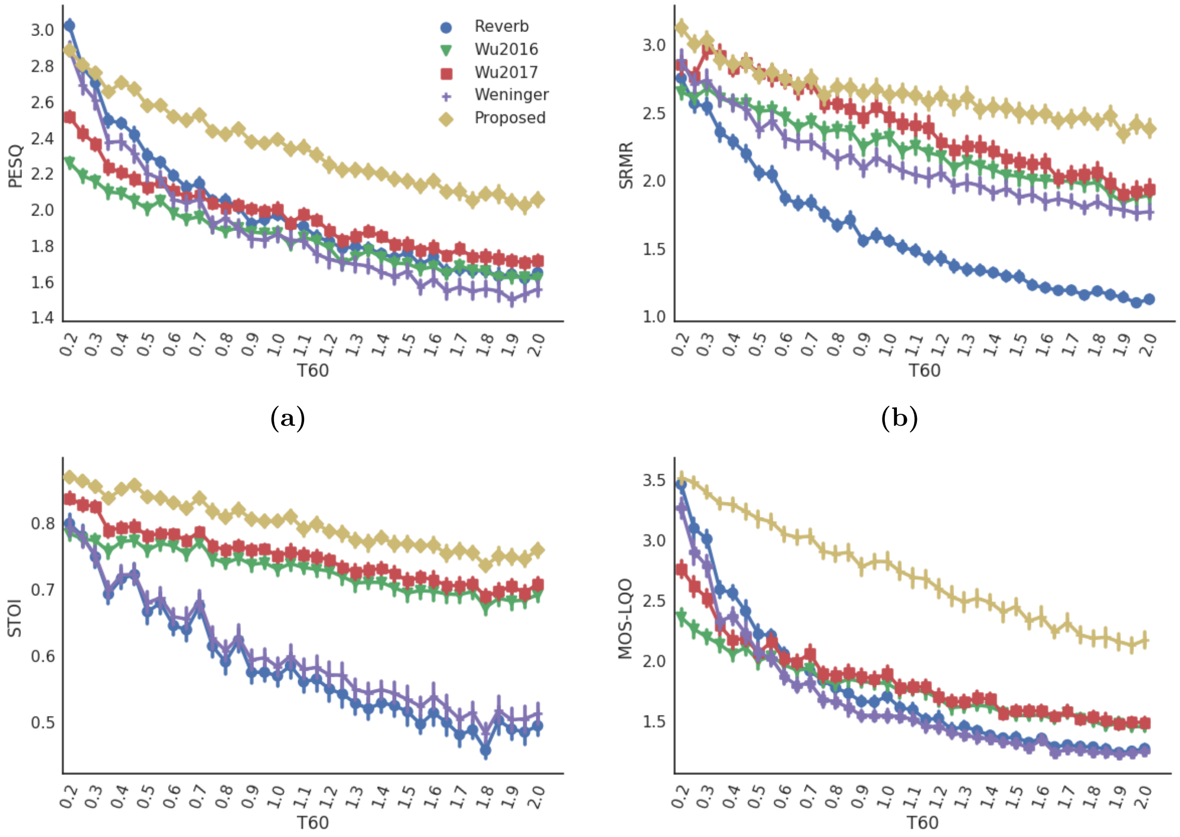
Figure 3.2 – Effect of context size on scores: (a) PESQ, (b) SRMR, (c) STOI, and (d) POLQA.

frames (5 past and 5 future frames) showed the best performance amongst all context sizes, in the next experiments we show only results with this context size for the proposed model.

3.5.2 Comparison with baseline models

Figure 3.3 shows a comparison between our best model, Wu2016 (fixed STFT window and hop sizes) and Wu2017 (STFT window and hop sizes dependent on the oracle T60 values). The interested reader is referred to this manuscript's supplementary material page to listen to audio samples generated by the proposed and benchmark algorithms [2]. A more complete audio demo can be found in [98]. It can be seen that our model outperforms these two feed-forward models in most scenarios and metrics, except for SRMR with T60 $<$ 0.7 s, where the results are very similar. It should be noted, however, that the SRMR metric is less accurate for lower T60s [95]. Even though it

[2] `http://www.seaandsailor.com/demo/index.html`



(a) **(b)**

(c) **(d)**

Figure 3.3 – Single speaker, simulated RIR scores: (a) PESQ, (b) SRMR, (c) STOI, and (d) POLQA.

uses oracle T60 information, the model Wu2017 does not have a large improvement in metrics when compared to Wu2016, especially in the STOI and POLQA metrics (which are more sensitive to the effects of reverberation than PESQ). Using T60 information to adapt the STFT representation seems to have a stronger effect for lower T60. It should also be noted that these models lead to a reduction in PESQ scores for lower reverberation times in the PESQ and POLQA metrics, which is probably due to the introduction of artifacts. This is also observed for the proposed model, but only in the PESQ metric and only for T60 < 0.4 s. On the other hand, STOI indicates all models lead to an improvement in intelligibility.

3.5.3 Unmatched vs. matched speakers

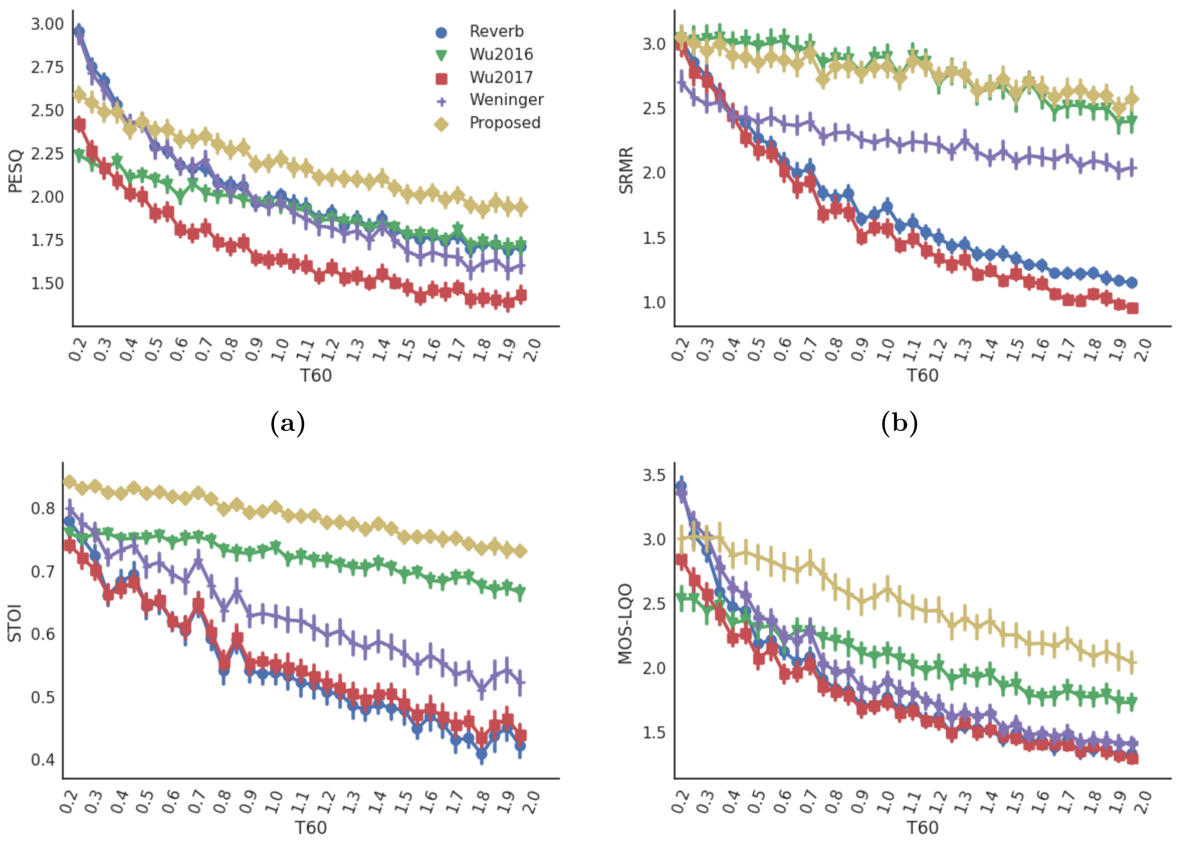
Although testing dereverberation models with single speaker datasets allows assessing basic model functionality, in real-world conditions, training a model for a single speaker is not practical and has very limited applications. It is important to evaluate the generalization capabilities of such models with mismatched speakers, as this is a more likely scenario. Figure 3.4 shows the results for Wu2016, Wu2017, the proposed model, as well as for the reverberant files in the test set. Between the baseline models, we can see that Wu2017 now underperforms Wu2016 in all metrics, and either reduces metric scores (as for PESQ and POLQA for low T60) or does not change them. However, the version of the model that does not depend on T60 leads to higher scores. Although we do not have a clear explanation for this behaviour, we believe the optimal scores for the STFT hop size and context might depend both on T60 and speaker, but the Wu2017 model uses fixed values that depend only on T60. Since the model has now less data from each speaker, it was not able to exploit these adapted features properly.

The proposed model, on the other hand, outperforms both baselines in 3 out of 4 metrics, only achieving similar scores in the SRMR metric. Compared to Wu2016, our model leads to improvements of around 0.4 in PESQ, 0.1 in STOI, and 0.5 in POLQA.

3.5.4 Real vs. simulated RIR

In our last experiment, we test the generalization capability of the models trained on simulated RIRs to real RIR. To that end, we used speech convolved with real RIRs from the ACE Challenge dataset, as specified in Section IV. The results are reported in Figure 3.5. Note that the x-axis in that figure does not have linear spacing in time, as we are showing the results for each T60 as a single point uniformly spaced from its nearest neighbours in the data. Note also the larger variability in scores for reverberant files, which is due to the scores here not being averaged across many RIRs.

In this test, the proposed method achieves the highest scores in all metrics. It is also the only method to improve PESQ and POLQA across all scenarios, while the baselines either decrease or do not improve such metrics. Although the baselines do improve STOI in most cases, in some scenarios they actually decrease STOI.



(a) (b) (c) (d)

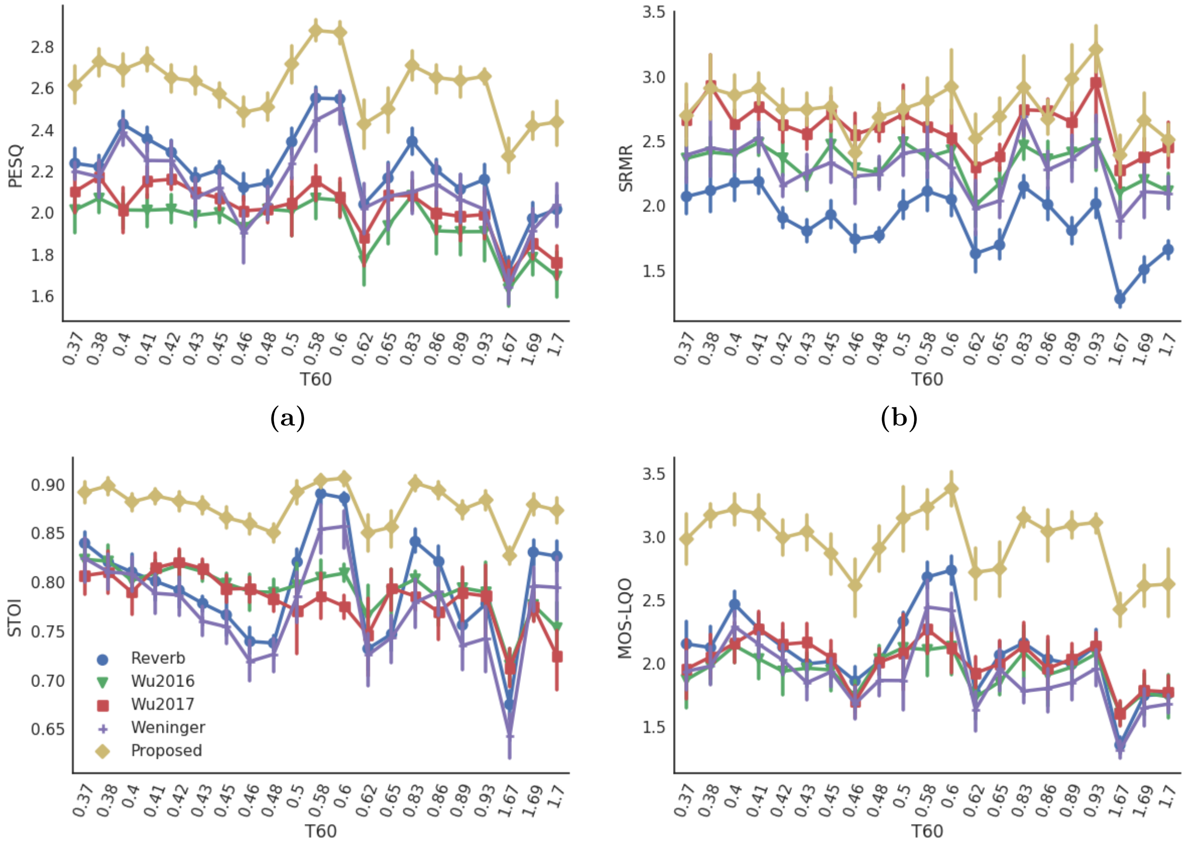
Figure 3.4 – Scores for experiment with unmatched speakers: (a) PESQ, (b) SRMR, (c) STOI, and (d) POLQA.

Table 3.2 – Results of the MUSHRAR test for reverberation perception for different T60 values. Lower is better.

Model	0.6 s	0.9 s	1.2 s	1.5 s
Reverb	42.62	60.06	71.73	82.24
Wu2016	32.27	38.08	46.27	43.89
Proposed	24.44	23.89	29.93	31.64

3.5.5 Subjective Listening Tests

The results of MUSHRAR tests are summarized in Table 3.2. The results for the original, unprocessed reverberant files are also reported. When rating how reverberant the enhanced stimuli were, participants rated the outputs of the proposed model as less reverberant than the baseline and reverberant signals for all T60 values, while the baseline model only achieved a lower reverberation perception for lower reverberation times.



(a)

(b)

(c)

(d)

Figure 3.5 – Scores for experiment with real RIRs: (a) PESQ, (b) SRMR, (c) STOI, and (d) POLQA.

3.6 Discussion

3.6.1 Effect of the context size and residual connections

In this chapter, we propose the addition of a few components to the architecture of deep neural networks for dereverberation. Using a context window for speech enhancement is not a new idea, and most studies using deep neural networks use similar context sizes. The novelty in our approach, however, is to use a convolutional layer as a local context encoder, which we believe helps the model to learn how to extract local features both in the time and the frequency axes in a more efficient way than just using a single context window as the input for a feedforward model. This local context, together with the residual input connections, is used as an input for the recurrent decoder, which is able to learn longer-term features. Similar architectures (save for the residual connections) have been successfully used for speech recognition tasks [99] but, to the best of our knowledge, this is the first work where such an architecture is used with the goal of estimating the clean speech magnitude spectrum.

Regarding the effect of the context size, our results agree with the intuition that including more past and future frames would help in predicting the current frame. However, we also show that there is a very small difference between using 7 vs. 11 frames as context. A window of 7 frames encompasses a total of 144 ms, vs. 208 ms for 11 frames. Both sizes are still much smaller than most of the reverberation times being used for our study, but they already start allowing the model to easily extract features related to amplitude modulations with lower frequencies, which are very important for speech intelligibility. We believe the combination of short- and long-term contexts, by having both the short-term context encoder and longer-term features through the recurrent layers, allows our architecture to benefit even from a shorter context window at the input. Lower modulation frequencies can still be captured through the recurrent layers, although we cannot explicitly control or assess how long these contexts are since they are learned implicitly through training and might be input-dependent.

We also introduce the use of residual and skip connections in the context of speech enhancement. Residual connections have a very close corresponding method in speech enhancement, namely spectral subtraction. The output of a speech enhancement model is likely to be very similar to the input, save for a signal that has to be subtracted from it (e.g. in the case of additive noise). In our case, our model is not restricted to subtraction due to how our architecture was designed. Consider the architecture as shown in Figure 1. The input to each recurrent layer is the sum of projections of the corrupted input (via the linear layers f_1, f_2, and f_3), and the output of the previous layers (via the linear layers g_1, g_2, and g_3). One possible solution to the problem, given this architecture, is for each recurrent layer to perform a new stage of spectral estimation given the difference between what the previous stage has predicted and the input at a given time. The output layer uses skip connections to the output of each recurrent layer and combines their predictions, allowing each layer to specialize on removing different types of distortions or to successively improve the signal. Our architecture was inspired by the work on generative models by Alex Graves [100], which uses a similar scheme of connections between recurrent layers. The recently proposed WaveNet architecture [67], which is a generative model for audio signals able to synthesize high-quality speech, also makes use of residual connections and skip connections from each intermediate layer and the output layer.

3.6.2 Comparison with baselines

As seen in the previous section, our model outperforms both baselines based on feed-forward neural networks. One clear advantage of our approach is that we do not need to predict T60, which is a hard problem in itself and adds another layer of complexity to the model, especially under the presence of other distortions such as background noise [86].

Regarding model size, as seen in Table 1, it is important to note that despite having a significantly smaller number of parameters than the baseline models, our proposed model consistently outperforms them in most experiments. Our best model (shown in the last line of Table 1) has approximately half the number of parameters but uses these parameters more efficiently because of its architectural characteristics.

Although the authors of [81] argue that their model leads to improvements in PESQ scores, the differences reported were not significant. Also, the authors have used PESQ scores for selecting the best models; however, that might be an issue, especially because PESQ is not a recommended metric for reverberant/dereverberated speech. In our experiments we used the best validation loss (MSE) for model selection for all the models, including the baselines. Using a proper speech quality or intelligibility metric as a function for model selection would be a good, but costlier solution because one has to generate/evaluate samples using that metric for all epochs, while MSE can be computed directly from the output of the model.

3.6.3 Generalization capabilities

Another important aspect of our study, when compared to [81], is that we train and test the models under several different room impulse responses, both real and simulated. The results we report for Wu2017 have much lower performance than those reported in [81] for similar T60 values; however, it must be noted we used random room geometries and random source and microphone locations for our RIRs, while in their experiments the authors used a fixed room geometry with different T60 values (corresponding to different absorption coefficients in the surfaces of the room) and a fixed location. Although they tried to show the model generalizes to different room sizes, they tested generalization by means of two tests: (a) a single different room size with the same source-microphone positioning and (b) and a single different source-microphone positioning with fixed room geometry. We believe those experiments were not sufficient to show the generalization capabilities of the model, and experiments 1 and 3 in our work confirm that hypothesis. In our work, we tried to expose the model to several different room geometries and source-microphone positioning, since this is closer to real-world conditions and helps the model to better generalize to unseen rooms and setups.

3.6.4 Anedoctal comparison with mask-based methods

Although we did not compare our method to masking alternatives in this study, we would like to briefly mention the results reported in the most recent paper with a masking approach [101], which used a similar single-speaker dataset (IEEE sentences uttered by a male speaker), simulated and real RIR for the dereverberation task. Although the range of T60s in our study and theirs is not similar, we can roughly compare the metrics of our model in the same range used in their study. The STOI improvements for our method are higher than the so-called CRM method proposed in that paper: our method has a STOI improvement of approximately 0.2 in all simulated T60s they report (0.3, 0.6, and 0.9 s), while their highest improvement is of 0.06 for 0.9 s. The CRM method, however, slightly decreased STOI for a T60 of 0.3 s.

3.6.5 Study limitations

Although we report both PESQ and SRMR scores in this work, these results should be taken cautiously. PESQ has been shown to not correlate well with reverberant and dereverberated speech. SRMR, on the other hand, is a non-intrusive metric, so it is affected by speech- and speaker-related variability [95]. It is also more sensitive to high reverberation times (0.8 s as shown in [102]), so measurements in low reverberation times might not be accurate enough for drawing conclusions about the performance of a given model.

A characteristic common to all DNN-based models and also other algorithms based on spectral magnitude estimation is a higher level of distortion due to reusing the reverberant phase, especially in higher T60. This is added to magnitude estimation errors, which are also higher in higher T60, since the input signal is very different from the target due to the combined effect of coloration and longer decay times. Although we do not try to tackle this issue here, there are recent developments in the field that could be applied jointly with our model, such as the method recently proposed in [103]. This exploration is left for future work.

3.7 Conclusion

We proposed a novel deep neural network architecture for performing speech dereverberation through magnitude spectrum estimation. We showed that this architecture outperforms current state-of-the-art architectures and generalizes over different room geometries and T60s (including real RIR), as well as to different speakers. Our architecture extracts features both in a local context (i.e., a few frames to the past/future of the frame being estimated) as well as long-term context. As future work, we intend to explore improved cost functions (e.g., incorporating sparsity in the outputs [104]) as well as applying the architecture to signals distorted with both additive noise and reverberation. We also intend to propose a multichannel extension of the architecture in the future. Finally, we intend to explore a number of solutions to the issue of reconstructing the signal using the reverberant phase.

Chapter 4

Interpreting and optimizing DNN-based speech enhancement models

4.1 Preamble

This chapter contains work that was presented at the thirty-second Neural Information Processing Systems workshop on Interpretability and Robustness in Audio, Speech and Language, as well as submitted to the IEEE/ACM Transactions on Audio, Speech, and Language Processing.

4.2 Introduction

While using deep neural networks for speech enhancement brings many advantages, as presented in previous chapters, such as not needing arbitrary models for speech and/or distortions, there are also some disadvantages. With data-driven approaches, it is often hard to extract the knowledge learned by such a model and interpret it to come up with new insights about the data itself. The model learns how to represent the data in a way that is useful for the problem at hand, but it is not necessarily meaningful to someone looking at it.

This chapter aims to present some ideas that bridge that gap by proposing methods that lead to more interpretable models, as well as trying to better understand the internal representation of the model presented on Chapter 3. Those insights led us to propose a novel method for model pruning that leads to significant model compression without an observed loss in performance.

4.3 Designing more interpretable models through the use of skip connections

Highway [105] and residual networks [106] have been proposed with the objective of improving activation and gradient flow in the training of deep neural networks. On the other hand, in tasks like image reconstruction or speech enhancement, the use of such skip connections serves a different purpose: if we model a corrupted signal $x = y + n$ as the addition of noise n to a clean signal y and x is the input to a neural network, we know that the task at hand is to predict n. In other words, to predict y, we have to alter the input x by subtracting n.

In speech enhancement, the two more commonly used approaches are spectral subtraction and spectral masking. In the first, a statistical model of n is used to predict its magnitude spectrum N, which is then subtracted from the input spectrum X to yield a clean magnitude spectrum estimate $\hat{Y}$. In spectral masking, instead of performing subtraction, we find a multiplicative mask M which aims at either blocking time-frequency cells dominated by noise (in the case of binary masks) or scaling down energies in such time-frequency cells to make them match that of the original clean signal. Recent work in speech enhancement has explored skip connections as a way of performing masking [107] and spectral estimation [76]. Time domain approaches, such as SEGAN [71], use a UNet-style network that employs multiple skip connections as well. Other works, such as [101], perform spectral masking but learn how to estimate an ideal mask instead of having the masking mechanism embedded in the neural network as a skip connection.

For better understanding of such models, we would like to understand whether there are any parallels between such connections and two traditional DSP approaches to speech enhancement, namely spectral subtraction and spectral masking. We also want to understand whether models using skip connections perform better for enhancement when such connections appear only once (resembling their DSP counterparts) or repeated as multiple blocks (like in highway and residual networks).

4.3.1 Residual and highway networks

Highway networks and residual networks are related, in the sense that residual networks can be seen as a special case of highway networks. A highway network block, as proposed in [105], can be described by the following equation:

$$\boldsymbol{y} = H(\boldsymbol{x}, \boldsymbol{\Theta_H}) \cdot T(\boldsymbol{x}, \boldsymbol{\Theta_T}) + \boldsymbol{x} \cdot (1 - T(\boldsymbol{x}, \boldsymbol{\Theta_T})). \tag{4.1}$$

Residual blocks [106], on the other hand, can be described by the following equation:

$$\boldsymbol{y} = H(\boldsymbol{x}, \boldsymbol{\Theta_H}) + \boldsymbol{x}, \tag{4.2}$$

which is equivalent (save for a multiplicative constant) to having a fixed gate T that outputs 0.5 for all inputs. In this work, we additionally test another type of skip connection, which we call here a masking block. In this type of connection, the computation in the block serves only to compute a multiplicative gating function which is then applied to the input of the block, as follows:

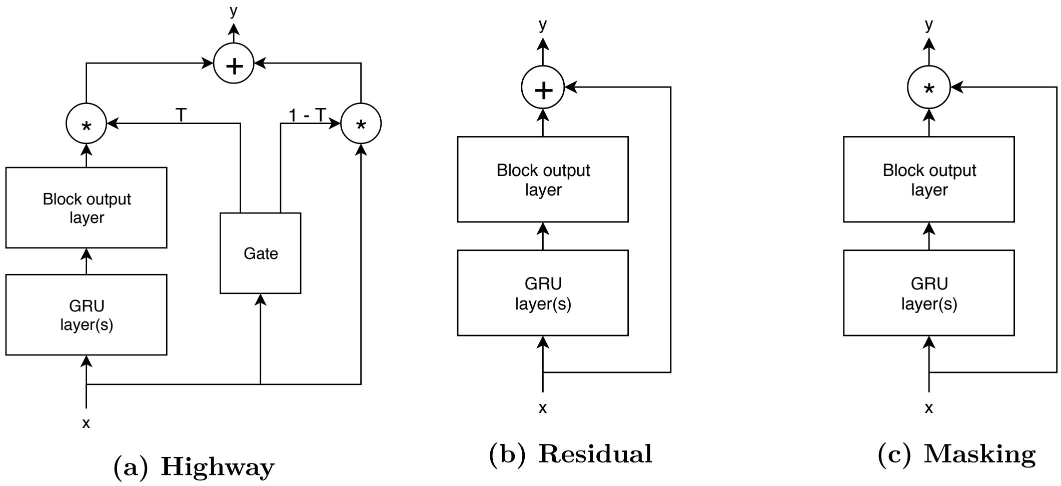
$$\boldsymbol{y} = M(H(\boldsymbol{x}, \boldsymbol{\Theta_H}), \boldsymbol{\Theta_M}) \cdot \boldsymbol{x}. \tag{4.3}$$

This block can be seen as a highway network without the residual connection.

In this work, we used the above mentioned skip connections with stacked gated recurrent units followed by a feedforward output layer with linear activation for each block, as illustrated in figure 4.1.

4.3.2 Skip connections in speech enhancement models

Spectral subtraction [3] speech enhancement models assume distortion is additive and try to predict the magnitude spectrum of the distortion and subtract it from the magnitude spectrum of the input



(a) Highway **(b) Residual** **(c) Masking**

Figure 4.1 – Diagrams for highway, residual, and masking blocks used in this work

to obtain an estimate of the clean signal:

$$\boldsymbol{y} = \boldsymbol{x} - N(\boldsymbol{x}, \boldsymbol{\Theta_N}), \tag{4.4}$$

where $N(\boldsymbol{x}, \boldsymbol{\Theta_N})$ is a noise estimation model with parameters $\boldsymbol{\Theta_N}$. In speech enhancement, such models usually predict the noise magnitude spectrum based on several past input frames. When applying the model based on residual blocks presented in this work to linear spectral amplitudes, the operations performed by the residual model are similar to the ones performed by spectral subtraction.

Spectral masking, on the other hand, is based on predicting time-frequency cells dominated by noise and creating a multiplicative mask that filters them out:

$$\boldsymbol{y} = \boldsymbol{x} \cdot M(\boldsymbol{x}, \boldsymbol{\Theta_M}). \tag{4.5}$$

Predicted masks can be either binary or ratio masks, although it has been shown that ratio masks potentially lead to better speech quality [61]. The residual model presented in the previous section, when applied to log-magnitude features, as well as the work done in [107, 66], perform enhancement using masking.

Highway networks, on the other hand, can be indirectly related to both spectral masking and subtraction (when applied to the linear amplitude spectrum): the output of each highway block is the sum of a masked input signal and a masked predicted signal.

4.3.3 Experiments

Datasets

For the experiments reported in this section, we used a single speaker dataset that is publicly available[1]. It is a relatively small dataset that is comprised of the IEEE sentence list. This list contains 720 phonetically balanced sentences, separated into 72 lists with 10 sentences each, uttered by the same male speaker. We split the list into 680 sentences for training and validation (sentence lists 01 to 67, and 50 sentences (sentence lists 68 to 72) for testing. Both datasets can be reproduced by running the dataset generation scripts in our code repository[2].

For the denoising dataset, we mixed the training and validation sentences with noises from the DEMAND dataset [108] at SNRs of 12, 6, 3, 0, -3, and -6 dB. The testing sentences were mixed with four noises (babble, factory1, factory2, and volvo) from the NOISEX dataset [109], at SNRs of 13, 7, 4, 1, -2, and -5 dB. For each sentence, a random noise segment with the same length as the sentence was picked. Signal energy for speech signals was computed according to the P.56 standard (which aims at only considering energy from speech segments and discarding silence segments), while the energy for noise signals was computed by its overall RMS value. For each (noise type, SNR) pair, we mixed all sentences in the training + validation set or the test set, accordingly. The training, validation, and test sets have 64923, 3417, and 1200 samples each.

The dereverberation dataset was constructed in a similar way, but using simulated room impulse responses obtained using the fast Image-Source Method [88]. We generated 20 RIRs for each reverberation time in the range 0.2 to 2.0, in 0.05 increments, and convolved 50 sentences to each RIR (given the large number of RIRs, we did not use the entire dataset for each). The training, validation, and test sets have 35150, 1850, and 3700 samples each.

Model architectures, hyperparameters, and training

For training all models we used the same input representation: the log-magnitude spectrum of the signal's short-time Fourier transform with a window of 32 ms and 50% overlap, which corresponds to a dimensionality of 257 coefficients per frame. The inputs were standardized, but the outputs were

[1] `https://www.crcpress.com/downloads/K14513/K14513_CD_Files.zip`
[2] `https://github.com/jfsantos/irasl2018`

not changed. Note that we tried to use linear features as well, but all models presented numerical issues when trained with linear features and did not converge.

All models used 3 layers of gated recurrent units with 256 hidden units each. We tried both having a single skip connection wrapping the 3 GRU layers as a single residual or highway block, and having 3 blocks with one GRU layer each. All models were trained using Adam [110] with standard hyper-parameters for 100 epochs, with a mini-batch size of 32.

4.3.4 Analysis

Visualizing the role of skip connections

To understand how each type of skip connection and each block in a model works, we used visualization of each output of a block. Since skip connections perform operations between the output of each block and its input, we can assume all operations happen in the domain of the transformed input, and can therefore be visualized by using the output layer of the model to bring them back to the STFT domain.

Figure 4.2 shows the intermediate and final outputs of the residual model with 3 blocks for a sentence corrupted by babble noise at 1 dB. The first row shows the input and respective target, the second row the residuals predicted in all blocks, and the last row the output of the residual block. The x and y axes represent the time and frequency bin respectively, with the color representing the log-magnitude (with blue representing low values, green intermediate values, and yellow high values). We can see that earlier blocks exhibit less frequency selectivity than later blocks, with the last block specializing strongly in certain frequencies. From block to block, we also see how the strongest components of the signal take shape, starting at the lower frequencies, followed by mid-bands and finally higher frequencies in the last stage.

Figure 4.4 shows a similar plot for the masking model, with the second row representing the predicted masks. Since we chose to use linear masks instead of bounded masks (such as the outputs of a tanh or sigmoid activation), these masks can flip the sign of time-frequency cells, which makes visualization a bit more complicated to follow; therefore, we used a different color map for plotting the mask, where values close to zero are shown in white, negative numbers in blue, and positive numbers in red. From the intermediate outputs, we can see in the first two plots that the

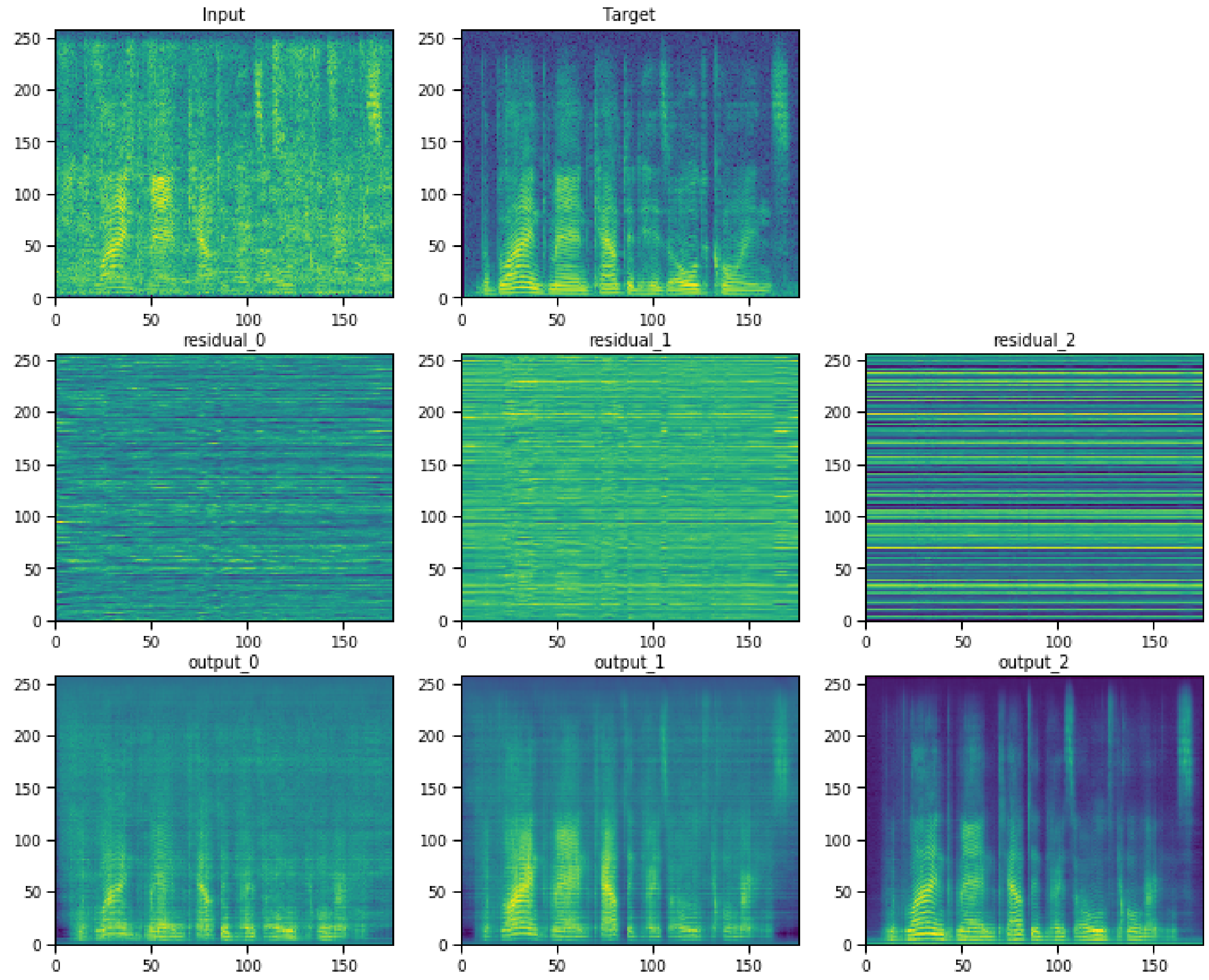


Figure 4.2 – Outputs of the residual model

model is increasingly improving the predicted speech structure in all layers, but the signs of many frequency bands might be flipped before we reach the last layer. This model is the most unrelated to both spectral subtraction and masking, since masking in the log-amplitude domain is equivalent to performing an exponentiation by the masks in the linear amplitude domain, which explains why the plots are harder to interpret than for the other models.

Figure 4.3 shows a similar plot for highway models, with the second and third rows representing the outputs of $\boldsymbol{H(x)}$ and $\boldsymbol{T(x)}$, respectively. The interpretation of this model is more complicated since the output is a linear combination of the elementwise products of $\boldsymbol{T(x)}$ and $1-\boldsymbol{T(x)}$ by $H(x)$ and x, respectively. We can interpret $T(x)$ in these plots as showing us which elements of $H(x)$ can we trust more than the current input, with those being represented by green and yellow (stronger confidence), while time-frequency cells marked in dark blue will be dominated by the input of the

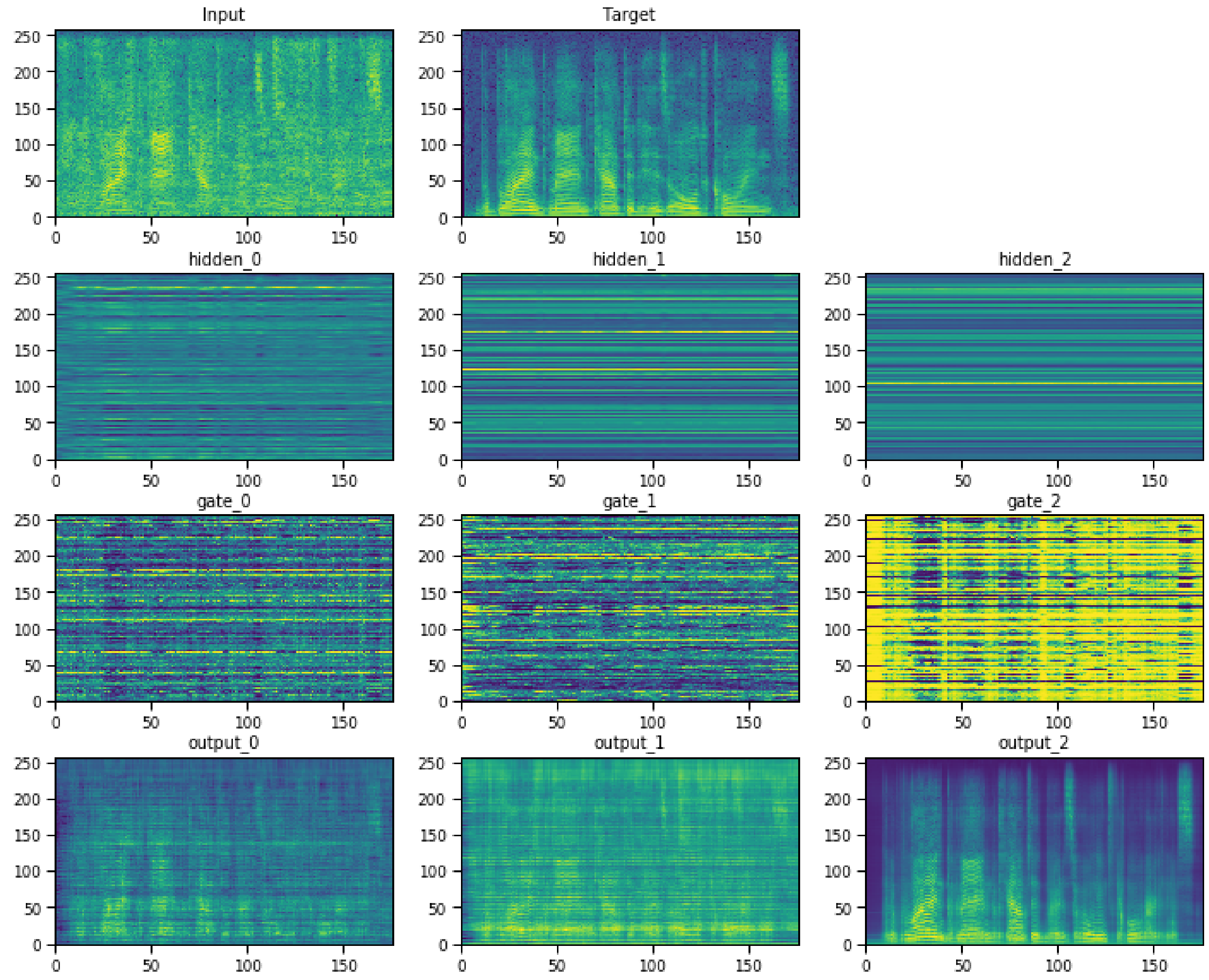


Figure 4.3 – Outputs of the highway model

current block. We can see the first block selects regions where the speech signal is stronger than the noise as having a low value for $T(x)$, and decides to reject the input and use its own internal estimate for a number of bands. In the last layer, though, the model uses its internal estimate for a large portion of the signal, as evidenced by the predominance of yellow in the last output for $T(x)$. This seems counter-intuitive, especially as we notice that the outputs of the second block seem to have a larger amount of noise than those of the first block.

Although all results reported here take into account a single example for the denoising tests, these characteristics are common to different sentences and distortion types. The same observations are also valid for the reverberation models, although these use a different set of models so specific frequency ranges might not be the same but the observations in general still hold.

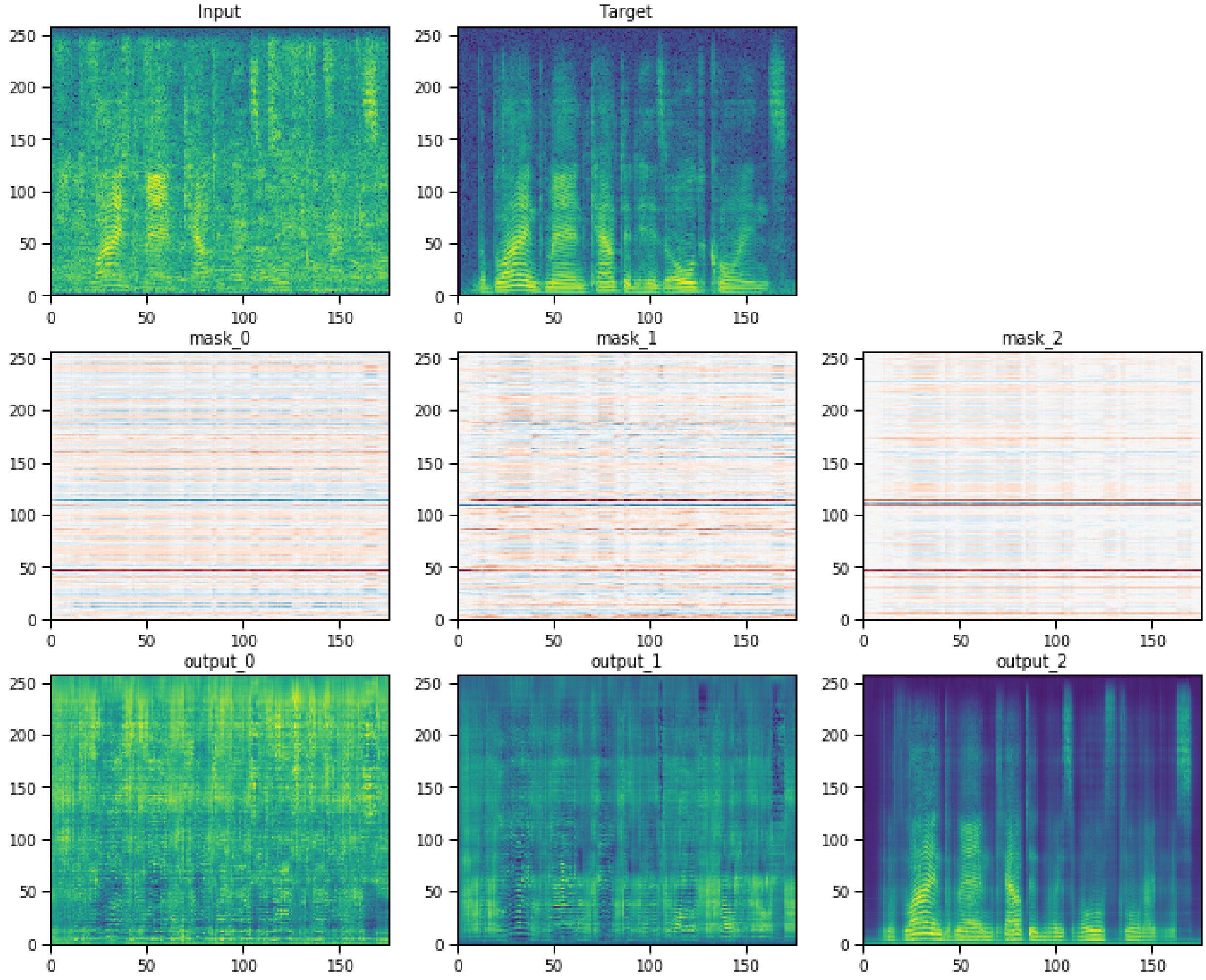


Figure 4.4 – Outputs of the masking model

Objective quality and intelligibility assessment

We also evaluated the models using the objective speech quality and intelligibility metrics PESQ and STOI [111]. For both metrics, higher scores are better. As can be seen in figures 4.5-4.8, for most of the models there is not a large improvement over the baseline. We can argue that a positive aspect of the use of residual or masking models is that they are more interpretable than the baseline. It should also be noted that direct comparison between these models is not completely fair since the models tested have different numbers of parameters, with the baseline model having the least parameters, followed by masking, residual, and highway, and models with 3 blocks have more parameters than their single block counterparts.

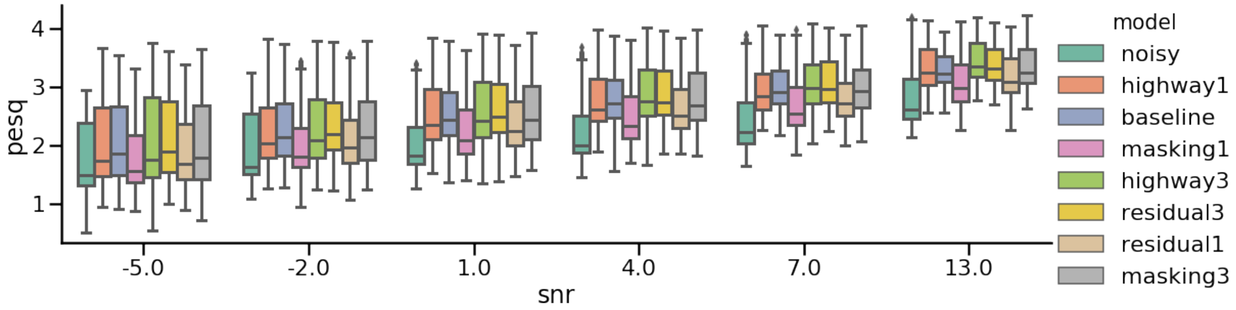


Figure 4.5 – Denoising models - PESQ

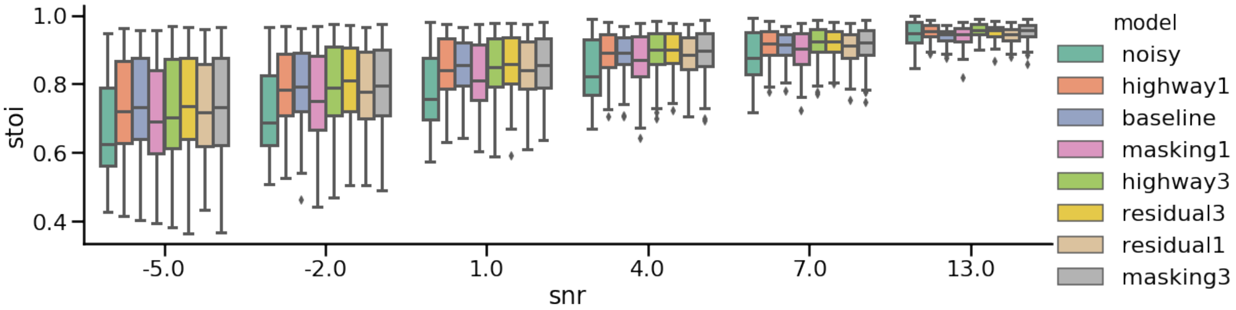


Figure 4.6 – Denoising models - STOI

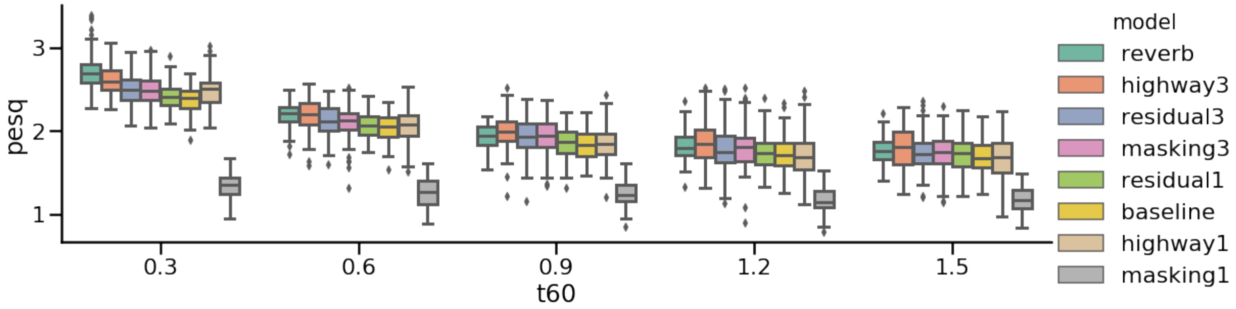


Figure 4.7 – Dereverb models - PESQ

We also note that these models are not competitive with state-of-the-art models presented more recently in the literature. In this thesis, we looked into these models as building blocks for more advanced models that achieve higher performance in these tasks, such as those previously cited.

4.3.5 Discussion

In this thesis we suggest that having skip connections makes models more interpretable. However, there are other alternatives in the literature we would like to bring up and compare to our approach.

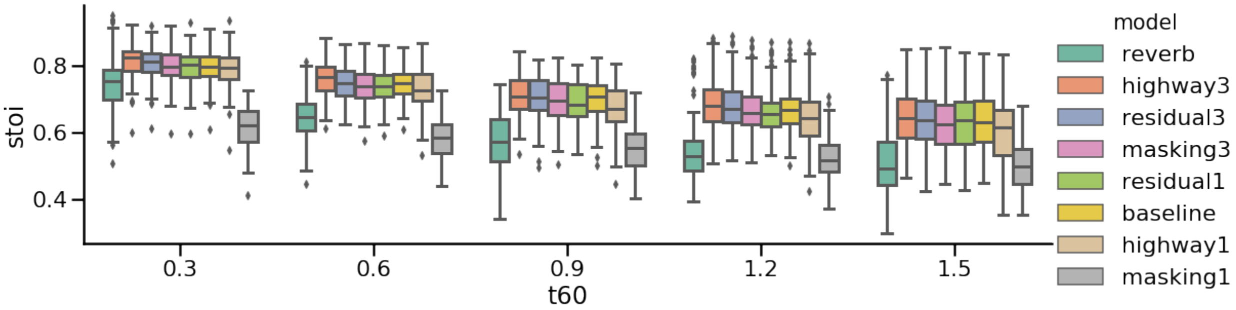


Figure 4.8 – Dereverb models - STOI

Although these were not developed with interpretability in mind, they are based in signal processing ideas, which makes them more interpretable than a DNN not using any domain knowledge.

The work by Sainath et al. in [112] proposes using a learned speech front-end with time-domain convolution layers that use much a longer kernel than usual for CNNs. The learned filterbank then replaces arbitrary filterbanks such as gammatone or mel filterbanks. Since the filters are learned jointly with a speech recognition model, one would expect those filters to achieve higher performance than arbitrary filters. However, the paper shows their performance matches the performance of a log-mel filterbank. Combining both representations, on the other hand, led to an improvement of 3% in word error rate (WER).

Followup work in [113, 77], however, used a similar strategy to learn beamformer filters for a multichannel speech recognition system. Instead of learning a filterbank for a single channel signal, in this work an FIR spatial filterbank (similar to a filter-and-sum beamformer) is learned during training. For each channel, a bank of P filters is learned, so the model can use different steering delays for each filter. The model then combines the outputs of all filters for all channels using a pooling layer and a non-linearity. This approach has led to an improvement of approximately 5% in WER compared to a model using only a log-mel filterbank for all channels and performs as well as a model given oracle knowledge of the true location of the source.

In [114], the authors propose a new DNN architecture called SincNet, which is based on learning parameterized sinc functions instead of learning all elements of a convolutional layer. SincNet models are more efficient to train since only two variables are trainable for each filter (low and high cutoff frequencies), and the authors show that SincNet converges faster and performs better in speaker identification and verification tasks than a standard CNN on raw waveforms.

A common element among all the abovementioned works is the learning of more conventional time domain filters that can be easily analysed using the mathematical tooling for linear systems (even though the entire models are not linear). This allows us to better understand what is represented in each of the features the model is learning. Standard CNNs usually employ a large number of small filters with non-linearities between the layers, which makes them much harder to interpret. As an example of models that perform operations in the time-domain with more conventional CNNs, we can cite Wavenet [67] and SEGAN [71]. Both employ convolutions in the time domain and skip connections, but interpreting their internal variables is much harder than for the abovementioned models or the models proposed in this section.

4.4 Pruning LSTM and GRU-based models through eliminating "stuck" and highly-correlated features

DNN-based speech processing methods are capable of performing several tasks at better performance than their statistical counterparts. However, many such methods come at a cost of more computational requirements, such as the need for more operations per second to be able to operate at real-time, or memory usage, since models have usually a large number of parameters. In order to be able to use these models in embedded applications or just for efficiency reasons, we would like to use the smallest possible models we can. There are many research efforts in that direction, going from the developments of methods for model compression and pruning to the development of more efficient architectures. The main idea of model compression and pruning is to achieve models with smaller number of parameters or storage requirement for such parameters. One can, for example, use quantization methods to store floating point parameters in more efficient formats [115], or distilling the knowledge of larger models into smaller, more efficient models [116].

Pruning methods work by identifying irrelevant parts of a model and removing them from the model, ideally without an impact (or with a minor impact) on performance. Pruning can be done either during training (ad hoc) or after training (post hoc), with the latter having the advantage of being useful even for models already in production. Ad hoc pruning has the advantage of allowing one to use sparsity-inducing regularizers to encourage the model to be sparse, while post hoc methods do not have any guarantees that the model will have parameters to be pruned.

We can classify pruning methods into two categories regarding what kind of information is removed from the model: weight pruning aims at making the weights of a model sparse, where neuron pruning aims at removing entire neurons from the model. While both methods reduce the storage needs for a model, sparse models require special linear algebra implementations (either in software or hardware) in order to be efficient, while neuron pruning keeps models dense and does not require any special operations. For both methods, a commonly used feature importance measure is the magnitude of the weights: for sparsifying a model, weights close to zero are set to zero, while for neuron pruning, neurons with all weights close to zero or activations close to zero are removed. Works such as optimal brain damage [117] propose more complex feature importance functions based on second-order methods; however, such methods are too computationally intense to be used during training, which led the community to continue using magnitudes.

Most feedforward and convolutional models use rectified linear units (ReLUs), which only saturate at 0, or other activation functions that do not saturate (e.g. leaky/parameterized ReLUs, exponential linear units). Most recent works on pruning aim at removing neurons whose activations have low magnitude, likely because saturation is not an issue in many of the models currently being used. However, recurrent models (such as models using LSTM and GRU layers) still use saturating activation functions.

Through visualization of the internal features of the model presented in the previous chapter, we observed that many features were saturated most of the time (we call such neurons "stuck" henceforth), seeming to be redundant to the task. We then applied several methods for model regularization and/or pruning, but these methods failed to cause the model to not have a large number of stuck neurons. Therefore, in this section we propose a novel method for neuron pruning in deep neural networks that is complementary to magnitude-based neuron pruning, but aimed at models using functions that saturate at non-zero values (such as models employing LSTM and GRU layers). We show that the proposed method is effective in pruning such neurons without significant changes in the model performance, allowing us to compress a GRU layer to 48% of its original size in the evaluated model.

4.4.1 Methods

Observing and understanding recurrent layer activations

As mentioned in chapter 2, practical recurrent networks usually employ gated connections to avoid gradient issues while training. Two popular architectures for gated neural networks are long short-term memory units [53] and gated recurrent units [54]. The main difference between both architectures is that LSTMs have four gates, namely the input i_t, output o_t, cell g_t, and forget f_t gates, and a memory cell c_t. The input and cell gates control how much information enters the cell, the cell gate controls how the output gate how much information leaves the cell through the hidden state, and the forget gate decides when the cell should reset its state. The operations performed at each timestep t are described by the equations below, where $*$ is the Hadamard product:

$$i_t = \sigma(W_{ii}x_t + b_{ii} + W_{hi}h_{(t-1)} + b_{hi}) \tag{4.6}$$

$$f_t = \sigma(W_{if}x_t + b_{if} + W_{hf}h_{(t-1)} + b_{hf}) \tag{4.7}$$

$$g_t = \tanh(W_{ig}x_t + b_{ig} + W_{hg}h_{(t-1)} + b_{hg}) \tag{4.8}$$

$$o_t = \sigma(W_{io}x_t + b_{io} + W_{ho}h_{(t-1)} + b_{ho}) \tag{4.9}$$

$$c_t = f_t * c_{(t-1)} + i_t * g_t \tag{4.10}$$

$$h_t = o_t * \tanh(c_t) \tag{4.11}$$

Gated recurrent units are a simplified variant of LSTMs that do not have an output gate. The gates in a GRU are called the reset gate r_t, update gate z_t, and new gate n_t. The operations performed at each timestep are described by the operations below:

$$r_t = \sigma(W_{ir}x_t + b_{ir} + W_{hr}h_{(t-1)} + b_{hr}) \tag{4.12}$$

$$z_t = \sigma(W_{iz}x_t + b_{iz} + W_{hz}h_{(t-1)} + b_{hz}) \tag{4.13}$$

$$n_t = \tanh(W_{in}x_t + b_{in} + r_t * (W_{hn}h_{(t-1)} + b_{hn})) \tag{4.14}$$

$$h_t = (1 - z_t) * n_t + z_t * h_{(t-1)} \tag{4.15}$$

We should note that both recurrent layers employ both sigmoid and hyperbolic tangent activation functions. These functions were used in feedforward and convolutional networks initially, but

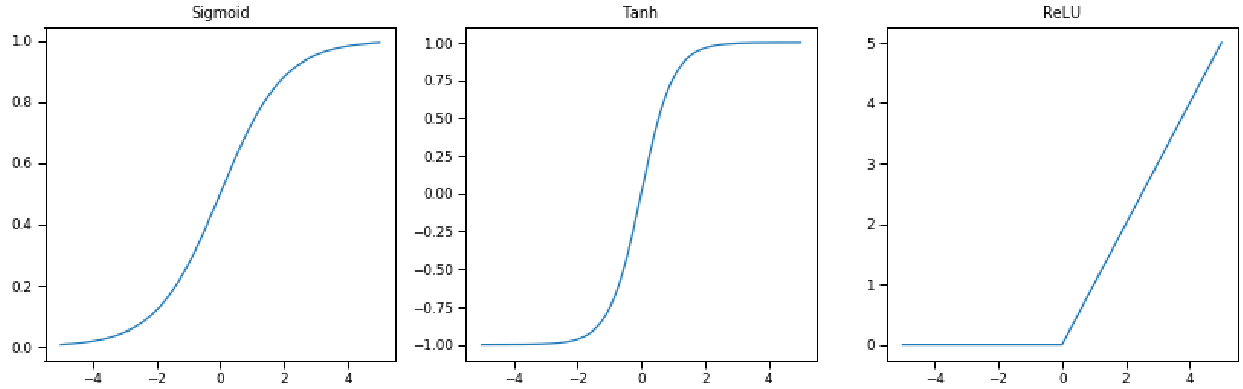


Figure 4.9 – Common activation functions in deep neural networks

replaced by simpler functions such as ReLUs and their variants as they are more computationally effective. These activation functions are computed as below:

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{4.16}$$

$$\tanh(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \tag{4.17}$$

$$\mathrm{ReLU}(x) = \max(0, x) \tag{4.18}$$

Plots of all three activation functions between the interval $x \in [-5, 5]$ are shown in figure 4.9.

In LSTMs and GRUs, however, it is important to use functions that saturate smoothly to -1 and 1 as these help mitigate the exploding gradient issue (as they bound the magnitude of all activations inside the layer). Combined with proper initialization methods, these activation functions lead to more numerically stable training of recurrent networks. Note that there are no guarantees that the weights and biases are not going to be too high, causing parts of a recurrent layer to be always in a saturated state.

Pruning "stuck" features

We tried a number of methods to eliminate such redundant/stuck features from the model. First, since those features seemed to not add any relevant information, we hypothesized training a model with reduced number of units could lead to less redundancy/irrelevant features, since the model would have to use whatever available capacity as best as possible. Second, we experimented with the ad hoc pruning strategy proposed in [118], which uses a binary mask to zero the smallest magnitude weights of the model incrementally over training, until a certain sparsity level is achieved. Finally, we used a dropout [119] strategy, which reduces model capacity during training time by zeroing certain hidden units (therefore simulating the effect of neuron pruning), forcing the rest of the model to learn a robust representation, but without reducing model capacity during inference.

As will be seen in the experiments presented in the next section, these methods have proven ineffective to solve the issue of stuck features. Therefore, we also propose here a novel post hoc method for neuron pruning based on a simple heuristic for detecting and pruning "stuck" neurons and replacing the need for such neurons by adjusting the biases of the layers previously connected to such neurons to simulate a constant input. Our criteria for detecting a stuck neuron is the following:

$$a(W^k_{i,:}x + b^k_i) \approx c \tag{4.19}$$

where $W^k_{i,:}$ is the vector corresponding to the i-th neuron of the k-th layer, a a saturating activation function used in the k-th layer, x the input to the layer, and c is one of the non-zero constants to which a can saturate (1 for sigmoids, -1 or 1 for hyperbolic tangents). Note that this excludes the case for magnitude pruning, for which $c \approx 0$. Note also that this could be done for any other values of c, but those are more unlikely than the values at which a saturates.

When we observe that equation 4.19 is valid for a large majority of x values in the data, we assume $W^k_{i,:}$ can be pruned. In practice, that can be computed from a subset of samples from the training set. In the experiments reported later in this section, we used $\text{median}(a(W^k_{i,:}x + b^k_i)) = c$ as the criteria to decide whether a neuron should be pruned. When that neuron is removed, unlike when we prune a neuron with low magnitude, we have to compensate for the loss of that neuron's input in the remainder of the network. For stacked LSTMs, GRUs, and feedforward linear layers, that means adjusting all affine transforms that depend on the pruned neuron as follows:

$$\delta_b = W^{k+1}_{:,i}c \tag{4.20}$$

$$b_{k+1} := b_{k+1} + \delta_b \tag{4.21}$$

and then dropping column W_i from W^{k+1}. Since layer k is recurrent, it also depends on its own hidden state, so all weight matrices with it as an input (denoted as $W_{h\square}$, where $\square$ can refer to any of the gates that depend on h) have to be updated as well. The algorithm is summarized below.

Algorithm 1 Algorithm for stuck neuron pruning

```
for all layers $k$ do
  for all neurons $i$ do
    for all saturation values $c$ do
      if $a(W^k_{i,:} + b^k) \approx c$ then
        Drop row $i$ from $W^k$ and $b^k$
        $\delta_b := W^{k+1}_{:,i} c$
        $b^{k+1} := b^{k+1} + \delta_b$
        Drop column $i$ from $W^{k+1}$
      end if
    end for
  end for
end for
```

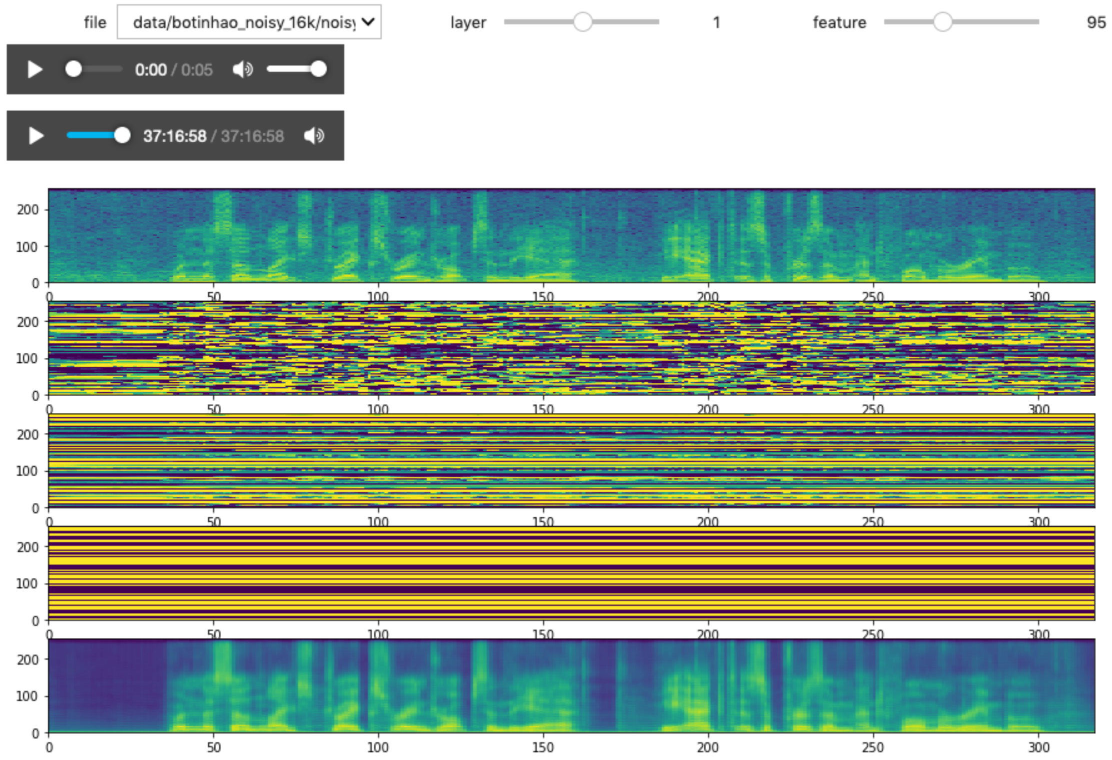
4.4.2 Experiments

Visualization

In order to better understand the behaviour of the internal workings of speech enhancement models we observe the hidden activations of the model proposed in Chapter 3. Here, instead of observing them through projecting its skip connections through the output layer to see the representation in the STFT domain, as done in the previous section, we investigate the learned representation directly.

Experiments were performed on a Jupyter notebook [120], using interactive widgets to allow us to quickly make changes to the model and observe results. This allowed for much faster visualization and experimentation. Our interactive interface supports four different actions. The first is a simple visualization for hidden activations in the model, while the other three are ablation experiments where a single feature or layer of the model is perturbed in one of three different ways: by injecting noise to all activations of a given layer, by clamping a feature to zero, and by clamping a feature to its median. We expected that by being able to visualize the effect of such ablations and also visualize hidden activations alongside the inputs and outputs of the model would provide some insights to how the model works and potential improvements. Figure 4.10 has a screenshot of the interface for one of the actions.

All visualizations in this section were performed with variations of the context-aware model presented in Chapter 3 trained for a denoising task with the dataset from [121]. The dataset



file data/botinhao_noisy_16k/noisy

layer 1

feature 95

Figure 4.10 – Interface of the tool for visualizing and changing hidden representations

consists of clean speech from the Voice Cloning Toolkit (VCTK) dataset [122] corrupted by noise from the DEMAND [123] noise corpus. Two speakers, one male and one female, were left out for testing, and the SNRs and noise types on the testing set do not match those of the training set.

Pruning

We tested multiple approaches to try to mitigate the stuck unit issue in the model. First, since the original model in [76] uses GRUs without biases for the input-to-hidden and hidden-to-hidden connections, we experimented with an alternative model with those bias vectors. We hypothesised that the stuck features were playing the role of fixed biases, and having learnable biases would alleviate the issue.

As a second hypothesis, we assumed the presence of a large number of stuck units means the model is overparameterized, so we trained a smaller version of the model where all layers are 50% of the original model size (i.e. 128 units/layer instead of 256). We also experimented with using 30% dropout during training and the weight sparsity method proposed in [118] (with a sparsity target of

0.85, meaning approximately 85% of the model parameters will be equal to zero). Dropout reduces the effective model capacity during each iteration of training but does not reduce it during inference time, while weight sparsity leads to an effective reduction in model capacity.

For the proposed method, we used the activations from 100 random features from the training set to select which neurons could be pruned. We only looked at the activations for hidden units and not for gates, since by removing the dependency on a hidden unit we can also remove coefficients from all the weight matrices and bias vectors that depend on the hidden state.

4.4.3 Results

Insights from visualizations

We used the proposed tools to experiment with different models and visualize the effects of the different types of ablations. One issue we immediately identified with all trained models was the large number of hidden units that seemed to be "stuck". While some of those units still had small variations, many were simply not changing despite changes in the input signal. Figure 4.11 shows one such visualization for the baseline model presented in chapter 3 and trained as described previously in this section. We can notice the large number of solid yellow and blue lines in the activations for h_2 and slighly less for h_3. h_3 clearly has some units that do not change while the input is silent and fewer units that do not change despite the input.

In order to better visualize how many features are stuck and/or highly correlated to each other, we also computed and plotted correlation matrices for each layer as heatmaps. Figure 4.12 shows the correlation matrices for the baseline model. Bright yellow or dark blue regions outside of the diagonal line indicate units whose output is highly correlated with other units. The majority of highly correlated units are in the second GRU layer. Upon observation and cross-verification with the stuck units in the unit activation maps, we noticed the majority of these correspond to stuck features. Using the criteria from equation 4.19, a total of 134 and 20 units were stuck in the second and third layers of the baseline model.

We also analyzed the correlation matrices for the biased and small versions of the baseline model, shown in figures 4.13 and 4.14, respectively. Both still presented a high number of stuck units, giving no evidence for the hypothesis that the model is using stuck units as biases or is overparameterized.

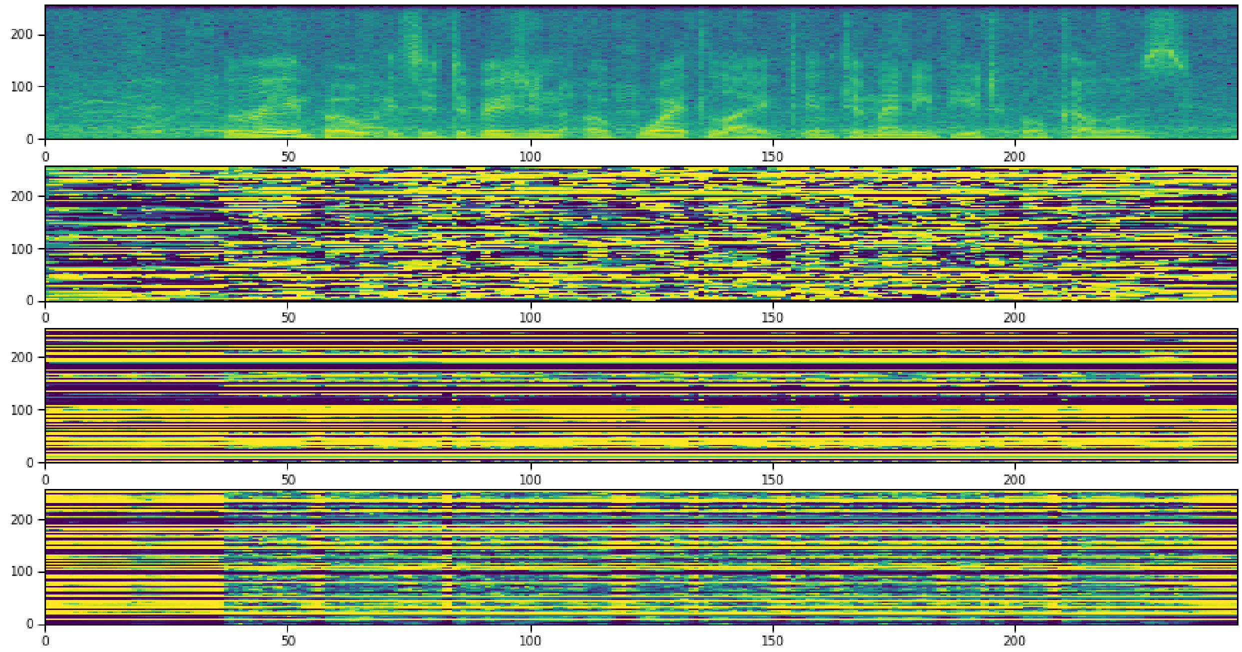


Figure 4.11 – Visualization of hidden unit activations for a noisy input. Top: spectrogram of input signal. Next three images show the activations for h_1, h_2, and h_3, respectively.

The biased model has more stuck features than the unbiased model (136 and 24 for the second and third GRUs), with a similar trend being observed on the small model as well (69 and 18 units, respectively).

Experiments with pruning and dropout

To evaluate the models trained using dropout and weight pruning methods, we first analyze their correlation matrices (figures 4.15 and 4.16). Dropout failed to reduce the amount of correlated and stuck features, with the model having a total of 75 stuck units on the second GRU layer and all features being stuck on the third layer. While this does not affect the performance of the model trained with weight pruning (as shown in the next section), it severely decreases the performance of the model trained with dropout, since it is basically not using any of the capacity in its third layer. The model trained with a weight pruning strategy has 146 stuck units on the second GRU layer and 5 stuck units on the third GRU layer.

By applying our pruning method on the second GRU only, we achieve a reduction of 52.9% in the size of the second GRU layer. The third GRU layer can be reduced by 10.5%. Although this

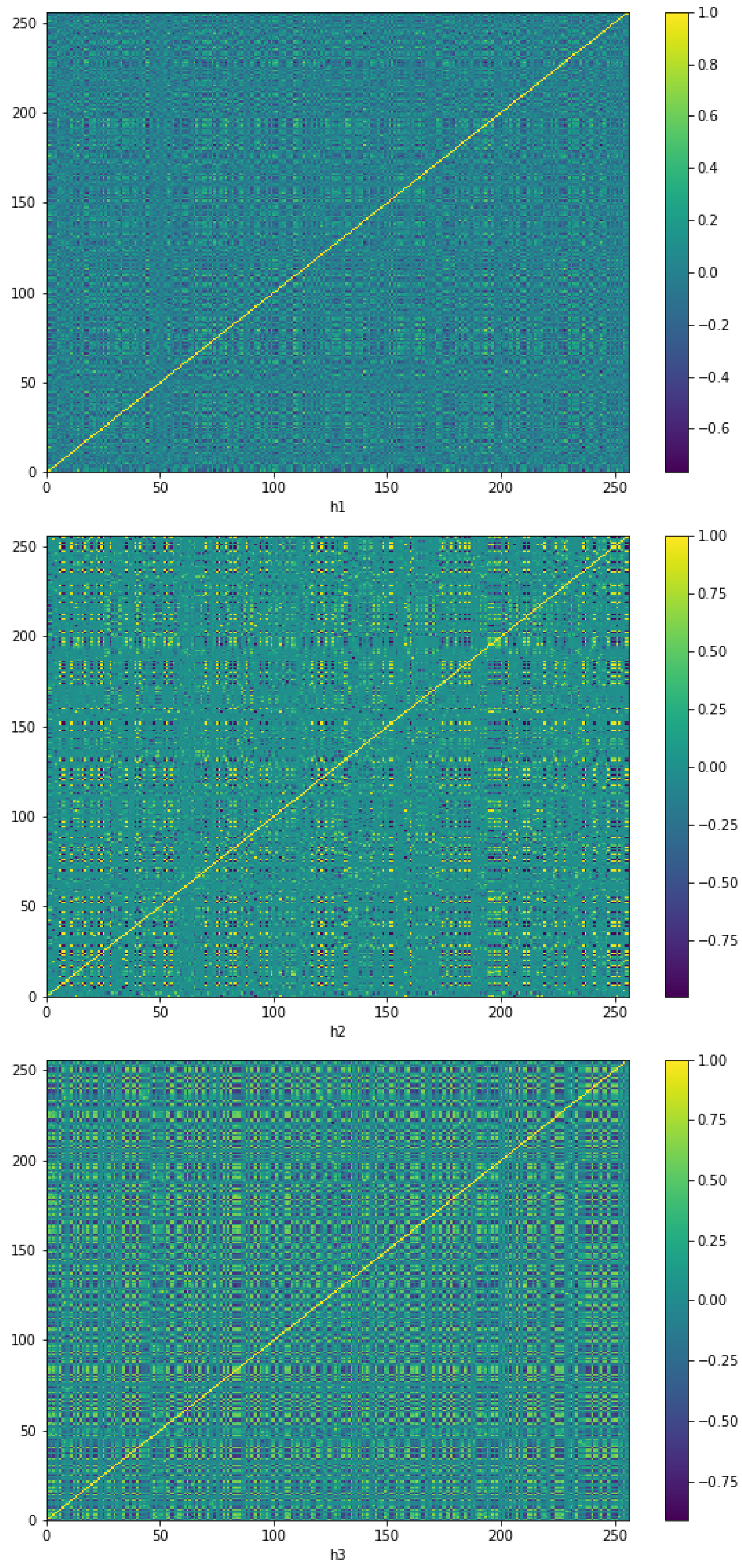


Figure 4.12 – Correlation matrices for the baseline model

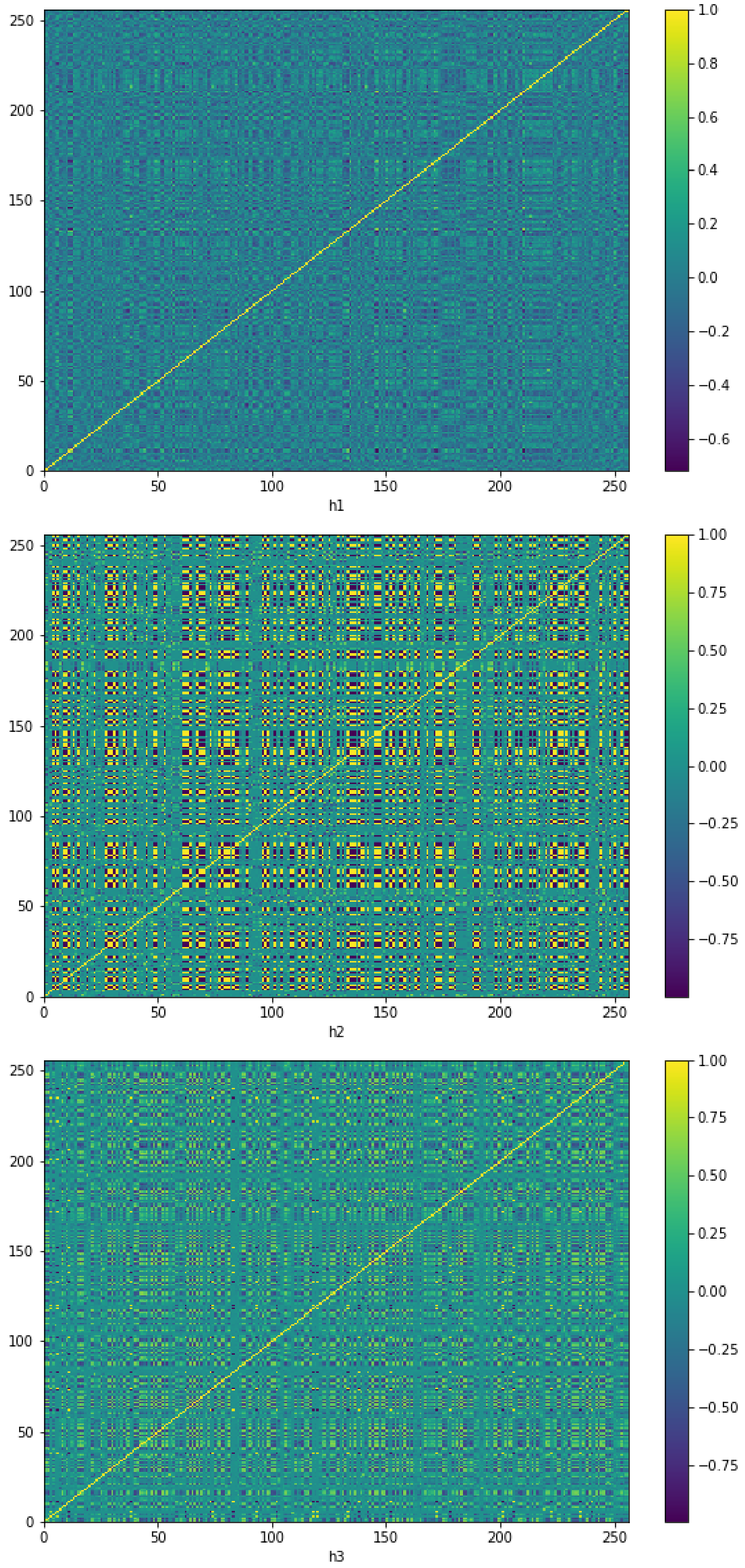


Figure 4.13 – Correlation matrices for the biased GRU model

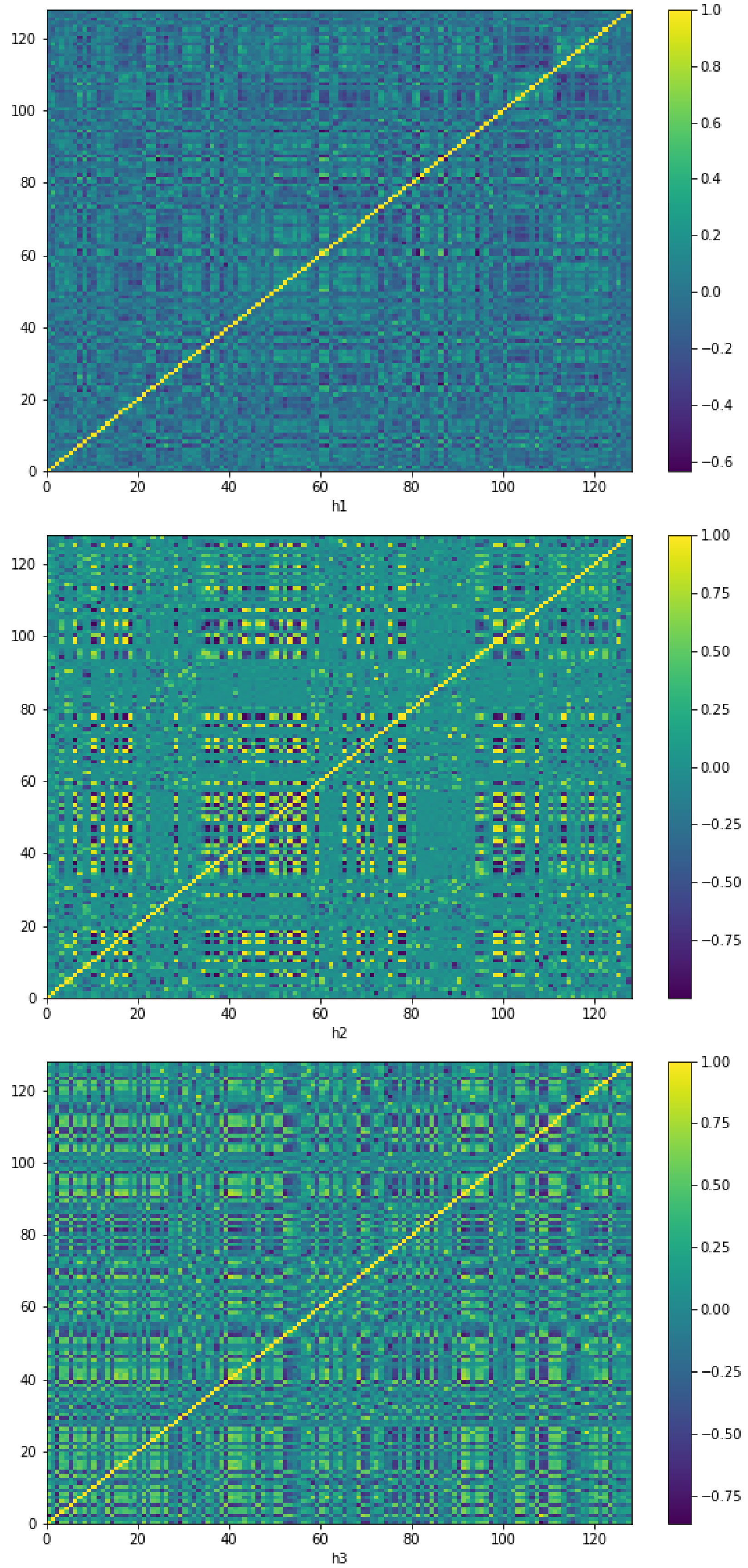


Figure 4.14 – Correlation matrices for the small (128 hidden units/layer) model

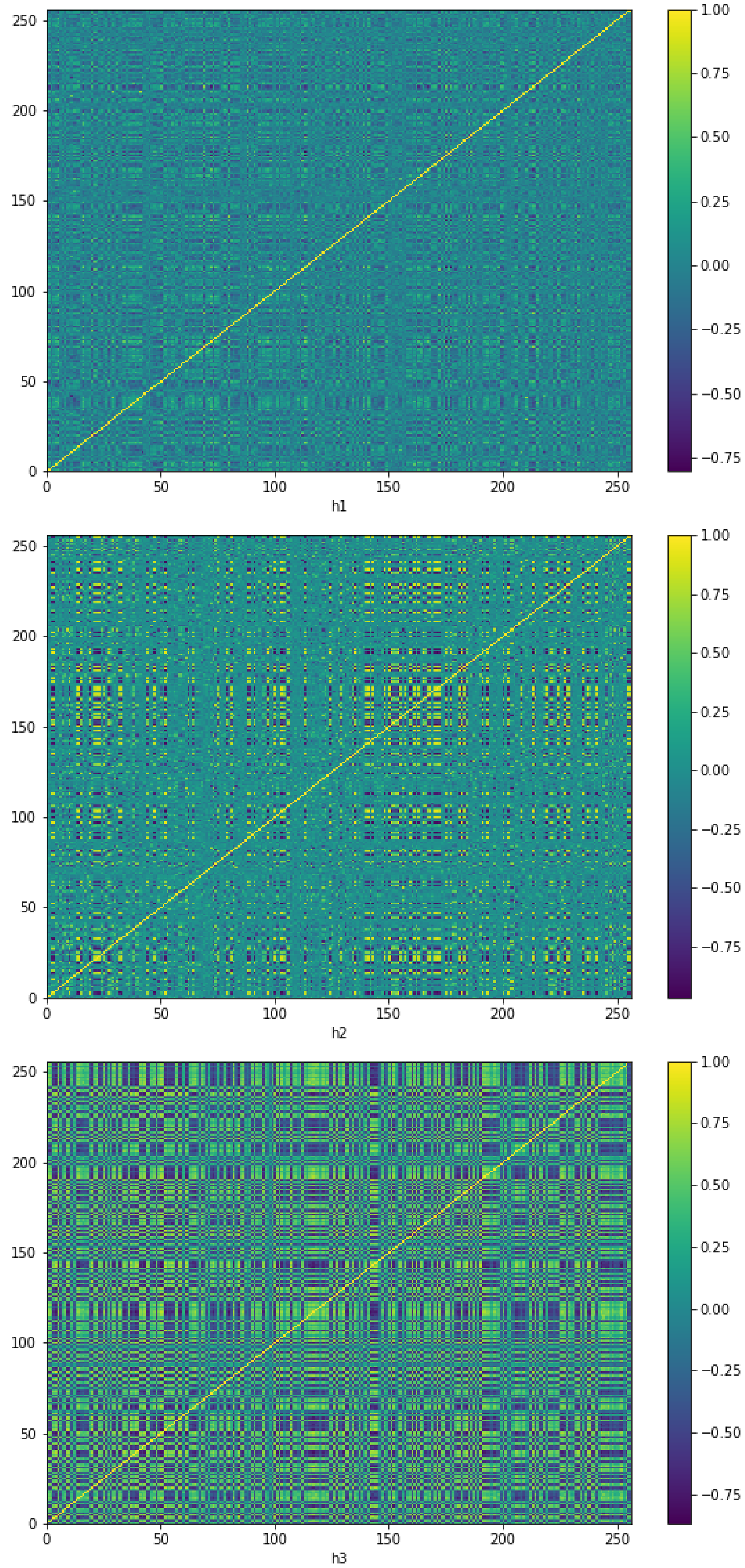


Figure 4.15 – Correlation matrices for the model trained using 30% dropout

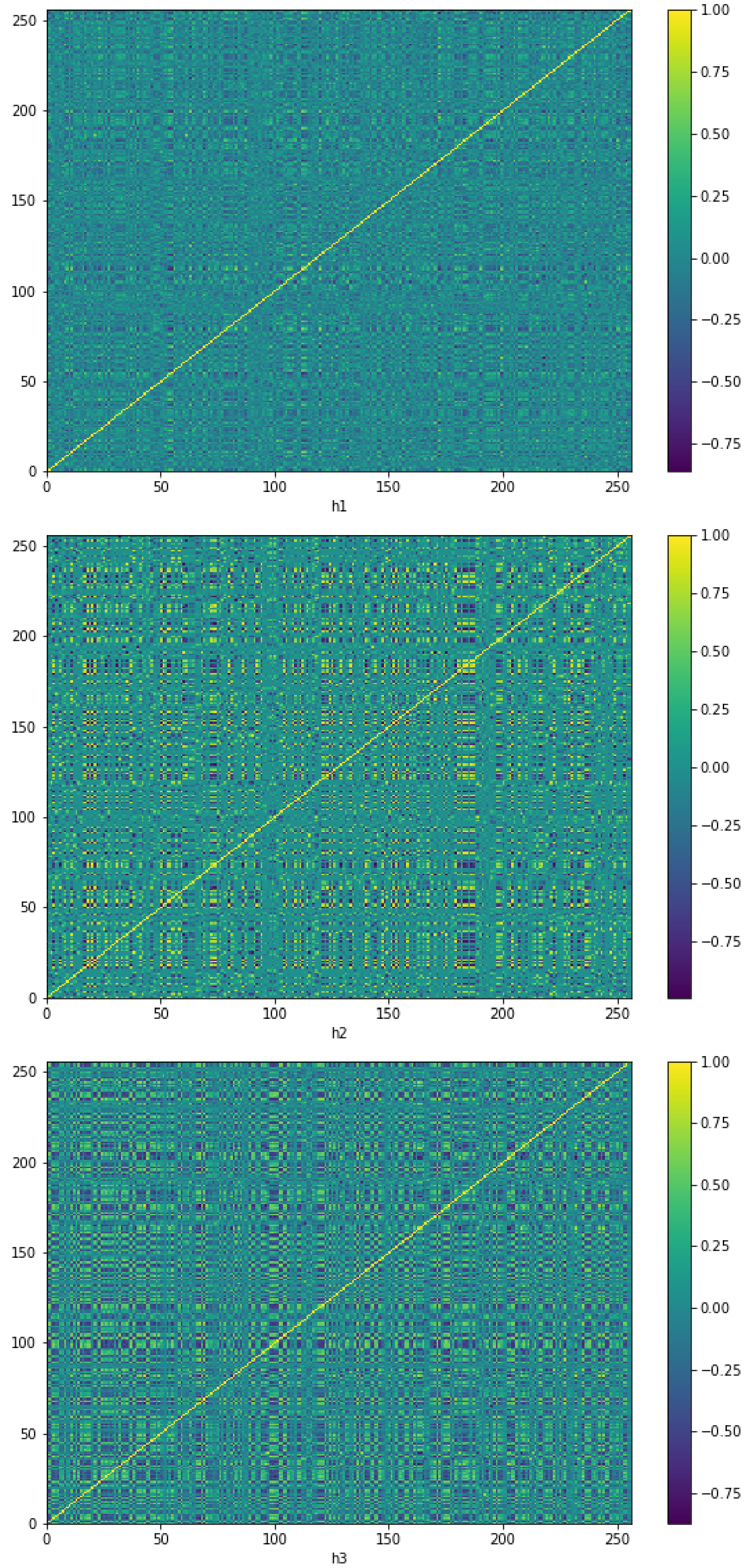


Figure 4.16 – Correlation matrices for the model trained using weight pruning

represents a small reduction in model size overall (3.3%), since many of the parameters of the model are concentrated in the first GRU layer, it should be noted that this has no impact whatsoever on the model training and also leads to a reduction in computation requirements. The weight pruning method achieves much higher compression because it also affects the first GRU (which has 6047232 parameters, of which 5138832 are equal to zero when using weight pruning). Both the second and the third GRUs have 394752 parameters, and training with weight pruning makes 334224 of those parameters equal to zero. Additional improvements in compression can be achieved by combining the proposed neuron pruning strategy to this weight pruning strategy, leading to a total of 92.9% weights equal to zero in the second GRU layer and an additional 0.03% zeroed weights on the third GRU layer.

To assess the impact of the different strategies, we computed two objective quality metrics (PESQ and SRMR) and one objective intelligibility metric (STOI) for all files in the test set. Results are reported in figures 4.17-4.19. As mentioned previously, both the smaller version of the model and the model trained with dropout showed a reduction in performance for all three metrics. Weight pruning and the proposed pruning method, when applied to the biased model, have similar performances to the baseline and biased models, with the weight pruning approach having a slightly lower performance than the proposed pruning method (which is expected, given the overall smaller effective model capacity). Combining the proposed method to weight pruning incurs no additional penalty to performance.

4.4.4 Discussion

Analysis of correlated features brought insights we were able to use to propose a new pruning method for recurrent models based on removing stuck neurons. We could not find any other approach in the literature that aimed at removing neurons whose output is constant because they are clamped to a non-zero saturation value of their activation function. We believe this to be the case because, as discussed previously, for computational efficiency reasons many recent models use simpler activation functions that do not saturate at values other than zero. Activations such as the sigmoid and hyperbolic tangent are still present in models with some type of gating, such as GRUs, LSTMs, and highway networks [105].

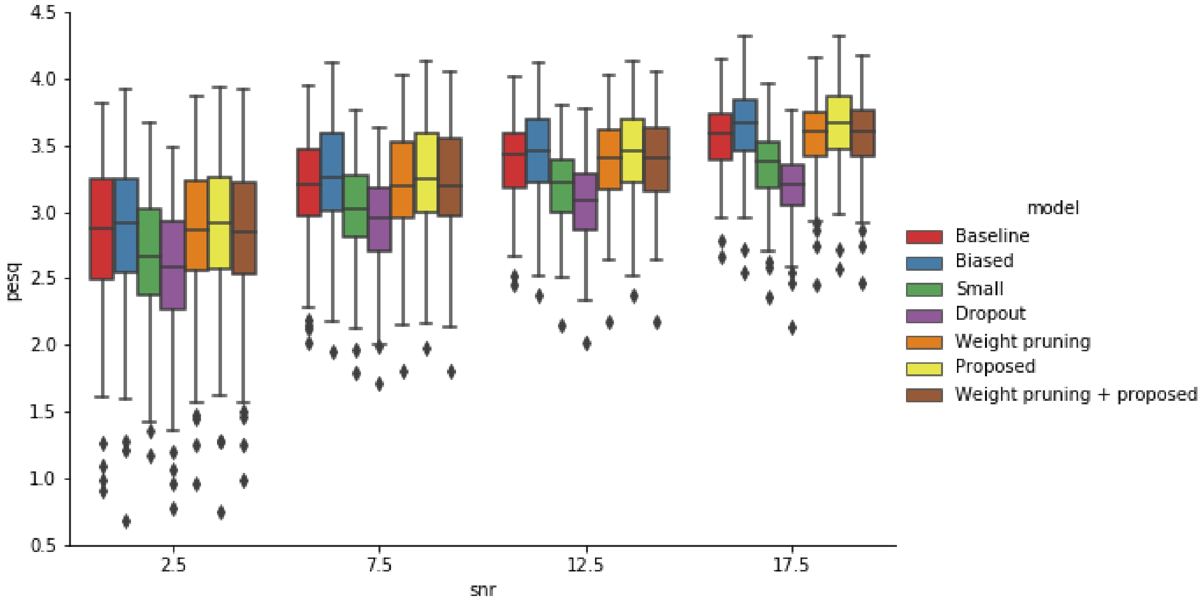


Figure 4.17 – PESQ distributions for all models, pruning strategies, and SNRs in the test set

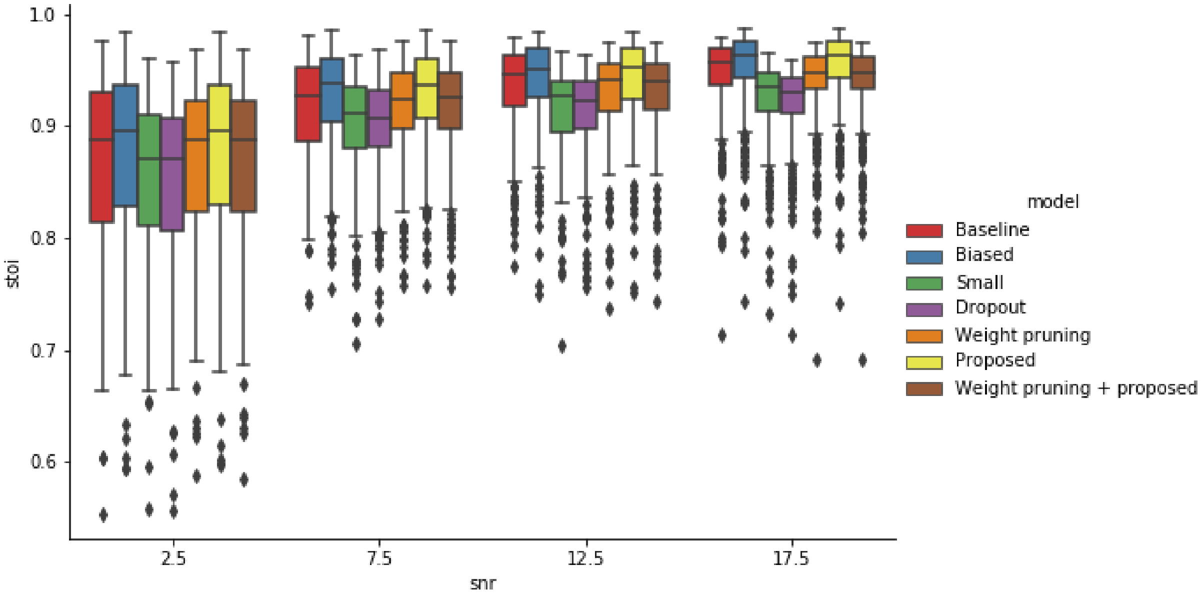


Figure 4.18 – STOI distributions for all models, pruning strategies, and SNRs in the test set

The work in [124] focuses specifically in reducing the computational cost of LSTMs. Their method is related to magnitude-based neuron pruning, but instead of pruning the neuron, they only prune activations with magnitude under a certain threshold T. This is performed during training. They observe that, in the evaluated tasks (image classification and word-level language modeling) 90% of the hidden states can be pruned without any accuracy degradation. This is in line to what

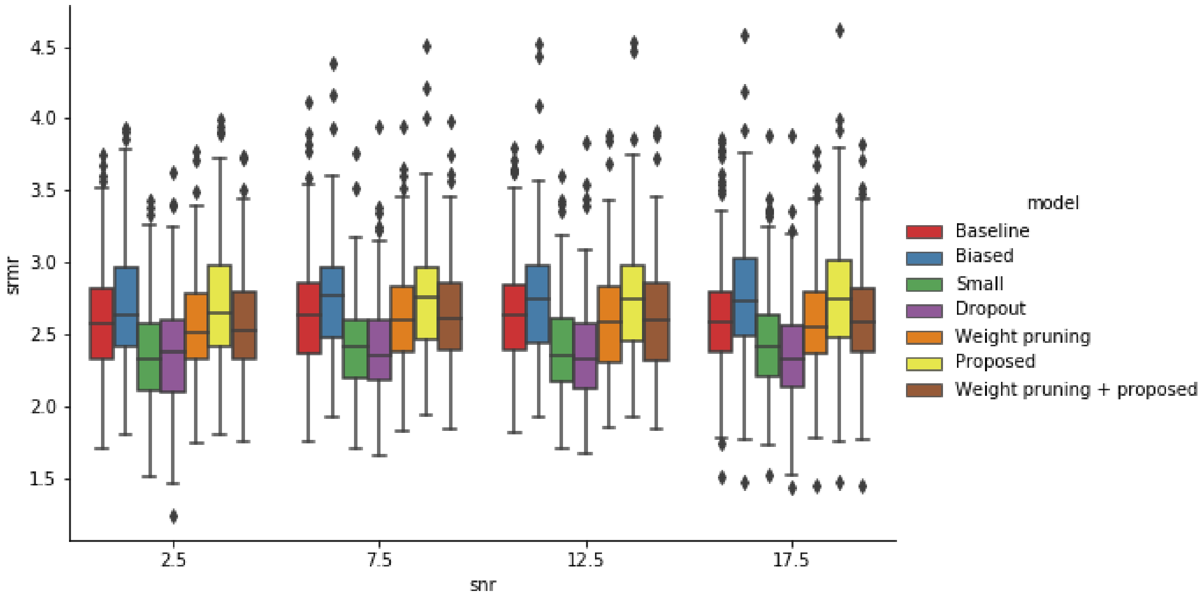


Figure 4.19 – SRMR distributions for all models, pruning strategies, and SNRs in the test set

we observed in this paper for GRUs. Their models are simpler (LSTMs with a single layer) and the tasks are different, and this might account for the difference in ratio of hidden states that can be pruned. The authors also propose a hardware accelerator to take advantage of sparse hidden vectors.

Another work that applies magnitude-based weight pruning to LSTMs is [125], where the authors use the method on neural machine translation models. The authors use the magnitude-based pruning scheme originally proposed in [126], but compare three different schemes for pruning based on how important they consider weights to be to the model, but their experiments show a class blind approach to perform best. They show that a neural machine translation model can be pruned by 40% with very little performance loss.

In [127], the authors propose a new method based on dropout, which they call targeted dropout, that promotes sparsity by dropping the least useful units during training (ad hoc), so that they can be safely pruned after training. This is opposed to standard dropout, where units are dropped stochastically. The criteria for dropping a unit is still magnitude-based, and they are only dropping units with activations close to zero.

Smallify [128] is another ad hoc pruning method, where network size and accuracy are simultaneously optimized during training. Training starts with an initially over-sized network, where each neuron is connected to an "on/off" switch (essentially, a gating function). During training, the model is trained both to achieve good performance on the task (by optimizing an appropriate cost function) and to minimize the number of switches turned "on". Since minimizing the l_0 norm is an NP-hard problem, the authors relax the problem by allowing the gates to instead assume real numbers between 0 and 1. The authors tested the model only on CNNs and fully connected models, reporting network size reductions of up to 35× with a 6× speedup in inference time.

4.5 Conclusions

This work shows results of our investigation on the role of skip connections in speech enhancement models. Our experiments show that, although they have no significant impact in the performance of the models, such connections might help making the models more interpretable, as we can identify the contribution of each individual layer to the task. In the future, we intend to investigate more complex models, such as models based on the UNet architecture, as well as models that employ a temporal context window at the input instead of a single frame (such as the work in [76]), since those are more in line with state-of-the-art models in the literature.

We have also proposed a post hoc pruning method for recurrent models based on insights gained from observations of activation patterns in such layers. Our method alone achieves more than 50% of compression of a recurrent layer in the model we experimented with, and a total model compression of approximately 3% in a totally post hoc manner. This enables this method to be employed in any pretrained models that utilize GRUs and LSTMs to reduce both computational and storage requirements. Additionally, the proposed method is complementary with methods based on magnitude importance functions for neurons, and can be easily combined with them.

As future work, we intend to evaluate the visualization methods proposed here with other speech-related models, such as models for speech recognition, speaker identification/verification, and speech synthesis. We also would like to evaluate how useful the pruning strategy proposed in this chapter is for other models that employ RNNs, such as for other speech-related tasks as above, and also for natural language processing.

Chapter 5

Towards quality-aware DNN-based speech enhancement

5.1 Preamble

The content of this chapter has been submitted to the 11th International Conference on Quality of Multimedia Experience and the IEEE Signal Processing Letters.

5.2 Analyzing objective speech quality metrics for DNN-based speech enhancement

Recently, several works have focused on leveraging the advances in research on deep neural networks to a variety of areas, including speech enhancement. Although there have been solid advances to the state of the art in areas such as speech recognition [129, 130, 131], the advantages of using modern DNN architectures for speech enhancement have, in comparison, not been explored as much. Although recent work has shown significant improvements in several objective quality and intelligibility metrics [76, 101, 81], such methods have not yet been widely adopted.

One of the reasons is that, as recently reported in [132], objective metrics such as the short term objective intelligibility metric are not very good predictors of speech quality or intelligibility. In

[133], the authors show that DNNs trained on unmatched conditions (noise type, SNR, or speaker) sometimes improve and sometimes degrade speech intelligibility. Although a few studies have performed subjective quality tests with DNN-based speech enhancement systems [134, 135, 78], there is no in-depth investigation on the performance of speech quality metrics for such systems.

Speech quality is multi-dimensional and takes into account several aspects of the signal. How quality is assessed for a signal will usually depend on the type of distortions and expected enhancement as well. Such measurements usually also reflect the listening effort of a speech signal, while intelligibility is only concerned with whether a given word or sentence is intelligible or not. Objective quality metrics are often developed with specific kinds of distortion in mind (e.g. the robotization internal metric in the ITU-T standard P.563); however, speech enhanced using DNN-based methods, especially methods that perform spectral/signal estimation instead of masking, can present unexpected types of artifacts that were not envisioned when such metrics were developed. Additionally, many metrics are phase-insensitive, but human listeners consider phase distortion an element of speech quality [136].

The majority of DNN-based speech enhancement is trained through supervised learning, requiring both a clean reference signal as a target and a distorted signal as input for the model. While it is relatively easy to simulate distortions by adding recorded noise and convolving signals with room impulse responses to make them reverberant, this does not necessarily reflect how natural signals are created. Having access to accurate non-intrusive models of speech quality and intelligibility could enable one to perform semi-supervised (requiring reference signals for a subset of the dataset used for training) or unsupervised training (requiring no reference signals at all). There is some work in the literature using semi-supervised learning for speech enhancement/separation using non-negative matrix factorization, such as [137] (where the noise model is learned in an unsupervised way), but we did not find any research on semi-supervised or unsupervised learning applied DNN-based speech enhancement.

In this section, we explore the results of online listening tests to assess speech quality of several recently-proposed DNN-based speech denoising and dereverberation systems, and compare the subjective ratings of such experiments with several different objective speech quality and intelligibility metrics. We investigate whether such metrics are fit for assessing speech quality for DNN-based speech enhancement models, as well as propose alternative combined metrics based on the non-intrusive internal metrics from the ITU-T Recommendation P.563 [138].

5.2.1 DNN-based speech enhancement models

Most of the work on DNN-based speech enhancement models so far has focused on spectral estimation or spectral masking, even though there are models that work directly on time-domain signals such as [69, 71]. In this section, we focus on architectures that operate on the frequency domain.

Most models that operate on spectral representations of speech predict either the magnitude spectrum of the signal or an ideal mask, which are defined based on which time-frequency bins are dominated by noise, and then applied to the corrupted signal to obtain its enhanced counterpart [61]. In the case of masking, both binary masks and ratio masks have been proposed, with ratio being shown to achieve higher speech quality than binary masks.

The remainder of this section presents the three models we used for our speech quality evaluation.

Spectral estimation using feed-forward model

The simplest model we explored in this section is an implementation of the model proposed by [80], which consists of a feed-forward model with 3 hidden layers with 2048 hidden units each. The input for the model is in the log-magnitude domain, using an STFT with a window size of 32 ms and 50% overlap. Instead of using 7 frames for context, we used 11 frames as for the other two models explored in this work. Unlike in subsequent work by the same authors [81], we did not use a reverberation-aware approach to change the STFT parameters since recent work using that model did not show any improvements [76]. Mean-variance normalization of the targets was used as in [80, 81]. This model is referred to as Wu16 in the remainder of this paper.

Spectral masking using feed-forward model and arbitrary features

The model we used for spectral masking is inspired by the work of [101]. In that work, assuming Y as the STFT of the input signal y and decomposing it into its real and imaginary components so that $Y = Y_r + iY_i$, and defining a complex ratio mask $M = M_r + iM_i$, one can estimate the real and imaginary components of the target signal S as

$$S_r = M_r Y_r - M_i Y_i \tag{5.1}$$

$$S_i = M_r Y_i + M_i Y_r. \tag{5.2}$$

In our work, we used a slightly different definition for the complex mask, where the real and imaginary components of the mask only affect the real and imaginary components of the input, as follows:

$$S_r = M_r Y_r \tag{5.3}$$

$$S_i = M_i Y_i. \tag{5.4}$$

This simplifies the definition of the complex ratio mask components to:

$$M_r = \frac{S_r}{Y_r + \epsilon} \tag{5.5}$$

$$M_i = \frac{S_i}{Y_i + \epsilon}, \tag{5.6}$$

where ϵ is only used to avoid division by zero and in our implementation is defined as 10^{-9}. While this definition of the complex mask is not as general as the complex ideal ratio mask proposed in [101], the estimated masks still perform a significant amount of spectral enhancement, although the phase enhancement is reduced. We used the same set of complementary features as presented in [101] as input, with a context window of 11 frames. This model is referred to as CRM (complex ratio masking) in the remainder of this paper.

Spectral estimation using a context-aware recurrent model

The last model to be evaluated is the model proposed in Chapter 3, which is a recurrent model based on gated recurrent units with residual connections and uses a convolutional layer as a context encoder. The inputs of the model, as the one for [80, 81], are the STFT coefficients for the distorted signal using a window length of 32 ms with 50% overlap. The context encoder used 64 kernels of size (21, 11), where 21 is the number of frequency bins and 11 the number of frames taken into account for each computed feature, with stride 2 in the frequency axis and 1 in the time axis. The

GRUs, as in Chapter 3, had 256 units each, with corresponding projection layers for the residual connections. This model is referred to as Santos18 in the remainder of this chapter.

5.2.2 Objective Speech Quality Metrics

The most basic objective metrics that are correlated to speech quality are the time and frequency domain segmental signal-to-noise ratios (segSNR) [139], which are also commonly used as loss functions for training DNN-based speech enhancement models. Such metrics are very simple and only consider raw features of the signal, without any consideration for psychoacoustic effects on speech perception. One first step in improving these metrics is the frequency-weighted segmental SNR (FWSegSNR) [140], which applies weights to different frequencies according to how relevant these frequencies are when a listener rates speech quality.

The ITU-T has proposed several metrics for objective speech quality estimation, namely PESQ [141], POLQA[93], and the recommendation P.563 [138]. All these metrics were designed for speech quality assessment in telephony applications, with PESQ and POLQA being intrusive metrics and designed both for narrowband and wideband, while P.563 is non-intrusive and designed for narrowband signals only. While many studies use these metrics for measuring the quality of enhanced speech (and they were shown several times to be highly correlated to enhanced speech quality), they were not originally developed for that purpose. The three metrics measure several different signal characteristics that are correlated to distortions and use regression models to convert such measurements into a mean opinion score. It is important to note that while PESQ and POLQA are intrusive metrics (i.e., require access to a clean reference signal), P.563 is non-intrusive (i.e. requires only the signal under test).

In addition to these metrics, we also investigated the composite metrics proposed in [140], which were designed specifically for assessing the performance of speech enhancement systems. These metrics were developed as a regression of multiple speech quality metrics (namely the Itakura-Saito distance, PESQ, cepstrum distance, log-likelihood ratio, and weighted spectral slope) for three quality categories: overall quality (C_{ovl}), signal quality (C_{sig}) and intrusiveness of the background noise (C_{bak}).

We also included the speech-to-reverberation modulation energy ratio (SRMR) metric, as proposed in [142], with the updates proposed in [143]. SRMR is a non-intrusive metric that is computed as the ratio between the modulation energies in low modulation frequency bands (assumed to be dominated by speech) and high energy bands (assumed to be dominated by distortions). The metric has been shown to be highly correlated to speech quality and intelligibility for noisy and reverberant speech.

5.2.3 Experimental Setup

We used the TIMIT corpus [89] as a source for the speech stimuli. The default training set (without the "SA" utterances, since these utterances were recorded by all speakers) was used for generating the training and validation sets, and the test set (with the SA utterances removed as well) was used for generating the test set. The training and test sets had a total of 462 and 168 speakers, respectively. The utterances were convolved with the same RIRs used for the single speaker experiments. A total of 3696 clean utterances were used for the training and validation set, and 1336 for the test set.

Reverberant utterances were generated by convolving randomly selected subsets of the utterances in the training set with 740 RIRs generated using a fast implementation of the image-source method [88], with T60 ranging from 0.2 s to 2.0 s in 0.05 s steps. Twenty different RIRs (with different room geometry, source-microphone positioning and absorption characteristics) were generated for each T60 value. Fifty random utterances from the training set were convolved with each of these 740 RIRs, resulting in 37,000 files. A random subset of 5% of these files was selected as a validation set and used for model selection and the remaining 35,150 files were used to train the models.

For the denoising dataset, we mixed the training and validation sentences with noises from the DEMAND dataset [108] at SNRs of 12, 6, 3, 0, -3, and -6 dB. The testing sentences were mixed with two noises (babble and factory noise) from the NOISEX dataset [109], at SNRs of 13, 7, 4, 1, -2, and -5 dB. For each sentence, a random noise segment with the same length as the sentence was picked. Signal energy for speech signals was computed according to the P.56 standard (which aims at only considering energy from speech segments and discarding silence segments), while the energy for noise signals was computed by its overall RMS value. For each (noise type, SNR) pair, we mixed all sentences in the training + validation set or the test set, accordingly. The training, validation, and test sets have 64923, 3417, and 1200 samples each.

For dereverberation, we generated a new set of 20 RIRs for T60 ranging from 0.3 to 1.5 s with 0.3 s steps. For each T60, we selected one RIR with low direct-to-reverberant ratio (DRR) and one with high DRR to use for generating the stimuli for the listening tests. Selections were based on the range of DRRs measured for each of the simulated RIR for each reverberation time.

Listening tests

We performed two different online listening tests, one for the dereverberation and one for the noise reduction models. Both tests were MUSHRA-style tests for evaluating speech quality by comparing it to a clean reference. In the tests, participants were presented with the outputs of all models (dereverberation or noise reduction) for a single speech stimulus simultaneously, as well as a hidden reference, the corrupted signal that was enhanced by the models, and an anchor, and asked to rate the signal quality using sliders, which have their positions quantized as integers in the 0-100 range. In the case of noise reduction conditions, the anchor was the same stimuli corrupted with the same type of noise but with an SNR 5 dB lower. In the case of dereverberation conditions, the anchor was a signal convolved with an RIR with a T60 of 2 s. In both cases, the reference was the clean anechoic signal. A total of 10 stimuli was used for each condition, where each stimuli consists of two concatenated sentences from the TIMIT dataset uttered by different speakers with a 2 seconds gap between the sentences. The minimum total stimuli length was 8 seconds.

Tests were carried out on Amazon Mechanical Turk using an online interface based on the Crowdsourced Audio Quality Evaluation (CAQE) [144]. A total of 245 participants took part in the denoising quality tests (mean age 33.94, standard deviation 9.68), evaluating a total of 12 conditions (120 stimuli). For the dereverberation quality tests, we had a total of 112 participants (mean age 35.46, standard deviation of 10.99) and 10 conditions (100 stimuli). Each participant could work on a maximum of 10 sessions, where each session corresponded to 5 random stimuli of a given distortion condition (we decided to split each condition in two sessions to reduce the duration of each task as recommended by Amazon Mechanical Turk). Presentation order of conditions and files was randomized for each participant. A minimum of 16 trials for each stimuli in each condition was performed.

Participants were asked about the audio setup they used and did a short listening test that aimed at checking if the frequency response of their setup was appropriate for the tests. The test

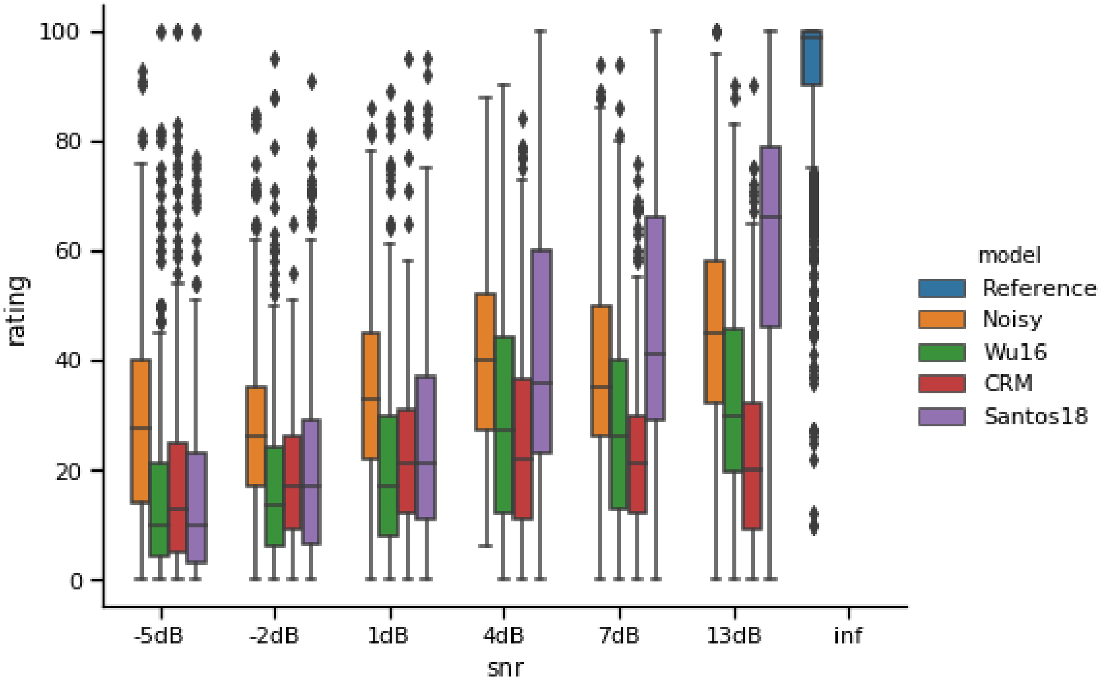


Figure 5.1 – Average quality ratings per SNR for the denoising models. Different colours for each boxplot indicate the model that generated the evaluated samples.

consisted of counting the number of "beeps" in a number of files, with beeps having a range of different frequencies and amplitudes. Only results from participants who passed the listening test were considered for further analysis.

5.2.4 Results and discussion

Listening test results

On the denoising test, results show the model proposed in Santos18 outperformed the other models for SNRs above 4 dB. There is no significant difference for lower SNRs, with participants always rating the noisy signals as having higher quality than processed ones. The masking model had the lowest quality scores of all models, and, upon inspection of the generated files, we noticed the amount of enhancement being performed by this model is very small. However, the model was still introducing some artifacts on speech even with higher SNR, which might explain why the scores for this model are lower than those for noisy files.

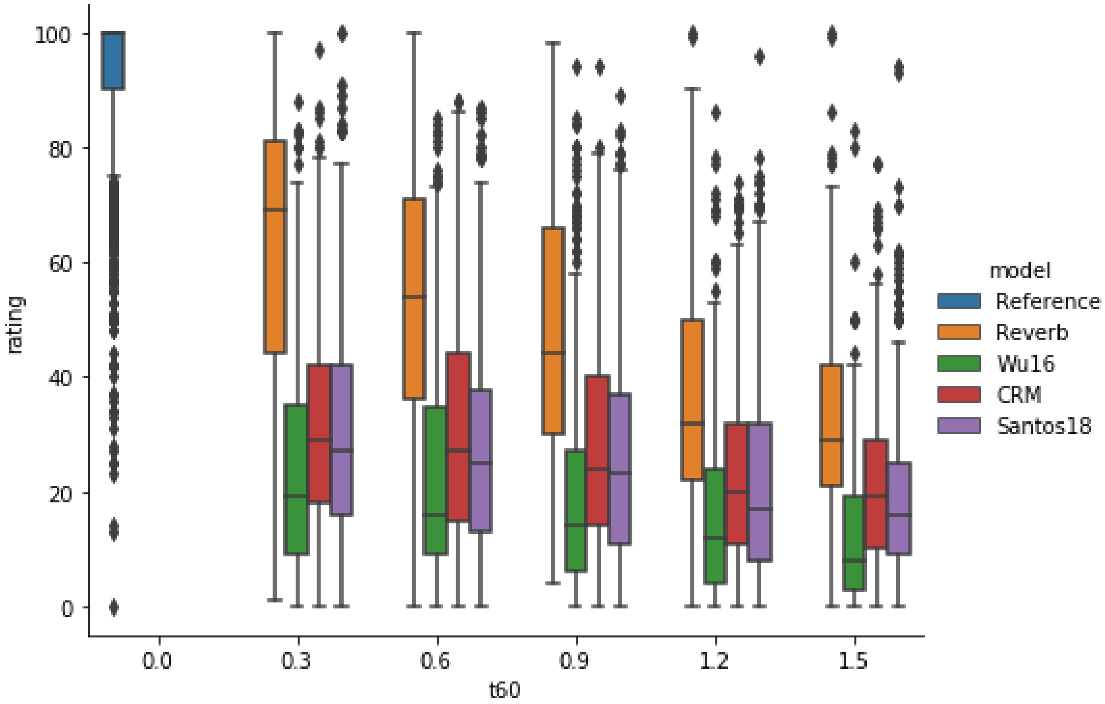


Figure 5.2 – Average quality ratings per T60 for the dereverberation models. Different colours for each boxplot indicate the model that generated the evaluated samples.

Dereverberation models decreased the signal quality in all scenarios, with the masking model and Santos18 scoring slightly higher than Wu2016. In the tested T60 range, reverberation does not affect speech quality, and reverberant speech with reverb might even be seen as more natural than anechoic speech. While it is not being measured here, intelligibility has been shown to only decrease with very large T60 values for normal hearing listeners.

Correlation between objective and subjective scores

We computed correlations for all objective metrics mentioned in section 5.2.2 for the files on both the denoising and dereverberation listening tests. Table 5.1 shows the results for the denoising tests, while table 5.2 shows the results for the dereverberation tests. The column "Sigmoid" shows the Pearson correlation scores after fitting a generalized linear model with logit as a link function instead of the identity (which maps better to speech quality due to the saturation at the lower and upper ends of the interval), and RMSE shows the root mean squared error for the predictions of this model.

Table 5.1 – Correlation between objective and subjective scores for the denoising task

Metric	Pearson	Spearman	Sigmoid	RMSE
SegSNR	0.948	0.784	0.953	7.963
FWSegSNR	0.934	0.711	0.937	9.132
C_{sig}	0.903	0.646	0.903	11.230
C_{bak}	0.940	0.723	0.940	8.915
C_{ovl}	0.916	0.711	0.917	10.402
PESQ_{adj}	0.944	0.685	0.945	8.571
PESQ_{MOS}	0.910	0.685	0.915	10.535
$\text{PESQ}_{\text{dist}}$	-0.857	-0.688	0.874	12.711
$\text{PESQ}_{\text{asym}}$	-0.927	-0.579	0.930	9.633
POLQA	0.940	0.777	0.937	9.116
$\text{SRMR}_{\text{norm}}$	0.630	0.633	0.646	19.962
P.563	0.760	0.660	0.777	16.470

Table 5.2 – Correlation between objective and subjective scores for the dereverberation task

Metric	Pearson	Spearman	Sigmoid	RMSE
SegSNR	0.901	0.423	0.914	10.439
FWSegSNR	0.922	0.656	0.900	11.246
C_{sig}	0.832	0.621	0.917	10.315
C_{bak}	0.893	0.409	0.560	21.981
C_{ovl}	0.904	0.624	0.743	17.737
PESQ_{adj}	0.910	0.446	0.741	17.667
PESQ_{MOS}	0.889	0.446	0.814	15.555
$\text{PESQ}_{\text{dist}}$	-0.864	-0.448	0.888	11.850
$\text{PESQ}_{\text{asym}}$	-0.907	-0.445	0.890	11.922
POLQA	0.874	0.431	0.900	11.266
$\text{SRMR}_{\text{norm}}$	0.351	0.082	0.895	11.520
P.563	0.727	0.475	0.810	15.338

Both tables 5.1 and 5.2 show that simpler, signal-based metrics such as SegSNR and FWSegSNR have similar performance to more complex metrics such as PESQ, POLQA, and the composite metrics from [140]. The performance of the metrics for dereverberated speech is slightly poorer than the performance for denoising, which might be due to the small range covered by processed speech in the rating axis (as can be seen in Figure 5.2), since all enhancement models had relatively poor performance when compared to the denoising task.

5.2.5 Investigating non-intrusive metrics for DNN-based speech enhancement

Discovering useful features

In order to discover features that could be useful for estimating speech quality for DNN-based enhancement in a non-intrusive way, we look into the correlations of the internal features of the P.563 metric with the mean opinion scores obtained in our listening tests. P.563 contains three modules that perform pitch-synchronous vocal tract modeling and linear predictive coding (LPC) analysis, speech reconstruction and full-reference perceptual modeling (using a virtual reference), and computing of distortion-specific parameters (noises, temporal clipping, and robotization). Based on the outputs of these modules, P.563 selects a dominant distortion category and perceptual weighting to compute its final MOS scores. Using the reference software provided by the ITU-T, we extracted the 46 internal parameters from all files in our dataset and averaged the features on a per-condition basis (where by condition we mean same distortion and processing).

The features with highest correlations for both datasets come from the LPC analysis and the distortion-specific parameters modules. Namely, the kurtosis and skewness of the LPC parameters were highly correlated to speech quality for both denoising (0.870 and 0.862) and dereverberation (0.837 and 0.846). Kurtosis and skewness measure the "tailedness" and asymmetry of a distribution, and are used in P.563 to assess speech naturalness by checking whether such values fall in the range expected for natural speech. In P.563, these statistics are computed on a per-frame basis over the LP coefficient vector, then averaged to yield final LP skewness and kurtosis scores.

For speech processed by the denoising models, the local background noise parameter was also highly correlated (-0.814). This parameter is estimated from the RMS energy in intervals between phonemes of a sentence, computed for 20 ms non-overlapping frames and averaged over all frames of the signal under test.

Two features computed for analyzing the effect of multiplicative noise were also found to be highly correlated, namely the spectral level range (Pearson correlation of 0.826 with denoising quality ratings) and deviation (Pearson correlation of 0.744 with dereverberation quality ratings). These two features are computed based on the range and standard deviation of short-term spectral power densities for active voice frames (as multiplicative noise is only present during speech activity).

Table 5.3 – Pearson correlations for the leave-one-model-out cross-validation for the denoising models.

Metric	Santos18	CRM	Wu16
SegSNR	0.935	0.469	0.817
FWSegSNR	0.900	0.324	0.759
$\mathrm{C_{sig}}$	0.131	-0.000	0.091
$\mathrm{C_{bak}}$	0.933	0.327	0.847
$\mathrm{C_{ovl}}$	0.643	-0.030	0.384
$\mathrm{PESQ_{adj}}$	0.935	0.226	0.788
$\mathrm{PESQ_{MOS}}$	0.912	0.349	0.820
$\mathrm{PESQ_{dist}}$	0.892	0.353	0.819
$\mathrm{PESQ_{asym}}$	0.917	0.215	0.705
POLQA	0.930	0.359	0.785
$\mathrm{SRMR_{norm}}$	0.530	0.454	0.706
P.563	0.820	0.247	0.694
Proposed	0.895	0.341	0.820

Leave-one-model-out cross validation analysis of the discovered features

In order to assess how effective such features are in predicting speech quality for DNN-enhanced speech, we perform here a leave-one-model-out (LOO) cross-validation analysis of a linear regression model trained to use such features. The reason to use LOO is that we wanted to make sure the results we found are not biased by our data. We trained separate linear regression models for the denoising and dereverberation tests, using the features described previously. For the denoising model, we used LPC kurtosis and skewness, log-energy of the local background noise, and spectral level range. For the dereverberation model, we used LPC kurtosis and skewness, spectral level deviation, and the P.563 MOS. We used Pearson correlations as a performance metric for these models.

From Table 5.3, we can see the proposed metric has correlation in line with intrusive metrics and outperforms other non-intrusive metrics. The correlations for the CRM model are lower than those for Santos18 and Wu16, which is probably due to all outputs from this model being rated in the same quality range independently of the condition being assessed, as can be seen in Figure 5.1.

As shown in Table 5.4, the correlations for the dereverberation tests are significantly lower than for denoising, even for intrusive metrics. This can be explained again by the short range covered by the results of all models: in each of the cross-validation training sets, correlation between the given features and the metric is high because a large increase/decrease in a given metric leads to a large

Table 5.4 – Pearson correlations for the leave-one-model-out cross-validation for the dereverberation models.

Metric	Santos18	CRM	Wu16
SegSNR	0.101	0.366	0.346
FWSegSNR	0.389	0.611	0.361
C_{sig}	0.359	0.658	0.288
C_{bak}	0.628	0.533	0.624
C_{ovl}	0.388	0.604	0.301
$\mathrm{PESQ}_{\mathrm{adj}}$	0.594	0.573	0.572
$\mathrm{PESQ}_{\mathrm{MOS}}$	0.601	0.606	0.616
$\mathrm{PESQ}_{\mathrm{dist}}$	0.578	0.613	0.587
$\mathrm{PESQ}_{\mathrm{asym}}$	0.164	0.423	0.622
POLQA	0.474	0.552	0.468
$\mathrm{SRMR}_{\mathrm{norm}}$	0.152	0.643	0.306
P.563	0.236	0.410	0.338
Proposed	0.341	0.510	0.610

increase/decrease in the rating. However, when applying the learned models to the test set, which only contains the metrics computed for the outputs of a given model, the target variable only covers a short range of ratings, which leads to poorer correlations. For dereverberation, the proposed metric does not perform as well as some of the intrusive metrics considered in this study, especially for the Santos18 model. The PESQ metrics had the highest correlations among all metrics, even though PESQ, unlike POLQA, was not originally designed for taking reverberation into account.

5.3 Quality-aware DNN-based speech enhancement using vocoder representations

Given the insights gained from our listening tests, we designed some experiments to try to include some of the alternative quality metrics we found as a cost function for training DNN-based speech enhancement models.

The simplest features with the highest correlations with speech quality from the set of features we analyzed were the LPC higher-order statistics skewness and kurtosis, with kurtosis having slightly higher correlations. However, we could not apply them directly to a model whose outputs are in the STFT domain. Converting from the STFT domain to the LPC involves either matrix inversions or an iterative solver (based on the Levinson-Durbin algorithm), both being costly operations when we need to compute the gradients relative to their inputs to be able to backpropagate through them for each iteration of a stochastic gradient descent solver.

In order to alleviate that issue and still have a proof-of-concept model to evaluate whether it is useful to incorporate features related to the statistics of the LPC representation, we updated our DNN-based speech enhancement model to generate outputs in two different vocoder representations. First, we used the reflection coefficients representation for the LPC, since they are more stable and bounded between -1 and 1. In order to be able to synthesize the enhanced signal, we also make the model predict the fundamental frequency (F0) for each frame. Second, we used a more recent vocoder, namely the WORLD vocoder, that uses mel-generalized cepstrum (MGC) coefficients instead of LPCs to represent the spectral envelope, as well as a mixed excitation signal.

5.3.1 Vocoders

LPC

Linear predictive coding is widely used for speech coding. A simple type of vocoder using LPC is a single excitation vocoder, where the LP coefficients represent the spectral envelope of the speech signal (which models the vocal tract) and an excitation signal, composed either of trains of periodic pulses (representing the vibration of the vocal folds for voiced sounds) or noise (for unvoiced sounds). While this simple vocoder tends to generate some artifacts, it is simple and does not require any special parameters besides some representation of the LPC coefficients and the fundamental frequency (or lack of F0) for each speech frame.

WORLD

The WORLD vocoder [145] is a more recent vocoder which has some fundamental differences when compared to a LPC-based vocoder. First, instead of using standard linear prediction, the WORLD vocoder uses a representation based on the mel-generalized cepstrum. This representation uses perceptual frequency warping followed by a discrete cosine transform, which can be truncated depending on the amount of compression required.

The WORLD vocoder is also a mixed excitation vocoder. For each frame, both the fundamental frequency and a measure of the aperiodicity of the signal are used to generate the excitation signal. This allows the generation of more natural-sounding speech.

Although features based on the MGC representation were originally not evaluated for the experiments in the previous section, using the same methods and data we found the standard deviation of the MGC coefficients to be highly correlated to speech quality ($\rho = 0.92$), so in this section we experiment with it as a replacement for higher-order statistics of the LPC representation. Like in P.563, these statistics were computed on a per-frame basis and then averaged.

5.3.2 Experimental setup

Model architecture and training

For the experiments reported in this section, we used a model similar to the one described in Chapter 3. However, the output layer dimensions were changed to make it suitable for generating vocoder features instead of STFT coefficients.

All models were trained with noisy data generated from the single speaker IEEE dataset and the DEMAND noise corpus as described in section 4.3.3. Vocoder features were generated from the clean files and used as targets. The LPC-based vocoder used 40 reflection coefficients, one broadband gain coefficient, and one pitch coefficient. All features were estimated with 32 ms windows with 75% overlap. For the WORLD vocoder, we used 60 MGCs, log-F0, and one band aperiodicity coefficient per frame, and we used windows of 40 ms with 5 ms step size.

We used off-the-shelf software libraries for our vocoders. Both WORLD and LPC features were extracted using the command-line tools provided by the Speech Toolkit (SPTK)[146]. For WORLD, we used the WORLD vocoder implementation provided by the Merlin project [147] and the scripts for synthesis. Synthesis for the LPC-based vocoder was performed by using the tools provided in the pysptk library[148]. SWIPE' [149] was used as the pitch estimator for the LPC-based vocoder and REAPER [150] for the WORLD vocoder.

All models were trained until the validation loss did not increase for 5 epochs in a row, up to a maximum of 100 epochs, using the Adam optimizer. This was done to avoid overfitting, since we used a relatively small dataset. To incorporate either the kurtosis of the LPC coefficients or the standard deviation of the MGC coefficients into the loss function, we simply added the value of the corresponding statistic as computed for an entire mini-batch, multiplied by a Lagrangian multiplier α, to the original MSE loss.

For each type of model, we report the distributions of the PESQ, STOI, and SRMR scores for each condition in the dataset. We used different values for α for each model (10^{-2} and 10^{-4} for the WORLD-based model and $10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}$ for the LPC-based model). No other hyperparameter search was performed.

5.3.3 Results

Objective quality and intelligibility metrics for each method can be seen in figures 5.3-5.8. Besides the results for each model, we also show the results for the original noisy signals and the target signals after analysis-synthesis with the respective vocoder (labeled as "target"). The latter gives us the maximum performance our models would achieve if they were able to generate perfect vocoder features for the clean signal given a corrupted signal. Including the LPC kurtosis with various different values for α did not improve objective metrics for the model with LPC vocoder outputs, with the scores approximating the baseline score for the model trained with standard MSE as α gets smaller. For the model based on the WORLD vocoder, however, we see a small increase in PESQ scores (figure 5.4) and no significant variation in the STOI and SRMR scores.

We also performed a small-scale online listening test[1] to verify whether listeners would prefer the WORLD model using the proposed loss as opposed to the baseline MSE loss. This was an ABX test where participants were asked to listen to the same sample enhanced by both models (without knowing which one came from which model) and select which sample they thought was more natural. A total of 12 participants took part in the test, having listened to the same 10 sentence pairs each. Both the order of sentences and order of models presented were randomized. Figure 5.9 summarizes the results. As we can see, over 50% of the participants preferred the baseline loss function, followed by approximately 25% having no preference between both.

[1] Test interface available at `https://peaceful-coast-17742.herokuapp.com/index.html`

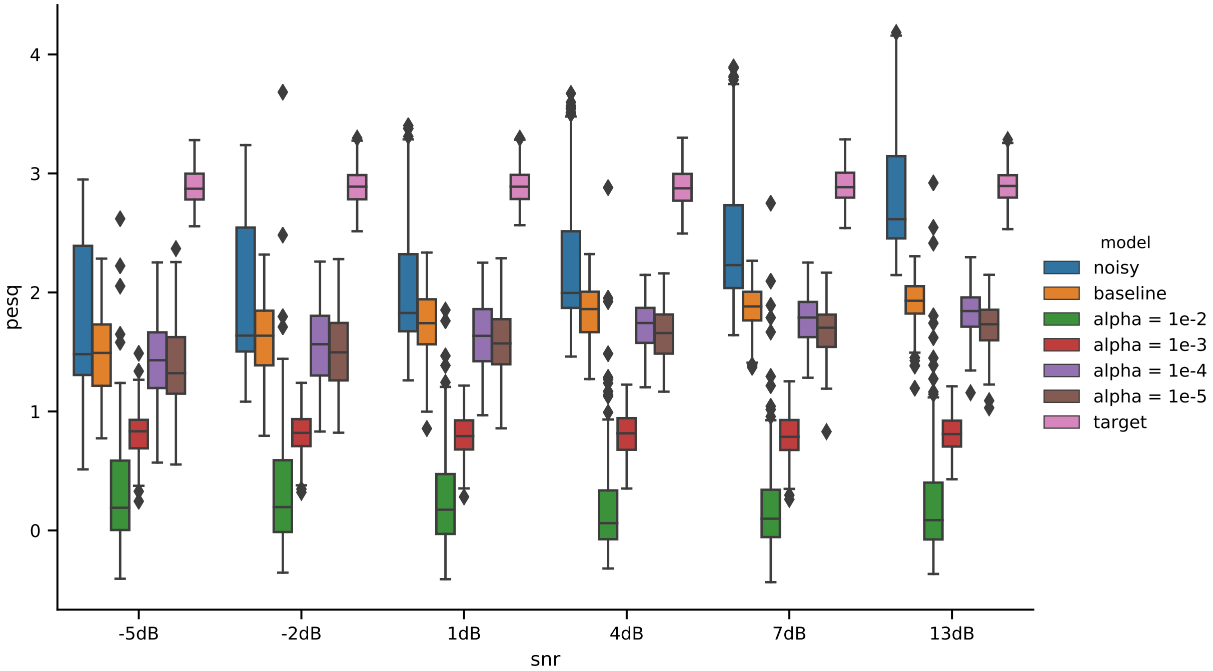


Figure 5.3 – PESQ scores for the LPC-based model

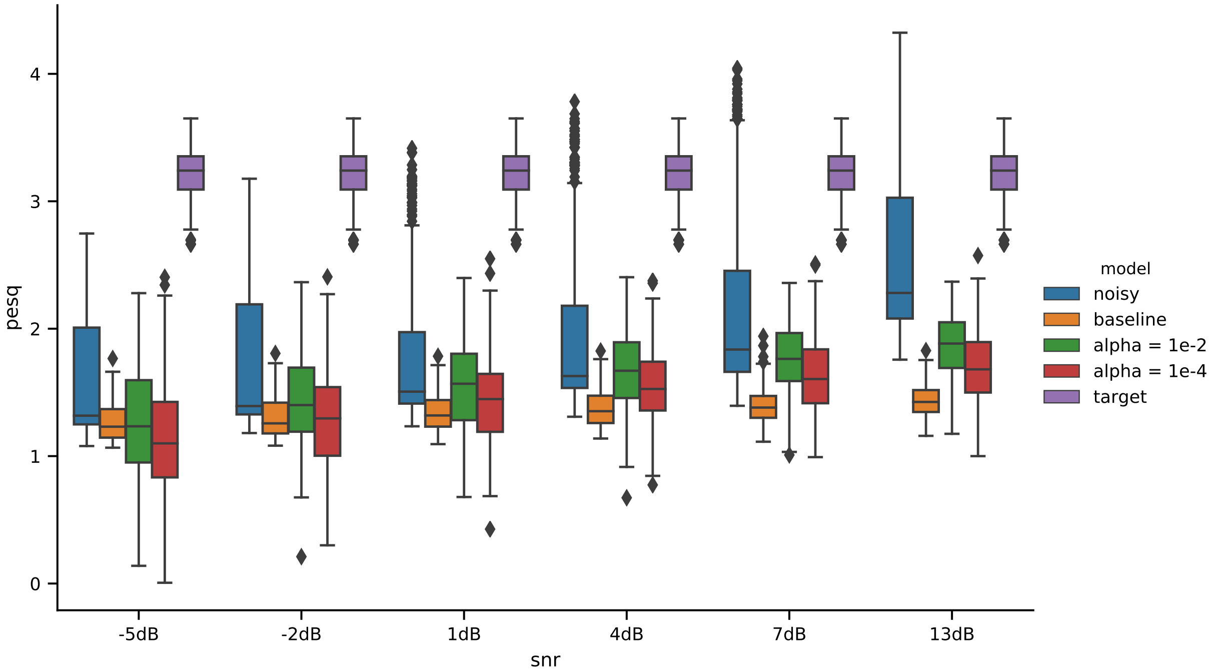


Figure 5.4 – PESQ scores for the WORLD-based model

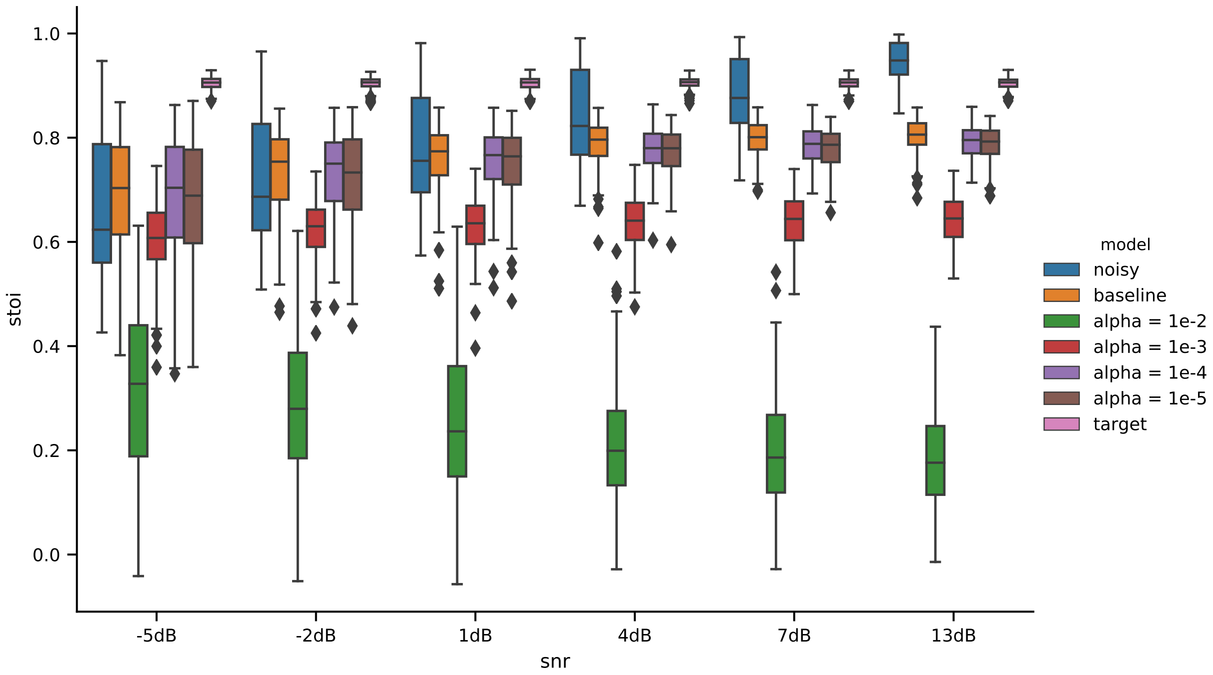


Figure 5.5 – STOI scores for the LPC-based model

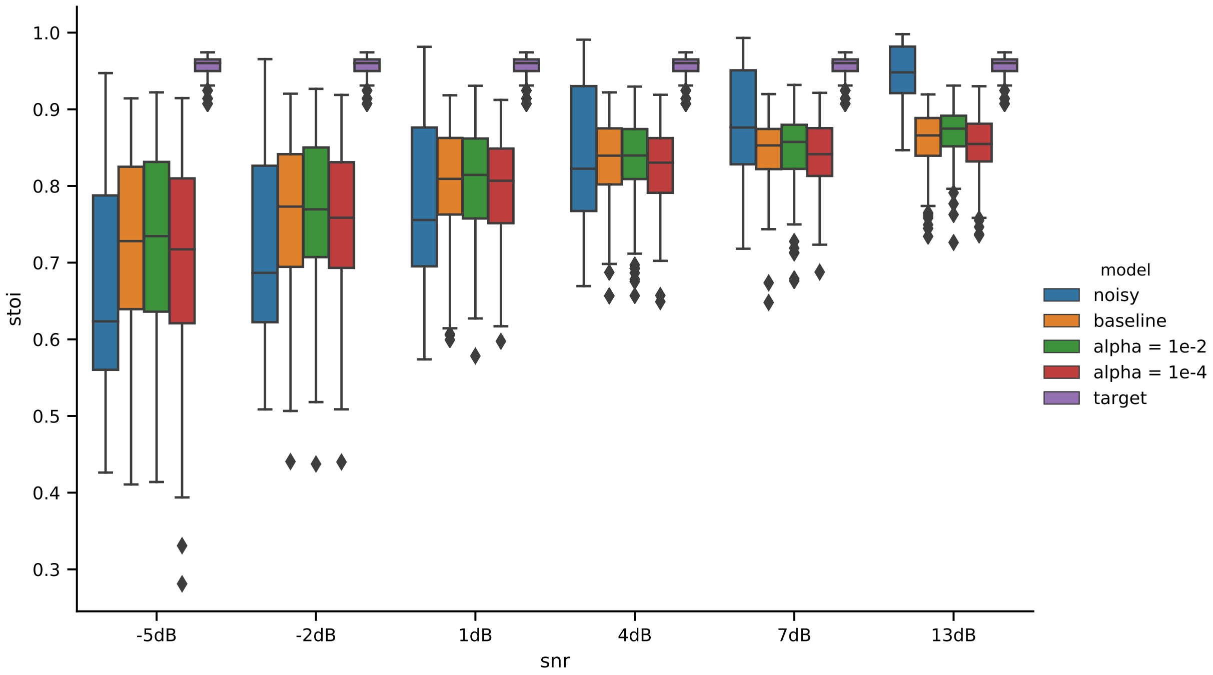


Figure 5.6 – STOI scores for the WORLD-based model

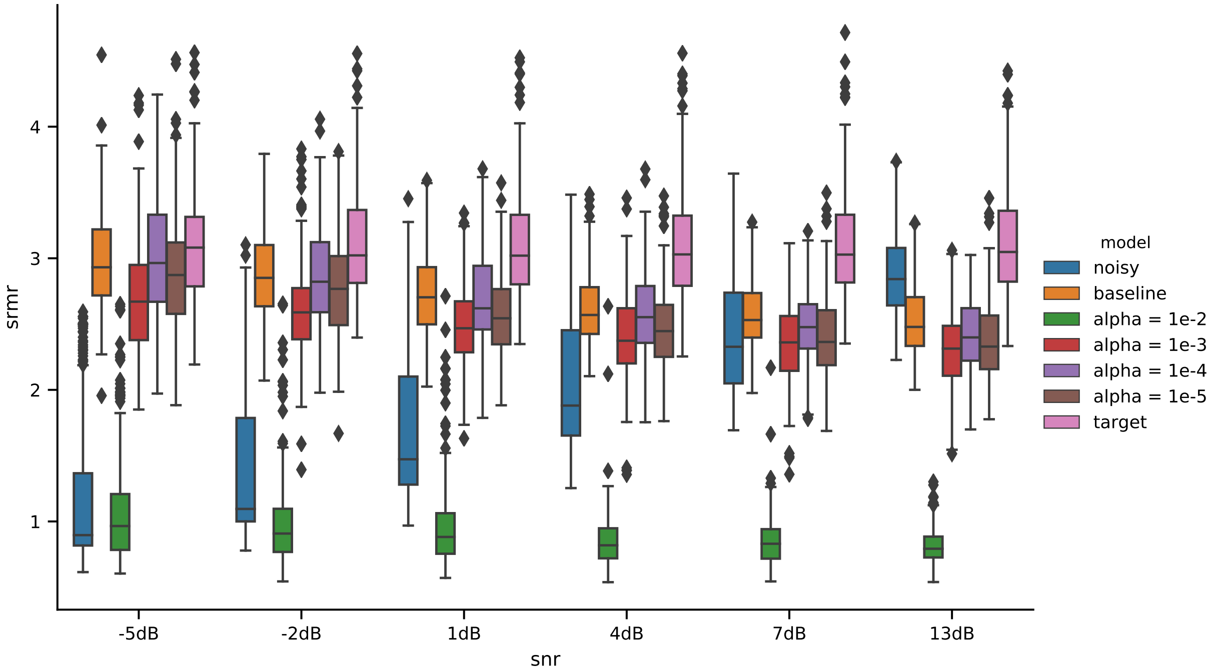


Figure 5.7 – SRMR scores for the LPC-based model

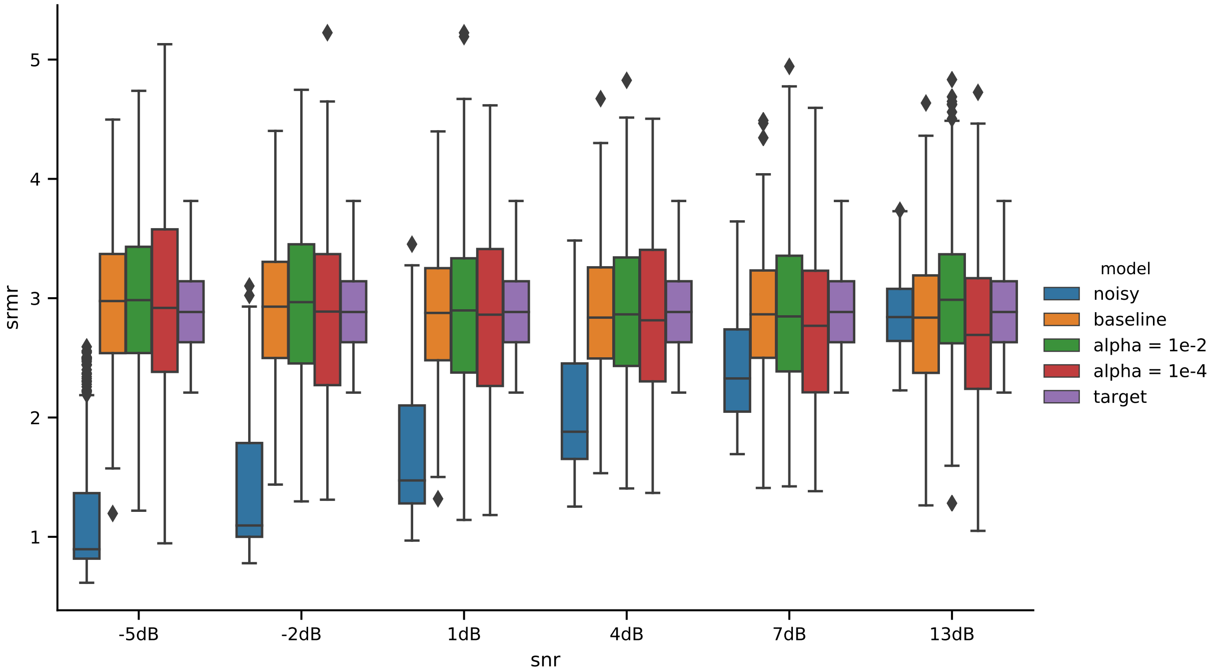


Figure 5.8 – SRMR scores for the WORLD-based model

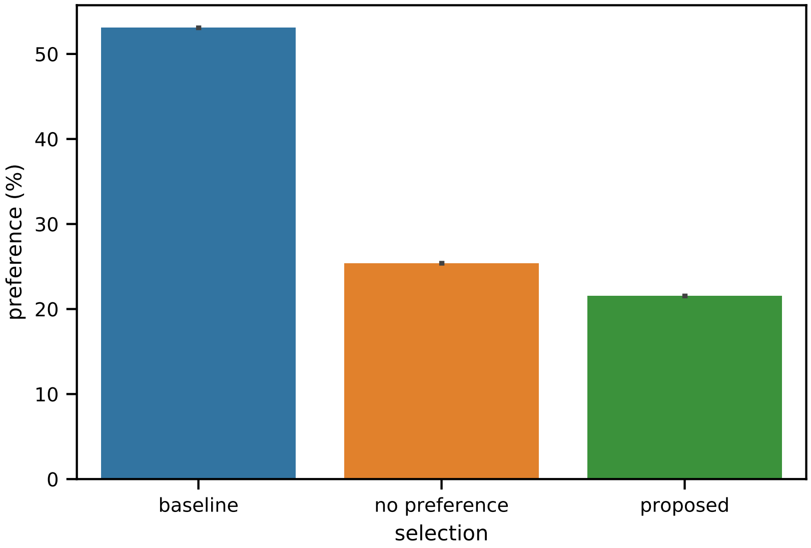


Figure 5.9 – Results of the preference test (in %)

5.3.4 Discussion

The results we observed do not fully support our hypothesis that incorporating quality metrics into the loss function leads to improvement of speech quality; notwithstanding, roughy half the listening test participants either preferred the proposed method or did not find a perceptual difference with the baseline. Moreover, there are some considerations about the limitations of this proof-of-concept study. First, using a vocoder representation for the output instead of an STFT output as for the models we evaluated in the last section causes a mismatch between the metrics we studied and the models where we were able to apply it. Additionally, using vocoders itself decreases speech naturalness and limits the maximum quality we can get, as can be observed in Figures 5.3-5.8. Additionally, the vocoders we used are very dependent on good predictions of F0, which is a challenging problem under noise [151]. We did not take any special considerations for pitch prediction in the models we used here. The pitch tracker used in the Merlin implementation of the WORLD vocoder, specifically, was designed for speech synthesis where the input is usually clean speech signals recorded on a studio setting.

We have also not performed any hyperparameter search for the DNN model we used, other than for the Lagrangian multiplier α. Since the output of the model does not match the outputs for which we performed hyperparameter searches in Chapter 3, more work would be needed in order to find an appropriate set of hyperparameters.

We would also like to acknowledge that during the development of this work, a similar approach based on combining MSE and an objective metric specific for speech was presented by Zhao et al. [152], which is based on early works of the author of this thesis. In their work, they use a modified implementation of the STOI metric and train the model to predict a ratio mask that approximates the IRM and maximizes STOI simultaneously. Their approach shows improvements in STOI and speech-to-distortion ratio, but not in PESQ. However, the authors did not perform a listening test, so comparison with our work is limited. Additionally, it should be noted that STOI is an intrusive metric, where in our work we explore using a non-intrusive metric.

5.4 Conclusion

The work presented in this section investigated the performance of multiple objective speech quality metrics for DNN-based speech enhancement and presented the first steps towards a non-intrusive metric for such systems. The proposed metric, based on internal metrics of the P.563 standard, performs in line with intrusive metrics for speech denoising, but does not perform as well for speech dereverberation. Even though the metric proposed is a work-in-progress, insights provided by the metrics that presented high correlations with speech quality can be taken into account for designing better cost functions or constraints for supervised training of speech enhancement models.

Along these lines, we have also presented proof-of-concept experiments on a quality-aware DNN-based speech enhancement model. Due to the numerical difficulties to backpropagate through vocoder parameter estimation, we experimented with training models to directly generate vocoder representations at their outputs, matching the domain of some of the features we encountered that were shown to be highly correlated to speech quality of DNN-based enhanced speech. While our approach with the WORLD vocoder increases PESQ scores, on a small-scale preference test participants did not show a clear preference for the proposed method and roughly half preferred the baseline. These experiments are limited in scope since we did not originally test quality metrics with vocoders.

As future work, we envision a similar listening test for speech intelligibility, which could lead to a similar study that would investigate the signal aspects connected to the intelligibility of DNN-enhanced speech. Future studies could also take into account recently proposed models for DNN-based enhancement that process signals in the time domain, such as [71], as these will likely generate a different set of artifacts than magnitude spectrum processing. For quality-aware models, next steps would be trying to improve the performance of the model for pitch prediction, likely with separate branches for MGC, log-F0, and aperiodicity coefficients. Other representations/quality metrics that avoid the numerical/computational issues we found with the LP kurtosis could also be explored. Finally, an extended version of the listening tests using models with outputs in vocoder domains could be useful for identifying other useful metrics to be used as complementary loss functions.

Chapter 6

Conclusions and future research directions

6.1 Conclusions

Since the relatively recent successes of deep learning in several areas, including multimedia processing such as speech, images, and video, there has been an enourmous amount of research efforts in the field. DNN-based models are now the state of the art in ASR, having recently achieved similar accuracy to that of human transcribers [153]. Text-to-speech models based on DNNs have also achieved MOS in line with those given to professional audio recordings [68].

The area of speech enhancement, including denoising and dereverberation, while very active, has not seen the same level of attention and performance. In this work we have attempted to bridge this gap by working on three separate topics we believed required more research efforts: single-channel speech dereverberation, interpretability for DNN-based speech processing models, and quality assessment/awareness for DNN-based speech enhancement.

First, we proposed a model for single-channel speech dereverberation that is able to exploit both short- and long-term context from the speech signal. The model is based on both convolutional and recurrent neural networks, where the local processing of CNNs is used to provide short-term context and long-term memory in GRUs for longer-term context. We have showed that this model outperformed other frequency-domain models presented in the literature, while requiring fewer parameters. While there are several recent advances in time-domain DNN-based speech processing, we believe frequency-domain processing is still relevant since current time-domain models have large amounts of parameters and are very computationally costly, and many of them cannot operate in real-time without specialized hardware.

Even though DNN-based speech processing has achieved state-of-the-art performance in many problems, we still do not understand how it operates very well, and this limits how research is done in the domain. Most recent advances have been achieved through inserting some domain knowledge into the models, or, more commonly, training models with more data and increasing their capacity. In this work, we have also started investigating some of the mechanisms that are common to many DNN-based speech processing models, namely skip connections and recurrent neural networks. We have shown that the use of skip connections leads to more interpretable speech enhancement models, since we can better understand what is being predicted by each part of the model. Through investigating internal workings of the single-channel model proposed in Chapter 3, we were also able to detect an issue with saturating functions in RNNs and proposed a method to prune those connections, which leads to a reduction in storage and computational requirements. The method we proposed is post hoc, being useful even for already trained models, and is complementary to other pruning/compression methods in the literature.

Finally, we showed through a series of listening tests that while current metrics are correlated with speech quality from DNN-enhanced speech, there is a significant gap in perceived quality and such metrics. This is especially the case for reverberant/dereveberated speech. We proposed a new non-intrusive metric based on relevant acoustic features that performs as well as the investigated intrusive metrics. This work led us into taking first steps into developing a quality-aware DNN-based speech enhancement model. The proposed model and quality-aware cost function achieve an improvement in the PESQ objective metric; however, a preference test shows that listeners consider speech generated by this model less natural than by a baseline model trained with a standard mean-squared error loss. While our results are limited in scope, we consider them a step forward in the research of quality-aware models.

6.2 Future research directions

There are multiple avenues for future research based on the work proposed in this thesis. For speech enhancement models, this thesis has only considered single-channel approaches. While there are already works on multi-channel approaches for ASR, we believe more research needs to be done for multi-channel approaches that aim at generating high quality and intelligibility speech. We also believe the model proposed in Chapter 3 can be improved by increasing the depth of the convolutional encoder and decreasing the amount of input features for the RNN-based decoder. This would make the model more efficient computationally and might also make the first RNN layer more amenable to pruning/compression. Finally, more research is needed on computationally-efficient ways of incorporating phase prediction and/or enhancement. This is currently a well-known issue and there are multiple research efforts in that regard [154, 155, 156].

Our work on model interpretability has only explored two aspects of DNN-based speech processing models, and only for speech enhancement. We would like to explore the approaches presented here, and also in the computer vision research community [157], for a wider variety of speech and audio processing models. Our early explorations have shown that the methods proposed for computer vision (such as activity maximization for neurons) do not necessarily map well to generative models for audio such as speech enhancement and synthesis. By using simple methods with speech recognition, enhancement, and synthesis models, we aim to better understand how these models work and to propose novel methods for interpreting their outputs, thus giving more tools for researchers to improve these models.

In the same lines, the method we proposed for weight pruning was only tested for speech enhancement models, and we would like to evaluate whether it is useful for other models in the literature that rely on RNNs such as speech recognition, synthesis, and natural language processing models such as those used in machine translation. The method we proposed is very simple and does not guarantee any units will be pruned since it is entirely post hoc, so we would like to investigate ad hoc methods based on the same principles. We would also like to investigate whether it is possible to find regularization functions that avoid neurons with saturating functions getting "stuck".

Finally, we would also like to extend the study done on speech quality metrics to more DNN-based enhancement models, including time-domain models, and also to speech synthesis. It would also be useful to investigate intelligibility-related metrics, since in some scenarios (such as assistive hearing devices) higher intelligibility might have higher priority than quality. The work on quality- and intelligibility-aware enhancement models reported here and in [152] are very preliminary and more work is needed, since there are clear limitations in our approach using outputs in the vocoder space and the improvements of the intelligibility-aware work on subjective ratings are unclear. By developing reliable non-intrusive quality metrics, we would also be able to incorporate unsupervised/semi-supervised learning into speech enhancement and synthesis models. In enhancement models, unsupervised and semi-supervised learning would allow us to leverage the immense amounts of speech data available online, while on synthesis it could be a way of improving model quality while keeping naturalness and variability.

Bibliography

[1] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 114–126, Nov. 2012. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6296524

[2] A. C. Neuman, M. Wroblewski, J. Hajicek, and A. Rubinstein, "Combined effects of noise and reverberation on speech recognition performance of normal-hearing children and adults," *Ear and Hearing*, vol. 31, no. 3, pp. 336–344, Jun. 2010.

[3] S. Boll, "A spectral subtraction algorithm for suppression of acoustic noise in speech," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '79.*, vol. 4, Apr. 1979, pp. 200–203.

[4] B. Cauchi, I. Kodrasi, R. Rehr, S. Gerlach, A. Jukić, T. Gerkmann, S. Doclo, and S. Goetze, "Combination of MVDR beamforming and single-channel spectral processing for enhancing noisy and reverberant speech," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 1, pp. 1–12, 2015. [Online]. Available: http://link.springer.com/article/10.1186/s13634-015-0242-x

[5] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015. [Online]. Available: http://www.nature.com/nature/journal/v521/n7553/full/nature14539.html

[6] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and others, "Deep neural networks for

acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6296526

[7] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2013, pp. 6645–6649.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-w

[9] M. Mimura, S. Sakai, and T. Kawahara, "Reverberant speech recognition combining deep neural networks and deep autoencoders augmented with a phone-class feature," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 1, p. 62, Jul. 2015. [Online]. Available: http://asp.eurasipjournals.com/content/2015/1/62/abstract

[10] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1–1, 2014. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6932438

[11] K. Han, Y. Wang, D. Wang, W. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 982–992, Jun. 2015.

[12] F. Weninger, J. Hershey, J. Le Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Dec. 2014, pp. 577–581.

[13] 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*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[14] T. Falk and W.-Y. Chan, "Temporal dynamics for blind measurement of room acoustical parameters," *IEEE Transactions on Instrumentation and Measurement*, vol. 59, no. 4, pp. 978–989, Apr. 2010.

[15] A. Nabelek, “Effects of room acoustics on speech perception through hearing aids by normal hearing and hearing impaired listeners,” in *Acoustical Factors Affecting Hearing Aid Performance*, G. Stubebaker and I. Hochberg, Eds. Needham Heights, USA: Allyn and Bacon, 1993, pp. 15–28.

[16] O. Hazrati and P. C. Loizou, “The combined effects of reverberation and noise on speech intelligibility by cochlear implant listeners,” *International journal of audiology*, vol. 51, no. 6, pp. 437–443, 2012. [Online]. Available: http://informahealthcare.com/doi/abs/10.3109/14992027.2012.658972

[17] P. C. Loizou, “Speech enhancement based on perceptually motivated Bayesian estimators of the magnitude spectrum,” *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 5, pp. 857–869, 2005. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1495469

[18] R. Martin, “Speech enhancement using MMSE short time spectral estimation with gamma distributed speech priors,” in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, May 2002, pp. I–253–I–256.

[19] ——, “Speech enhancement based on minimum mean-square error estimation and supergaussian priors,” *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 845–856, Sep. 2005.

[20] O. Räsänen, “Average spectrotemporal structure of continuous speech matches with the frequency resolution of human hearing.” in *INTERSPEECH*, 2012. [Online]. Available: http://users.spa.aalto.fi/orasanen/papers/IS12_freqresolution.pdf

[21] Y. Ephraim and H. Van Trees, “A signal subspace approach for speech enhancement,” *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 251–266, Jul. 1995.

[22] F. Jabloun and B. Champagne, “Incorporating the human hearing properties in the signal subspace approach for speech enhancement,” *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 700–708, Nov. 2003. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1255456

[23] L. Parra and P. Sajda, “Blind source separation via generalized eigenvalue decomposition,” *The Journal of Machine Learning Research*, vol. 4, pp. 1261–1269, 2003. [Online]. Available: http://dl.acm.org/citation.cfm?id=964305

[24] P. Smaragdis, “Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs,” in *Independent Component Analysis and Blind Signal Separation.* Springer, 2004, pp. 494–499. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-540-30110-3_63

[25] T. O. Virtanen, “Monaural sound source separation by perceptually weighted non-negative matrix factorization,” *Tampere University of Technology, Tech. Rep*, 2007. [Online]. Available: http://www.researchgate.net/profile/Tuomas_Virtanen2/publication/228626834_Monaural_sound_source_separation_by_perceptually_weighted_non-negative_matrix_factorization/links/0c96051cc0515a1514000000.pdf

[26] T. Barker and T. Virtanen, “Non-negative tensor factorisation of modulation spectrograms for monaural sound source separation.” in *INTERSPEECH*, 2013, pp. 827–831. [Online]. Available: http://ee301iitk.wdfiles.com/local--files/student-project-list/TP22.PDF

[27] M. Davies and C. James, “Source separation using single channel ICA,” *Signal Processing*, vol. 87, no. 8, pp. 1819–1832, Aug. 2007. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0165168407000151

[28] B. Yegnanarayana, C. Avendano, H. Hermansky, and P. S. Murthy, “Speech enhancement using linear prediction residual,” *Speech Communication*, vol. 28, no. 1, pp. 25–42, 1999. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639398000703

[29] B. Yegnanarayana and P. S. Murthy, “Enhancement of reverberant speech using LP residual signal,” *Speech and Audio Processing, IEEE Transactions on*, vol. 8, no. 3, pp. 267–281, 2000. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=841209

[30] Mingyang Wu and DeLiang Wang, “A two-stage algorithm for one-microphone reverberant speech enhancement,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 3, pp. 774–784, May 2006. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1621193

[31] T. H. Falk, S. Stadler, W. B. Kleijn, and W.-Y. Chan, “Noise suppression based on extending a speech-dominated modulation band.” in *INTERSPEECH*, 2007, pp. 970–973. [Online]. Available: http://20.210-193-52.unknown.qala.com.sg/archive/archive_papers/interspeech_2007/i07_0970.pdf

[32] K. Paliwal, K. Wójcicki, and B. Schwerin, “Single-channel speech enhancement using spectral subtraction in the short-time modulation domain,” *Speech Communication*, vol. 52, no. 5, pp. 450–475, May 2010. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0167639310000282

[33] E. A. P. Habets, *Single- and multi-microphone speech dereverberation using spectral enhancement*. Citeseer, 2007, vol. 68, no. 04.

[34] M. Delcroix, T. Hikichi, and M. Miyoshi, “Blind dereverberation algorithm for speech signals based on multi-channel linear prediction,” *Acoustical Science and Technology*, vol. 26, no. 5, pp. 432–439, 2005.

[35] M. Miyoshi and Y. Kaneda, “Inverse filtering of room acoustics,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 36, no. 2, pp. 145–152, Feb. 1988.

[36] T. Nakatani, K. Kinoshita, and M. Miyoshi, “Harmonicity-based blind dereverberation for single-channel speech signals,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 80–95, Jan. 2007. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4032782

[37] C. S. J. Doire, M. Brookes, P. A. Naylor, C. M. Hicks, D. Betts, M. A. Dmour, and S. H. Jensen, “Single-Channel Online Enhancement of Speech Corrupted by Reverberation and Noise,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 572–587, Mar. 2017. [Online]. Available: https://ieeexplore.ieee.org/document/7795155/

[38] S. Gannot, D. Burshtein, and E. Weinstein, “Signal enhancement using beamforming and nonstationarity with applications to speech,” *IEEE Transactions on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.

[39] E. Habets, J. Benesty, I. Cohen, S. Gannot, and J. Dmochowski, “New insights into the MVDR beamformer in room acoustics,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 158–170, Jan. 2010.

[40] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B. Juang, "Blind speech dereverberation with multi-channel linear prediction based on short time fourier transform representation," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, Mar. 2008, pp. 85–88.

[41] B. Li, T. N. Sainath, A. Narayanan, J. Caroselli, M. Bacchiani, A. Misra, I. Shafran, H. Sak, G. Pundak, K. Chin, K. C. Sim, R. J. Weiss, K. W. Wilson, E. Variani, C. Kim, O. Siohan, M. Weintraub, E. McDermott, R. Rose, and M. Shannon, "Acoustic Modeling for Google Home," in *Interspeech 2017*. ISCA, Aug. 2017, pp. 399–403. [Online]. Available: http://www.isca-speech.org/archive/Interspeech_2017/abstracts/0234.html

[42] M. Wu and D. Wang, "A one-microphone algorithm for reverberant speech enhancement," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, vol. 1. IEEE, 2003, pp. I–892. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1198925

[43] J. Erkelens, J. Jensen, and R. Heusdens, "A data-driven approach to optimizing spectral speech enhancement methods for various error criteria," *Speech Communication*, vol. 49, no. 7-8, pp. 530–541, Jul. 2007. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0167639306001427

[44] T. Fingscheidt, S. Suhadi, and S. Stan, "Environment-optimized speech enhancement," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 4, pp. 825–834, 2008. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4472220

[45] I. Tashev and M. Slaney, "Data driven suppression rule for speech enhancement," in *Information Theory and Applications Workshop (ITA), 2013*, Feb. 2013, pp. 1–6.

[46] Y. Hu and P. C. Loizou, "Environment-specific noise suppression for improved speech intelligibility by cochlear implant users," *The Journal of the Acoustical Society of America*, vol. 127, no. 6, pp. 3689–3695, Jun. 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2896410/

[47] K. Kokkinakis, O. Hazrati, and P. C. Loizou, "A channel-selection criterion for suppressing reverberation in cochlear implants," *The Journal of the Acoustical Society of America*, vol. 129, no. 5, pp. 3221–3232, May 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3108395/

[48] C. H. Taal, R. C. Hendriks, and R. Heusdens, “Matching pursuit for channel selection in cochlear implants based on an intelligibility metric,” in *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European.* IEEE, 2012, pp. 504–508. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6334221

[49] Y. Bengio, I. J. Goodfellow, and A. Courville, “Deep learning,” 2015, book in preparation for MIT Press. [Online]. Available: http://www.iro.umontreal.ca/~bengioy/dlbook

[50] I. Sutskever, O. Vinyals, and Q. V. V. Le, “Sequence to sequence learning with neural networks,” in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 3104–3112. [Online]. Available: http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf

[51] F. J. Weninger, F. Eyben, and B. Schuller, “Single-channel speech separation with memory-enhanced recurrent neural networks,” in *Proceedings 39th IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, 2014. [Online]. Available: http://www.mmk.ei.tum.de/publ/pdf/14/14wen4.pdf

[52] R. Pascanu, T. Mikolov, and Y. Bengio, “On the difficulty of training recurrent neural networks,” in *Proceedings of The 30th International Conference on Machine Learning*, 2013, pp. 1310–1318. [Online]. Available: http://jmlr.org/proceedings/papers/v28/pascanu13.html

[53] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, bibtex: Hochreiter1997. [Online]. Available: http://www.mitpressjournals.org/doi/abs/10.1162/neco.1997.9.8.1735

[54] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, “Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).* Doha, Qatar: Association for Computational Linguistics, 2014, pp. 1724–1734. [Online]. Available: http://aclweb.org/anthology/D14-1179

[55] Y. LeCun and Y. Bengio, “Convolutional networks for images, speech, and time series,” *The handbook of brain theory and neural networks*, vol. 3361, no. 10, 1995. [Online]. Available: http://www.iro.umontreal.ca/~lisa/pointeurs/handbook-convo.pdf

[56] O. Abdel-Hamid, L. Deng, and D. Yu, “Exploring convolutional neural network structures and optimization techniques for speech recognition.” in *INTERSPEECH*, 2013, pp. 3366–3370. [Online]. Available: http://research-srv.microsoft.com/pubs/200804/CNN-Interspeech2013_pub.pdf

[57] C. v. d. Malsburg and W. Schneider, “A neural cocktail-party processor,” *Biological Cybernetics*, vol. 54, no. 1, pp. 29–40, May 1986. [Online]. Available: http://link.springer.com/article/10.1007/BF00337113

[58] D. Wang, “Auditory stream segregation based on oscillatory correlation,” in *Neural Networks for Signal Processing [1994] IV. Proceedings of the 1994 IEEE Workshop.* IEEE, 1994, pp. 624–632. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=366003

[59] Y. Wang and D. Wang, “Towards Scaling Up Classification-Based Speech Separation,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1381–1390, Jul. 2013.

[60] Y. Jiang, D. Wang, R. Liu, and Z. Feng, “Binaural classification for reverberant speech segregation using deep neural networks,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 2112–2121, Dec. 2014.

[61] Y. Wang, A. Narayanan, and D. Wang, “On training targets for supervised speech separation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.

[62] Y. Wang and D. Wang, “A deep neural network for time-domain signal reconstruction,” in *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing, to appear*, vol. 14, 2015, p. 138. [Online]. Available: http://web.cse.ohio-state.edu/~dwang/papers/Wang-Wang.icassp15.pdf

[63] E. Grais, M. Sen, and H. Erdogan, “Deep neural networks for single channel source separation,” in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 3734–3738.

[64] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, “Deep learning for monaural speech separation,” in *Acoustics, Speech and Signal Processing (ICASSP), 2014*

IEEE International Conference on. IEEE, 2014, pp. 1562–1566. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6853860

[65] S. Araki, T. Hayashi, M. Delcroix, M. Fujimoto, K. Takeda, and T. Nakatani, "Exploring multi-channel features for denoising-autoencoder-based speech enhancement," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2015, pp. 116–120.

[66] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for joint enhancement of magnitude and phase," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2016, pp. 5220–5224.

[67] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," *arXiv:1609.03499 [cs]*, Sep. 2016, arXiv: 1609.03499. [Online]. Available: http://arxiv.org/abs/1609.03499

[68] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, apr 2018. [Online]. Available: https://doi.org/10.1109%2Ficassp.2018.8461368

[69] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, apr 2018. [Online]. Available: https://doi.org/10.1109%2Ficassp.2018.8462417

[70] Y. Luo and N. Mesgarani, "TaSNet: Time-domain audio separation network for real-time, single-channel speech separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, apr 2018. [Online]. Available: https://doi.org/10.1109%2Ficassp.2018.8462116

[71] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech enhancement generative adversarial network," in *Interspeech 2017*. ISCA, aug 2017. [Online]. Available: https://doi.org/10.21437%2Finterspeech.2017-1428

[72] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 2672–2680. [Online]. Available: http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf

[73] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Lecture Notes in Computer Science.* Springer International Publishing, 2015, pp. 234–241. [Online]. Available: https://doi.org/10.1007%2F978-3-319-24574-4_28

[74] Y. C. Subakan and P. Smaragdis, "Generative adversarial source separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, apr 2018. [Online]. Available: https://doi.org/10.1109%2Ficassp.2018.8461671

[75] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," *arXiv:1701.07875 [cs, stat]*, Jan. 2017, arXiv: 1701.07875. [Online]. Available: http://arxiv.org/abs/1701.07875

[76] J. F. Santos and T. H. Falk, "Speech Dereverberation With Context-Aware Recurrent Neural Networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 7, pp. 1236–1246, Jul. 2018.

[77] T. N. Sainath, R. J. Weiss, K. W. Wilson, A. Narayanan, and M. Bacchiani, "Factored spatial and spectral multichannel raw waveform CLDNNs," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on.* IEEE, 2016, pp. 5075–5079. [Online]. Available: http://ieeexplore.ieee.org/abstract/document/7472644/

[78] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, "A Regression Approach to Speech Enhancement Based on Deep Neural Networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, Jan. 2015.

[79] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, Apr. 1984.

[80] B. Wu, K. Li, M. Yang, and C. H. Lee, "A study on target feature activation and normalization and their impacts on the performance of DNN based speech dereverberation systems," in *2016*

Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), Dec. 2016, pp. 1–4.

[81] ——, "A Reverberation-Time-Aware Approach to Speech Dereverberation Based on Deep Neural Networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 102–111, Jan. 2017.

[82] A. Keshavarz, S. Mosayyebpour, M. Biguesh, T. A. Gulliver, and M. Esmaeili, "Speech-Model Based Accurate Blind Reverberation Time Estimation Using an LPC Filter," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1884–1893, Aug. 2012. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6171837

[83] X. Xiao, S. Zhao, D. H. H. Nguyen, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "Speech dereverberation for enhancement and recognition using dynamic features constrained deep neural networks and feature adaptation," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, p. 4, 2016. [Online]. Available: http://link.springer.com/article/10.1186/s13634-015-0300-4

[84] M. M. S. S. T. Kawahara, "Speech dereverberation using long short-term memory," 2015. [Online]. Available: https://pdfs.semanticscholar.org/b382/72d3b17c43452862943c40376e8fc2ad5eb5.pdf

[85] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, "On the Properties of Neural Machine Translation: Encoder–Decoder Approaches," in *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation.* Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 103–111. [Online]. Available: https://www.aclweb.org/anthology/W14-4012

[86] J. Eaton, N. Gaubitch, A. Moore, and P. Naylor, "The ACE challenge #x2014; Corpus description and performance evaluation," in *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2015, pp. 1–5, bibtex: eaton_ace_2015.

[87] M. Karjalainen, P. Ansalo, A. Mäkivirta, T. Peltonen, and V. Välimäki, "Estimation of modal decay parameters from noisy response measurements," *Journal of the Audio Engineering Society*, vol. 50, no. 11, pp. 867–878, 2002. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=11059

[88] E. A. Lehmann and A. M. Johansson, “Diffuse reverberation model for efficient image-source simulation of room impulse responses,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 6, pp. 1429–1439, 2010. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5299028

[89] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, “TIMIT acoustic phonetic continuous speech corpus LDC93S1 (web download),” 1993.

[90] A. Paszke, S. Gross, S. Chintala, and G. Chanan, “PyTorch.” [Online]. Available: https://www.pytorch.org

[91] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *International Conference for Learning Representations*, 2015.

[92] ITU-T P.862, “Perceptual evaluation of speech quality: An objective method for end-to-end speech quality assessment of narrow-band telephone network and speech coders,” ITU Telecommunication Standardization Sector (ITU-T), Tech. Rep., 2001.

[93] ITU-T P. 863, “Perceptual Objective Listening Quality Assessment (POLQA),” ITU Telecommunication Standardization Sector (ITU-T), Tech. Rep., 2011.

[94] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time–frequency weighted noisy speech,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 7, pp. 2125–2136, 2011.

[95] J. F. Santos, M. Senoussaoui, and T. H. Falk, “An updated objective intelligibility estimation metric for normal hearing listeners under noise and reverberation,” in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, September 2014, pp. 55–59.

[96] T. H. Falk, C. Zheng, and W.-Y. Chan, “A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1766–1774, Sep. 2010.

[97] B. Cauchi, H. Javed, T. Gerkmann, S. Doclo, S. Goetze, and P. Naylor, “Perceptual and instrumental evaluation of the perceived level of reverberation,” in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2016, pp. 629–633.

[98] Santos, J. F. and Falk, T. H., "Supplementary materials for the paper "Speech dereverberation using context-aware recurrent neural networks"," accessed on 24 July 2017. [Online]. Available: https://figshare.com/s/c639aed9049d5de00c1b

[99] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and others, "DeepSpeech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014. [Online]. Available: http://arxiv.org/abs/1412.5567

[100] A. Graves, "Generating sequences with recurrent neural networks," *arXiv preprint arXiv:1308.0850*, 2013. [Online]. Available: http://arxiv.org/abs/1308.0850

[101] D. Williamson and D. Wang, "Time-Frequency Masking in the Complex Domain for Speech Dereverberation and Denoising," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. PP, no. 99, pp. 1–1, 2017.

[102] M. Senoussaoui, J. F. Santos, and T. H. Falk, "Speech temporal dynamics fusion approaches for noise-robust reverberation time estimation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2017, pp. 5545–5549.

[103] F. Mayer, D. S. Williamson, P. Mowlaee, and D. Wang, "Impact of phase estimation on single-channel speech separation based on time-frequency masking," *The Journal of the Acoustical Society of America*, vol. 141, no. 6, pp. 4668–4679, Jun. 2017. [Online]. Available: http://asa.scitation.org/doi/abs/10.1121/1.4986647

[104] A. Jukic, T. van Waterschoot, T. Gerkmann, and S. Doclo, "A general framework for incorporating time–frequency domain sparsity in multichannel speech dereverberation," *Journal of the Audio Engineering Society*, vol. 65, no. 1/2, pp. 17–30, 2017.

[105] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training Very Deep Networks," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 2377–2385. [Online]. Available: http://papers.nips.cc/paper/5850-training-very-deep-networks.pdf

[106] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2016. [Online]. Available: https://doi.org/10.1109%2Fcvpr.2016.90

[107] S. I. Mimilakis, K. Drossos, J. F. Santos, G. Schuller, T. Virtanen, and Y. Bengio, “Monaural singing voice separation with skip-filtering connections and recurrent inference of time-frequency mask,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, apr 2018. [Online]. Available: https://doi.org/10.1109%2Ficassp.2018.8461822

[108] J. Thiemann, N. Ito, and E. Vincent, “The diverse environments multi-channel acoustic noise database: A database of multichannel environmental noise recordings,” *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 3591–3591, May 2013. [Online]. Available: http://scitation.aip.org/content/asa/journal/jasa/133/5/10.1121/1.4806631

[109] A. Varga and H. J. M. Steeneken, “Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems,” *Speech Communication*, vol. 12, no. 3, pp. 247–251, Jul. 1993. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0167639393900953

[110] D. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, May 2015.

[111] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “A short-time objective intelligibility measure for time-frequency weighted noisy speech,” in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on.* IEEE, 2010, pp. 4214–4217. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5495701

[112] T. N. Sainath, R. J. Weiss, A. Senior, K. W. Wilson, and O. Vinyals, “Learning the speech front-end with raw waveform cldnns,” *Proc. Interspeech*, 2015. [Online]. Available: http://www.ee.columbia.edu/~ronw/pubs/interspeech2015-waveform_cldnn.pdf

[113] T. N. Sainath, R. J. Weiss, K. W. Wilson, A. Narayanan, M. Bacchiani, and Andrew, “Speaker location and microphone spacing invariant acoustic modeling from raw multichannel waveforms,” in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Dec. 2015, pp. 30–36.

[114] M. Ravanelli and Y. Bengio, “Speaker recognition from raw waveform with SincNet,” in *2018 IEEE Spoken Language Technology Workshop (SLT).* IEEE, dec 2018. [Online]. Available: https://doi.org/10.1109%2Fslt.2018.8639585

[115] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, “Quantized Neural Networks: Training Neural Networks with Low Precision Weights and Activations,” p. 30.

[116] G. Hinton, O. Vinyals, and J. Dean, “Distilling the Knowledge in a Neural Network,” *NIPS 2014 Deep Learning Workshop*, Dec. 2014.

[117] Y. LeCun, J. S. Denker, and S. A. Solla, “Optimal Brain Damage,” p. 8, 1990.

[118] M. Zhu and S. Gupta, “To prune, or not to prune: exploring the efficacy of pruning for model compression,” *2017 NIPS Workshop on Machine Learning of Phones and other Consumer Devices*, Dec. 2017.

[119] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Improving neural networks by preventing co-adaptation of feature detectors,” *CoRR*, vol. abs/1207.0580, 2012. [Online]. Available: http://arxiv.org/abs/1207.0580

[120] “Project Jupyter.” [Online]. Available: https://www.jupyter.org

[121] C. Valentini-Botinhao, “Noisy speech database for training speech enhancement algorithms and TTS models,” *http://parole.loria.fr/DEMAND/*, Aug. 2017. [Online]. Available: https://datashare.is.ed.ac.uk/handle/10283/2791

[122] C. Veaux, J. Yamagishi, and K. MacDonald, “CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit,” *The Rainbow Passage which the speakers read out can be found in the International Dialects of English Archive: (http://web.ku.edu/~idea/readings/rainbow.htm).*, Apr. 2017. [Online]. Available: https://datashare.is.ed.ac.uk/handle/10283/2651

[123] J. Thiemann, N. Ito, and E. Vincent, “The diverse environments multi-channel acoustic noise database: A database of multichannel environmental noise recordings,” *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 3591–3591, May 2013. [Online]. Available: http://scitation.aip.org/content/asa/journal/jasa/133/5/10.1121/1.4806631

[124] A. Ardakani, Z. Ji, and W. J. Gross, “Learning to skip ineffectual recurrent computations in LSTMs,” in *2019 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, mar 2019. [Online]. Available: https://doi.org/10.23919%2Fdate.2019.8714765

[125] A. See, M.-T. Luong, and C. D. Manning, “Compression of neural machine translation models via pruning,” in *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning.* Association for Computational Linguistics, 2016. [Online]. Available: https://doi.org/10.18653%2Fv1%2Fk16-1029

[126] S. Han, J. Pool, J. Tran, and W. Dally, “Learning both Weights and Connections for Efficient Neural Network,” p. 9, 2015.

[127] A. N. Gomez, I. Zhang, K. Swersky, Y. Gal, and G. E. Hinton, “Learning Sparse Networks Using Targeted Dropout,” *arXiv:1905.13678 [cs, stat]*, May 2019, arXiv: 1905.13678. [Online]. Available: http://arxiv.org/abs/1905.13678

[128] G. Leclerc, M. Vartak, R. C. Fernandez, T. Kraska, and S. Madden, “Smallify: Learning Network Size while Training,” *arXiv:1806.03723 [cs, stat]*, Jun. 2018, arXiv: 1806.03723. [Online]. Available: http://arxiv.org/abs/1806.03723

[129] K. J. Han, A. Chandrashekaran, J. Kim, and I. Lane, “The CAPIO 2017 Conversational Speech Recognition System,” *arXiv:1801.00059 [cs]*, Dec. 2017, arXiv: 1801.00059. [Online]. Available: http://arxiv.org/abs/1801.00059

[130] G. Kurata, B. Ramabhadran, G. Saon, and A. Sethy, “Language modeling with highway LSTM,” in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU).* IEEE, dec 2017. [Online]. Available: https://doi.org/10.1109%2Fasru.2017.8268942

[131] G. Saon, G. Kurata, T. Sercu, K. Audhkhasi, S. Thomas, D. Dimitriadis, X. Cui, B. Ramabhadran, M. Picheny, L.-L. Lim, B. Roomi, and P. Hall, “English conversational telephone speech recognition by humans and machines,” in *Interspeech 2017.* ISCA, aug 2017. [Online]. Available: https://doi.org/10.21437%2Finterspeech.2017-405

[132] F. B. Gelderblom, T. V. Tronstad, and E. M. Viggen, “Subjective Intelligibility of Deep Neural Network-Based Speech Enhancement,” *1968-1972*, 2017. [Online]. Available: https://brage.bibsys.no/xmlui/handle/11250/2467464

[133] M. Kolbæk, Z. H. Tan, and J. Jensen, “Speech Intelligibility Potential of General and Specialized Deep Neural Network Based Speech Enhancement Systems,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 153–167, Jan. 2017.

[134] K. Kinoshita, M. Delcroix, S. Gannot, E. A. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, and others, “A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research,” *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, pp. 1–19, 2016. [Online]. Available: http://link.springer.com/article/10.1186/s13634-016-0306-6

[135] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, “An Experimental Study on Speech Enhancement Based on Deep Neural Networks,” *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, Jan. 2014.

[136] K. Paliwal, K. Wójcicki, and B. Shannon, “The importance of phase in speech enhancement,” *Speech Communication*, vol. 53, no. 4, pp. 465–494, Apr. 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639310002086

[137] N. Mohammadiha, P. Smaragdis, and A. Leijon, “Supervised and unsupervised speech enhancement using nonnegative matrix factorization,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2140–2151, 2013.

[138] I.-T. P.563, “Single ended method for objective speech quality assessment in narrow-band telephony applications,” Intl. Telecom Union, Tech. Rep., 2004.

[139] J. H. Hansen and B. L. Pellom, “An effective quality evaluation protocol for speech enhancement algorithms,” in *Fifth International Conference on Spoken Language Processing*, 1998.

[140] Y. Hu and P. C. Loizou, “Evaluation of Objective Quality Measures for Speech Enhancement,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, Jan. 2008. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4389058

[141] I.-T. P.862, “Perceptual evaluation of speech quality: An objective method for end-to-end speech quality assessment of narrow-band telephone network and speech coders,” ITU-T, Tech. Rep., 2001.

[142] T. H. Falk and W.-Y. Chan, “A non-intrusive quality measure of dereverberated speech,” in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2008. [Online]. Available: http://musaelab.ca/pdfs/C24.pdf

[143] J. F. Santos, S. Cosentino, O. Hazrati, P. C. Loizou, and T. H. Falk, “Objective speech intelligibility measurement for cochlear implant users in complex listening environments,” *Speech Communication*, vol. 55, no. 7–8, pp. 815–824, Sep. 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639313000435

[144] M. Cartwright, B. Pardo, G. J. Mysore, and M. Hoffman, “Fast and easy crowdsourced perceptual audio evaluation,” in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2016, pp. 619–623.

[145] M. Morise and Y. Watanabe, “Sound quality comparison among high-quality vocoders by using re-synthesized speech,” *Acoustical Science and Technology*, vol. 39, no. 3, pp. 263–265, May 2018. [Online]. Available: https://www.jstage.jst.go.jp/article/ast/39/3/39_E1779/_article

[146] S. W. Group, “Speech Signal Processing Toolkit (SPTK).” [Online]. Available: http://sp-tk.sourceforge.net/

[147] Z. Wu, O. Watts, and S. King, “Merlin: An Open Source Neural Network Speech Synthesis System,” Sep. 2016, pp. 202–207. [Online]. Available: http://www.isca-speech.org/archive/SSW_2016/abstracts/ssw9_PS2-13_Wu.html

[148] R. Yamamoto, “r9y9/pysptk: A python wrapper for Speech Signal Processing Toolkit (SPTK).” [Online]. Available: https://github.com/r9y9/pysptk/

[149] A. Camacho and J. G. Harris, “A sawtooth waveform inspired pitch estimator for speech and music,” *The Journal of the Acoustical Society of America*, vol. 124, no. 3, pp. 1638–1652, Sep. 2008. [Online]. Available: https://asa.scitation.org/doi/abs/10.1121/1.2951592

[150] D. Talkin, “REAPER,” Jun. 2019, original-date: 2014-12-22T23:30:40Z. [Online]. Available: https://github.com/google/REAPER

[151] B. S. Lee, “Noise Robust Pitch Tracking by Subband Autocorrelation Classification,” Ph.D. dissertation, Columbia University, 2012. [Online]. Available: https://doi.org/10.7916/D8SJ1SPJ

[152] Y. Zhao, B. Xu, R. Giri, and T. Zhang, “Perceptually Guided Speech Enhancement Using Deep Neural Networks,” in *2018 IEEE International Conference on Acoustics, Speech and*

Signal Processing (ICASSP). Calgary, AB: IEEE, Apr. 2018, pp. 5074–5078. [Online]. Available: https://ieeexplore.ieee.org/document/8462593/

[153] W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, and A. Stolcke, "The microsoft 2017 conversational speech recognition system," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, apr 2018. [Online]. Available: https://doi.org/10.1109%2Ficassp.2018.8461870

[154] N. Zheng and X.-L. Zhang, "Phase-Aware Speech Enhancement Based on Deep Neural Networks," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 27, no. 1, pp. 63–76, Jan. 2019. [Online]. Available: https://doi.org/10.1109/TASLP.2018.2870742

[155] S. Wisdom, J. R. Hershey, K. Wilson, J. Thorpe, M. Chinen, B. Patton, and R. A. Saurous, "Differentiable Consistency Constraints for Improved Deep Speech Enhancement," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 900–904.

[156] X. Du, M. Zhu, X. Shi, X. Zhang, W. Zhang, and J. Chen, "End-to-End Model for Speech Enhancement by Consistent Spectrogram Masking," *arXiv:1901.00295 [cs, eess]*, Jan. 2019, arXiv: 1901.00295. [Online]. Available: http://arxiv.org/abs/1901.00295

[157] C. Olah, A. Satyanarayan, I. Johnson, S. Carter, L. Schubert, K. Ye, and A. Mordvintsev, "The Building Blocks of Interpretability," *Distill*, vol. 3, no. 3, p. e10, Mar. 2018. [Online]. Available: https://distill.pub/2018/building-blocks