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

ENVIRONMENT-AWARE KNOWLEDGE DISTILLATION FOR IMPROVED RESOURCE-CONSTRAINED EDGE SPEECH RECOGNITION

By

Arthur Santos Pimentel

A thesis submitted in conformity with the requirements for the degree of
Master of Science, M.Sc.
in Telecommunications

Evaluation Committee

External evaluator	Prof. Alessandro L. Koerich École de Technologie Supérieure
Internal evaluator	Prof. Douglas O'Shaughnessy INRS-EMT
Research supervisor	Prof. Tiago H. Falk INRS-EMT

Acknowledgements

I would like to thank my supervisor, Prof. Tiago H. Falk, for all the support and guidance throughout my academic journey and for the expertise and insights that have been instrumental in shaping the direction of this thesis. I would also like to acknowledge Prof. Anderson Avila, whose insightful discussions have played a crucial role in developing this project.

I would like to express my deepest gratitude to my family, who always supports and encourages me. Their love and dedication have been a constant source of inspiration, and for that, I am truly thankful.

I thank my colleagues, whose collaboration has been a cornerstone of this academic pursuit. A special mention goes to Heitor Guimarães, whose dedication and collaborative approach have left an indelible mark on this project.

To my circle of friends, both in and outside of academia, I am grateful for the unwavering support, understanding, and moments of levity that have lightened the challenges along the way. Your friendship has been a cornerstone of my personal and academic life.

Résumé

Les avancées récentes dans l'apprentissage auto-supervisé ont permis aux systèmes de reconnaissance automatique de la parole (ASR) d'atteindre l'etat de l'art des taux d'erreurs de reconnaissance de mots (WER) tout en ne nécessitant qu'une fraction des données étiquetées nécessaires à leurs prédécesseurs. Néanmoins, bien que de tels modèles atteignent des résultats de pointe dans des scénarios d'entraînement/test correspondants, leurs performances se dégradent considérablement lorsqu'ils sont testés dans des conditions non vues. Pour palier à ce problème, des stratégies telles que l'augmentation de données et/ou l'adaptation au domaine ont été explorées. Cependant, les modèles disponibles sont encore trop volumineux pour être considérés pour des applications vocales sur des appareils aux ressources limitées ; ainsi, des outils de compression de modèle tels que l'élagage de paramètres, la quantification de poids et la distillation de connaissances sont nécessaires.

Dans notre étude sur l'apprentissage de la représentation de la parole auto-supervisée (S3RL), nous abordons d'abord les effets de non-correspondance d'entraînement/test sur les modèles compressés, en investiguant l'impact de la quantification des paramètres et de l'élagage du modèle sur le modèle robust wav2vec 2.0 dans des conditions bruyantes, réverbérantes et mixtes. De plus, nous améliorons la recette de distillation DistilHuBERT avec des têtes de prédiction optimisées, un enrichissement du jeu de données spécifique pour différents environnements, et un estimateur d'environnement en temps réel pour la sélection du modèle lors de l'inférence. Des expériences sur l'ensemble de données LibriSpeech, corrompu par différents niveaux de bruit et de réverbération, démontrent qu'une diversité de données pendant l'entraînement renforce la robustesse du modèle contre la compression, le bruit et la réverbération. Notre méthode proposée surpassent les modèles de références jusqu'à 48,4% et 89,2% dans le taux de réduction d'erreurs de mots dans des conditions extrêmes, tout en réduisant les paramètres de 50%.

Mots-clés reconnaissance automatique de la parole ; élagage des paramètres ; quantification des poids ; distillation des connaissances ; apprentissage auto-supervisé ; spectre de modulation ; sensibilité au contexte

Abstract

Recent advances in self-supervised learning have allowed automatic speech recognition (ASR) systems to achieve state-of-the-art (SOTA) word error rates (WER) while requiring only a fraction of the labeled data needed by its predecessors. Notwithstanding, while such models achieve SOTA results in matched train/test scenarios, their performance degrades substantially when tested in unseen conditions. To overcome this problem, strategies such as data augmentation and/or domain adaptation have been explored. Available models, however, are still too large to be considered for edge speech applications on resource-constrained devices; thus, model compression tools, such as parameter pruning, weight quantization and knowledge distillation, are needed.

In our study on self-supervised speech representation learning (S3RL), we first address train/test mismatch effects on compressed models, investigating parameter quantization and model pruning impact on robust wav2vec 2.0 under noisy, reverberant, and mixed conditions. Additionally, we enhance the DistilHuBERT distillation recipe with optimized prediction heads, targeted data augmentation for various environments, and a real-time environment estimator for model selection during inference. Experiments on the LibriSpeech dataset, corrupted with different noise and reverberation levels, demonstrate that diverse data during training enhances model robustness against compression, noise, and reverberation, with our proposed method outperforming benchmarks by up to 48.4% and 89.2% in word error reduction rate under extreme conditions, while reducing parameters by 50%.

Keywords automatic speech recognition; parameter pruning; weight quantization; knowledge distillation; self-supervised learning; modulation spectrum; context awareness

Contents

List of Figures

List of Tables

List of Acronyms and Abbreviations

ASR	Automatic speech recognition
BIRD	Big Impulse Response Dataset
CNN	Convolutional neural network
DCASE	Detection and Classification of Acoustic Scenes and Events
MAC	Multiply–accumulate
QAT	Quantization-aware training
RBF	Radial basis function
RIR	Room impulse response
RT60	Reverberation time
S3RL	Self-supervised speech representation learning
SNR	Signal to noise ratio
SOTA	State-of-the-art
SRMR	Speech-to-reverberation modulation energy ratio
SUPERB	Speech Processing Universal Performance Benchmark
SVM	Support vector machine
WADA	Waveform Amplitude Distribution Analysis
WER	Word error rate

SYNTHÈSE: DISTILLATION DE CONNAISSANCES SENSIBLE À L'ENVIRONNEMENT POUR UNE RECONNAISSANCE VOCALE AMÉLIORÉE EN BORDURE DE RESEAU AVEC CONTRAINTES DE RESSOURCES

0.1 Introduction

Les grands modèles d'apprentissage profond ont récemment connu un grand succès dans les tâches de reconnaissance de la parole (Ao *et al.*, 2021; Babu *et al.*, 2021). Cependant, ces modèles utilisent une quantité considérable de ressources computationnelles, ce qui peut être irréalisable pour de nombreuses applications en périphérie. Ces applications se concentrent sur le traitement des données au plus proche de leur source afin de réduire la latence et l'utilisation de la bande passante. Cela peut être particulièrement important pour les applications de reconnaissance vocale, où des données vocales privées et/ou sensibles peuvent nécessiter d'être envoyées sur le réseau pour être traitées à distance sur de grands clusters de traitement de données hébergeant des modèles très

grands et complexes. Déployer des modèles aussi grand en périphérie peut être un défi, car certains dispositifs en périphérie peuvent avoir des capacités de stockage et de traitement limités. De plus, les applications en périphérie sont affectées par plusieurs facteurs environnementaux, tels que le bruit ambiant et/ou la réverbération de la pièce, qui sont connus pour être préjudiciables aux applications basées sur la parole. Ainsi, une étude plus détaillée sur l'impact de la compression de modèle et de l'efficacité de l'inférence pour les grands modèles de reconnaissance vocale est nécessaire. Notre étude a pour objectif de répondre à cette problematique.

Automatic Speech Recognition (ASR) vise à convertir un signal vocal continu en une représentation textuelle discrète. Alors que la parole est l'une des méthodes de communication les plus efficaces pour les humains (O'shaughnessy, 1987), le texte est une représentation importante pour les machines, et un grand nombre de techniques peuvent être plus facilement appliquées pour structurer et comprendre les données. Récemment, de grands modèles d'apprentissage profond ont connu un grand succès en ASR (Baevski *et al.*, 2020b; Hsu *et al.*, 2021a; Chen *et al.*, 2022; Ao *et al.*, 2021; Babu *et al.*, 2021; Radford *et al.*, 2023), avec des méthodes qui rivalisent avec la reconnaissance vocale humaine dans une grande variété d'environnements acoustiques (Spille *et al.*, 2018).

L'apprentissage de la représentation de la parole auto-supervisée (S3RL) s'est imposé comme une force motrice derrière ces innovations. Avec ce type d'apprentissage, l'objectif principal est de générer des représentations informatives à partir de vastes ensembles de données non étiquetées, en tirant parti de leurs caractéristiques multimodales. Par la suite, le modèle est affiné avec des données étiquetées. Aujourd'hui, wav2vec 2.0 (Baevski *et al.*, 2020b), HuBERT (Hsu *et al.*, 2021a), et WavLM (Chen *et al.*, 2022) constituent les représentations generalisées de la parole les plus largement déployées, atteignant des résultats de pointe non seulement pour l'ASR, mais aussi pour d'autres tâches telles que la reconnaissance des locuteurs et des émotions (Feng *et al.*, 2023).

Bien que ces modèles représentent une percée en termes de performances, il existe toujours un écart à combler entre les systèmes ASR basés sur de grands modèles déployés sur le cloud et les systèmes ASR destinés au déploiement en périphérie.

Cependant, les méthodes universelles existantes de S3RL présentent deux limitations significatives dans le contexte des applications en périphérie : (i) leur taille importante, ce qui peut être irréalisable pour de nombreuses applications en périphérie ; et (ii) leur robustesse, au moment de l'inférence, face à des conditions environnementales inconnues, telles que le bruit et la réverbération.

Par exemple, le modèle HuBERT a généralement entre 95 millions et 1 milliard de paramètres, ce qui est prohibitif sur des dispositifs en périphérie avec une capacité de stockage et de traitement limitée. Des techniques de compression de modèle, telles que la quantification, l'élagage de modèle et la distillation des connaissances, ont été explorées, la première produisant les résultats les plus prometteurs, comme le montre le récent développement du modèle "DistilHuBERT" (Chang *et al.*, 2022).

Concernant la robustesse environnementale, des études antérieures ont montré que l'utilisation de techniques d'amélioration basées sur le signal est généralement insuffisante, car les distorsions introduites par ces algorithmes peuvent également dégrader les performances du modèle (par exemple, Kshirsagar *et al.* (2023)). Les niveaux de bruit et de réverbération non vus, par exemple, sont connus pour réduire considérablement l'exactitude même des systèmes ASR de pointe (Zhang *et al.*, 2023) et peuvent être très sensibles à différentes conditions environnementales (Pimentel *et al.*, 2023a; Li *et al.*, 2022). Bien que les techniques d'adaptation au domaine, telles que celles proposées dans "Robust HuBERT" (Huang *et al.*, 2022b) et "deHuBERT" (Ng *et al.*, 2023), puissent atténuer ce problème, elles ne sont pas directement applicables à la compression.

Dans ce travail, nous présentons deux contributions au domaine de S3RL. Tout d'abord, nous explorons les effets que les conditions de divergence entre les jeux d'entraînement et de test ont sur l'exactitude de la reconnaissance de la parole basée sur des modèles compressés de parole auto-supervisée. En particulier, nous rendons compte des effets que la quantification des paramètres et l'élagage du modèle ont sur l'exactitude de la reconnaissance de la parole basée sur le modèle résilient appelé wav2vec 2.0 dans des conditions bruyantes, réverbérantes et mixtes. Nos résultats montrent que l'entraînement avec une plus grande diversité de données améliore significativement la robustesse des modèles de reconnaissance de la parole non seulement contre le bruit et la réverbération, mais aussi face à la réduction de la taille du modèle. Ensuite, nous proposons trois innovations en plus du protocole de distillation DistilHuBERT existante : optimiser les têtes de prédiction, utiliser une méthode d'enrichissement des données ciblée pour différents scénarios environnementaux, et utiliser un estimateur de conditions environnementales en temps réel pour la sélection adaptative de modèles compressés en inférence. Des expériences avec l'ensemble de données LibriSpeech, contaminé par différents types de bruit et niveaux de réverbération, montrent que la méthode proposée surpasse plusieurs méthodes de référence, à la fois originales et compressées, jusqu'à 48,4% et 89,2% de réduction du taux d'erreur de reconnaissance de mots dans des conditions extrêmement bruyantes

et réverbérantes, respectivement, tout en réduisant de 50% le nombre de paramètres. Ainsi, la méthode proposée est bien adaptée aux applications de reconnaissance vocale en périphérie dans un contexte de ressources disponibles limitées.

0.2 Fondement théorique

0.2.1 Compression

La compression de modèle consiste à transformer un modèle volumineux et gourmand en ressources en une version compacte adaptée au stockage sur des appareils mobiles à capacités limitées. De plus, elle peut impliquer l'optimisation du modèle pour une exécution plus rapide avec une latence minimale ou l'atteinte d'un compromis entre ces objectifs (Zhu *et al.*, 2023)

La quantification consiste à compresser le réseau d'origine en réduisant le nombre de bits nécessaires pour représenter chaque paramètres du modèle (Cheng *et al.*, 2018). Ce concept peut également être étendue à la propagation du gradient et à l'activation vers une forme quantifiée. Les paramètres dans un réseau neuronal sont généralement stockés sous forme de nombres flottants sur 32 bits. Ces paramètres peuvent être quantifiés en 16 bits, 8 bits, 4 bits ou 1 bit. La quantification peut être effectuée pendant l'entraînement, appelée entraînement conscient de la quantification (QAT) (Nagel *et al.*, 2022), ou après l'entraînement (quantification post-entraînement (Liu *et al.*, 2021)).

L'élagage du modèle vise à réduire le nombre d'opérandes. Il peut réduire le nombre d'accès à la mémoire et les opérations de calcul, permettant ainsi d'augmenter la vitesse d'inference (Deng *et al.*, 2020). Dans les DNN, de nombreux paramètres sont redondants et ont une faible contribution à la reduction de l'erreur pendant l'entraînement et sur la généralisation du modèle. La suppression de ces paramètres redondant après l'entraînement du réseau n'aura qu'un impact negligeable sur l'exactitude du modèle. Le principal objectif de l'élagage est de rendre le modèle d'apprentissage profond plus compact pour son utilisation dans un stockage à capacité réduite (Choudhary *et al.*, 2020b). Les formes les plus courantes d'élagage sont l'élagage des poids et l'élagage des neurones, où le premier réduit le nombre de connexions, tandis que le second réduit le nombre de nœuds dans un réseau neuronal intégralement interconnecté.

La Distillation de Connaissances (DC) est une technique largement utilisée pour transférer les connaissances d'un modèle volumineux (enseignant) à un modèle plus petit (élève) afin d'atteindre tous types d'efficacité. La DC nécessite généralement la conception d'une fonction de perte pour minimiser la distance de la sortie ou des caractéristiques intermédiaires entre l'élève et l'enseignant (Xu & McAuley, 2023).

Le processus de DC peut être hors ligne, en ligne, ou par auto-distillation. La distillation hors ligne extrait les connaissances du modèle enseignant pré-entraîné, tandis que la distillation en ligne extrait pendant que les modèles enseignant et élève sont en cours d'entraînement. Dans l'auto-distillation, le réseau élève est formé progressivement en utilisant ses propres connaissances sans nécessiter un modèle enseignant pré-entraîné (Ghimire *et al.*, 2022).

0.2.2 Apprentissage de la Représentation de la Parole Auto-Supervisé

S3RL permet aux modèles de réseaux de neurones profonds de capturer des facteurs importants et distinctement isolés à partir des formes d'onde vocales brutes (Guimarães *et al.*, 2023). S3RL peut être considéré comme un cas particulier de l'apprentissage non supervisé, car les deux schémas apprennent sans annotations. Bien que les méthodes non supervisées conventionnelles reposent sur des objectifs de reconstruction ou d'estimation de densité, les approches S3RL reposent sur des tâches préliminaires qui exploitent la modalité des données utilisée pour l'entraînement. Bien que les méthodes d'apprentissage supervisé aient tendance à apprendre des caractéristiques plus fortes que les approches d'apprentissage non supervisé, elles nécessitent un travail coûteux et chronophage d'étiquetage de données de la part des annotateurs humains. Les techniques S3RL visent à réunir le meilleur des deux mondes: former un extracteur de caractéristiques puissant en utilisant un apprentissage discriminatif sans avoir besoin d'une annotation manuelle des exemples d'entraînement (Ericsson *et al.*, 2022). Les représentations vocales universelles apprises peuvent ensuite être utilisées dans de nombreuses tâches (Mohamed *et al.*, 2022).

Dans le contexte de S3RL, les « modèles en amont » sont conçus pour la préformation et l'apprentissage de représentations générales à partir de données non étiquetées ; essentiellement, ils agissent comme des « extracteurs de caractéristiques » pour les tâches en aval. L'objectif est de capturer des caractéristiques et des motifs de haut niveau dans les données vocales en entrée. Ces modèles sont formés sur un grand ensemble de données sans étiquettes spécifiques à la tâche. À

leur tour, les « modèles en aval » sont des modèles spécifiques à la tâche affinés sur des données étiquetées pour une tâche vocale particulière, telle que la reconnaissance vocale ou la classification des émotions. Les modèles en aval sont formés sur un ensemble de données plus petit avec des étiquettes spécifiques à leur tâche. Le modèle en amont pré-formé est donc affiné sur ces données pour adapter ses représentations apprises aux exigences spécifiques de la tâche cible.

Le benchmark Speech processing Universal Performance Benchmark (SUPERB) a été proposé dans (Yang *et al.*, 2021) dans le but d'offrir à la communauté une base d'essai standard et complète pour évaluer la généralisabilité des modèles pré-formés sur diverses tâches couvrant tous les aspects de la parole.

0.2.3 Modèles de Reconnaissance de la Parole

Le modèle *wav2vec 2.0* (Baevski *et al.*, 2020a) apprend des unités vocales de base utilisées pour résoudre une tâche auto-supervisée. L'architecture se compose d'un encodeur de caractéristiques convolutionnel à plusieurs couches qui prend en entrée un signal audio brut et produit des représentations vocales latentes à chaque interval temporel, qui sont ensuite transmises à un réseau contextuel. L'encodeur se compose de plusieurs blocs contenant une convolution temporelle suivie d'une normalisation de couche (Ba *et al.*, 2016) et d'une fonction d'activation GELU (Hendrycks & Gimpel, 2016), tandis que le réseau contextuel est composé de couches de Transformer empilées, créant des représentations contextualisées de l'ensemble de la séquence.

La couche d'encodeur du réseau d'encodeur se compose de six couches convolutionnelles empilées. Le composant suivant important est le réseau contextuel, composé d'un réseau Transformer de 12 couches pour le modèle de base et d'un réseau de 24 couches pour le modèle large. Les représentations fournies par le réseau contextuel sont les caractéristiques d'entrée pour les applications subséquentes en aval. La propriété contrastive se produit dans la fonction de perte associée à cette approche. L'idée est de comparer des échantillons (d'où la notion de propriété contrastive) et de distinguer un échantillon futur d'un ensemble d'échantillon distracteurs échantillonnés à partir d'une distribution uniforme p_n. En utilisant les représentations du réseau contextuel, il est possible d'entraîner des modèles pour des tâches en aval en utilisant moins de données étiquetées tout en conservant de bonnes performances. De plus, la fonction de perte dans ce modèle inclut à la fois la fonction de perte intègre à la fois une composante contrastive et une composante visant à minimiser

la perte de diversité du codebook acquis par le module de quantification. Cette dernière est réalisée par la maximisation de l'entropie selon la distribution Gumbel-Softmax, favorisant l'utilisation par le modèle d'une variété étendue de codes du codebook, évitant ainsi sa concentration sur un nombre restreint de codes.

Le modèle *robust wav2vec 2.0* (Hsu *et al.*, 2021b) est une variante récente développée pour offrir une robustesse améliorée contre les changements de domaine (par exemple, dus au bruit, à des ensembles de données variables, ou d'autres facteurs) au moment du test. Utilisant la même architecture que son prédécesseur, robust wav2vec utilise des données de domaine cible lors de la phase de pré-entraînement, conduisant ainsi à des améliorations significatives des performances en reconnaissance vocale hors domaine.

Le modèle HuBERT (Hsu *et al.*, 2021a) représente une technique de préformation vocale conçue pour apprendre des représentations vocales efficaces au moyen de la reconstruction masquée de caractéristiques. Cette architecture se compose de deux composants majeurs : l'encodeur et le réseau contextuel. Le réseau d'encodeur fonctionne comme une fonction de mappage de l'échantillon de forme d'onde brut à une représentation de caractéristiques intermédiaires. Il est structuré comme un réseau de convolution 1D, comprenant sept couches avec 512 cartes de caractéristiques. Les tailles de noyau et de pas pour chaque couche varient en conséquence. Après le réseau de convolution, une fonction d'activation non linéaire GELU (Hendrycks & Gimpel, 2016) est appliquée. Le réseau contextuel joue un rôle crucial dans l'architecture HuBERT. Son principe fondamental consiste à mapper des caractéristiques diverses en un vecteur contextuel unifié à travers plusieurs encodeurs Transformer, capturant ainsi des informations à longue portée. Le lecteur intéressé peut se référer à (Vaswani *et al.*, 2017) pour plus de détails sur l'architecture Transformer.

Différentes versions du modèle HuBERT ont été développées, notamment les versions HuBERT *Base*, *Large*, et *X-Large*. La principale différence entre elles réside dans la taille de leurs réseaux Transformer respectifs, comprenant 12, 24 et 48 couches, respectivement. De plus, il est à noter que le modèle HuBERT Base a été formé sur un ensemble de données comprenant 960 heures du jeu de données LibriSpeech (Panayotov *et al.*, 2015), tandis que les versions Large et X-Large ont été formées sur un ensemble de données nettement plus important composé de 60 000 heures du jeu de données LibriLight (Kahn *et al.*, 2020).

Bien que les versions plus grandes de HuBERT aient montré une précision accrue en reconnaissance vocale dans des conditions bruyantes (Guimarães *et al.*, 2023), pour les applications en périphérie, de tels modèles de grande taille peuvent poser problème, tandis que des modèles plus petits peuvent être plus sensibles au bruit. À cette fin, les auteurs de (Huang *et al.*, 2022b) ont proposé une méthode appelée *Robust HuBERT*, où l'entraînement adversarial de domaine a été utilisé pour rendre le système plus robuste aux facteurs environnementaux. Plus spécifiquement, un discriminateur de domaine est responsable de classer la source des distorsions appliquées à l'extrait vocale. Le modèle HuBERT Base a été continuellement formé sur des données plus variées, avec une probabilité de recevoir des données vocales contaminées par différents types et niveaux de bruit, y compris le bruit gaussien et des bruits enregistrés provenant de la base de données MUSAN (Snyder *et al.*, 2015). De plus, une autre technique pour augmenter la robustesse est la désentrelacement du bruit, comme le montre (Ng *et al.*, 2023). Les auteurs proposent *deHuBERT*, un nouveau cadre d'entraînement auto-supervisé qui encourage l'invariance au bruit dans les représentations contextuelles intégrées de HuBERT en introduisant une deuxième intégration à partir de signaux augmentés de bruit différents, en utilisant un encodeur CNN partagé et une nouvelle paire de fonctions objectifs de perte auxiliaires.

Un sujet émergent dans le traitement vocal est celui de l'ASR en périphérie. Pour ces applications, des modèles plus petits sont nécessaires, et la distillation des connaissances a donc été explorée récemment. Le travail récent dans (Chang *et al.*, 2022), par exemple, a proposé le modèle appelé *DistilHuBERT*, où les représentations cachées d'un modèle HuBERT étaient directement distillées pour réduire la taille du modèle et le temps d'inférence.

Plus précisément, DistilHuBERT propose un nouveau cadre enseignant-élève pour l'apprentissage des représentations vocales par distillation de connaissances multitâches. Le modèle se compose d'un extracteur de caractéristiques CNN et d'un encodeur Transformer de taille reduite. L'idée est d'apprendre à générer plusieurs représentations cachées de l'enseignant à partir de représentations partagées. Cela se fait en prédisant les représentations cachées de l'enseignant avec des têtes de prédiction distinctes. Cet objectif est un paradigme d'apprentissage multitâche et encourage l'encodeur Transformer à produire des représentations compactes pour plusieurs têtes de prédiction. DistilHuBERT prend ensuite un modèle HuBERT Base pré-entraîné figé et utilise uniquement trois têtes de prédiction pour prédire respectivement les sorties des 4^{e}, 8^{e} et 12^{e} couches cachées de HuBERT. Avant la préformation, l'élève est initialisé avec les paramètres de l'enseignant. Ensuite, les têtes de

prédiction de l'élève apprennent à générer les représentations cachées de l'enseignant en minimisant la fonction de perte $\mathcal{L}^{(l)} = \mathcal{L}^{(4)} + \mathcal{L}^{(8)} + \mathcal{L}^{(12)}$. Après l'entraînement, les têtes sont retirées, car le paradigme d'apprentissage multitâche force le modèle DistilHuBERT à apprendre des représentations contenant des informations riches. Ces couches, cependant, sont choisies pour maximiser les performances dans diverses tâches liées à la parole et peuvent ne pas être optimales pour l'ASR. Dans cette étude, nous explorons comment différentes têtes de prédiction peuvent améliorer les performances du modèle.

Une autre méthode récemment proposée pour des représentations universelles distillées nommée *RobustDistiller* est décrite dans (Guimarães *et al.*, 2023). Il s'agit d'une méthode qui combine enrichissement de données et apprentissage multitâche de débruitage incorporés dans le processus de distillation des connaissances. Dans l'étape d'enrichissement des données, une contamination en ligne des données est effectuée pendant le processus de distillation. En particulier, le modèle élève reçoit les données bruyantes en entrée, mais l'objectif du réseau est de reconstruire les représentations propres du modèle enseignant. Ainsi, dans l'étape d'apprentissage multitâche de débruitage, outre l'apprentissage pour reconstruire les représentations de l'enseignant, une tête d'amélioration supplémentaire est responsable de reconstruire la forme d'onde vocale propre à partir des représentations apprises. Contrairement aux techniques d'amélioration habituelles, l'objectif est de contraindre le modèle amont à porter suffisamment d'informations sur la parole elle-même et non sur les composants de bruit. Bien que RobustDistiller ait montré des performances supérieures à DistilHuBERT sur des tâches de repérage de mots-clés, de classification des intentions et de reconnaissance des émotions, avec du bruit et/ou de la réverbération, démontrant, tout de même, une certaine sensibilité à différentes conditions environnementales.

Enfin, nous explorons si la distillation d'un modèle enseignant conçu pour être robuste aboutit également à un modèle compressé robuste. Il s'agit d'une étape importante pour décider si la robustesse doit être mise en œuvre au niveau de l'enseignant ou pendant le processus de distillation, comme proposé ici. À cette fin, nous appliquons le même processus de distillation utilisé dans DistilHuBERT, mais nous l'appliquons au modèle enseignant *Robust HuBERT* décrit dans la Section 2.2.2. Par la suite, nous appelons ce modèle *DistilRobustHuBERT*.

0.2.4 Scénarios en Conditions Réelles

Le bruit, dans le contexte de l'acoustique, est tout son indésirable, soit désagréable, soit interférant avec d'autres sons écoutés. En électronique et en théorie de l'information, le bruit fait référence à des signaux aléatoires, imprévisibles et indésirables, ou à des changements dans les signaux, qui masquent le contenu informationnel souhaité. Un type de bruit pertinent est le 'bruit blanc'. Défini comme un signal ou un son complexe qui couvre l'ensemble des fréquences audibles, toutes ayant une intensité égale (Britannica, 2013). Le bruit blanc est généralement ajouté aux enregistrements de signaux vocaux propres, ainsi qu'à d'autres bruits environnementaux, pour simuler la manière dont ce signal serait capté dans des conditions réelles. Le bruit environnemental fait référence aux sons indésirables causés par l'activité humaine, tels que le bruit de la route, des avions, du trafic ferroviaire, les bruits de chantier de construction, des installations de bâtiments (par exemple, la ventilation ou le refroidissement), ou le bruit du voisinage (par exemple, musique forte, terrains de sport) (NCCEH, 2022).

La réverbération est la persistance du son après son arrêt en raison de multiples réflexions provenant de surfaces à l'intérieur d'un espace clos. Ces réflexions augmentent progressivement à chaque nouvelle réflexion et décroissent progressivement en intensité à mesure qu'elles sont absorbées par les surfaces des objets dans l'espace clos. Le temps de réverbération est un paramètre important pour caractériser la qualité d'un espace auditif. Les sons dans des environnements réverbérants sont sujets à la coloration. Cela affecte l'intelligibilité de la parole et la localisation du son. Historiquement, le temps de réverbération a été désigné comme le temps RT60, qui est le temps nécessaire pour que le son décroisse à 60 dB en dessous de sa valeur à l'arrêt (Ratnam *et al.*, 2003).

Dans ce travail, nous utilisons l'estimateur de rapport signal sur bruit nomée WADA, un algorithme initialement proposé par (Kim & Stern, 2008). Cet algorithme fournit une correspondance directe entre un signal et une valeur estimée du rapport signal sur bruit (RSB). Il suppose que la distribution d'amplitude de la parole propre peut être approximée par la distribution Gamma avec un paramètre de mise en forme de 0,4, et que le signal est contaminé par un bruit additif avec une distribution gaussienne. Même avec ces hypothèses, les expériences realisées par les auteurs de WADA montrent que l'algorithme se comporte bien lorsqu'il est testé pour trois types de bruit différents : bruit additif gaussien blanc, segments musicaux de la base de données DARPA HUB 4 Broadcast News, et bruit d'un locuteur interférant unique. Les signaux de bruit ont été ajoutés

artificiellement aux signaux vocaux à différents RSB allant de -10 dB à 30 dB. Nous avons choisi d'utiliser cet algorithme en raison de son efficacité computationnelle et de sa facilité d'utilisation.

Afin d'estimer le temps de réverbération d'un signal sans connaître les propriétés de la pièce dans laquelle il a été enregistré, nous adoptons la méthode décrite par (Falk & Chan, 2010) et (Falk *et al.*, 2007). Les auteurs étudient l'utilisation de l'information sur la dynamique temporelle pour la mesure à l'aveugle des paramètres acoustiques d'une pièce. L'information sur la dynamique à long terme est obtenue par l'analyse spectrale des enveloppes temporelles de la parole, un processus communément appelé traitement du spectre de modulation. Dans cette étude, nous nous appuyons sur la métrique d'énergie de modulation parole-réverbération (SRMR) décrite dans (Falk & Chan, 2010). Le SRMR est inversement proportionnel au RT60, des valeurs plus basses indiquant des niveaux de réverbération plus élevés. De plus, nous utilisons un algorithme de machine à vecteurs de support (SVM) pour fournir une correspondance entre l'inverse du SRMR et le RT60.

0.3 Quantification et Élagage des Modèles pour les Tâches de Reconnaissance de la Parole en Conditions "Réelles"

0.3.1 Introduction

Initialement, le modèle évalué proposé, le robust wav2vec 2.0 (Hsu *et al.*, 2021b), a été évalué en termes de performances dans différentes conditions hors domaine. Cependant, les niveaux de bruit variables, une condition importante dans les applications embarquées, n'ont pas été explorés de manière exhaustive. Nous visons ici à combler cette lacune et à évaluer l'impact que la compression des modèles peut avoir. Dans nos expériences, des modèles pré-entraînés et affinés de la plateforme *Hugging Face* ont été utilisés. Plus précisément, le modèle wav2vec 2.0 est pré-entraîné et affiné sur 960 heures du jeu de données Librispeech, tandis que le robuste wav2vec 2.0 est pré-entraîné sur les ensembles de données Libri-Light, Common Voice, Switchboard et Fisher, puis affiné sur 960 heures du jeu de données Librispeech.

0.3.2 Méthodes et Matériaux

Deux techniques classiques de compression de modèle sont explorées pour évaluer le potentiel des applications de reconnaissance de la parole embarquées. La première est la quantification, où le nombre de bits requis pour stocker chaque paramètres est réduit, réduisant ainsi considérablement la taille du modèle, économisant de la mémoire et accélérant le calcul. La méthode peut également être étendue pour représenter le gradient et l'activation sous forme quantifiée (**?**). Ici, nous explorons l'impact de la quantification sur 8 bits sur toutes les couches linéaires des modèles de parole et le comparons à sa version originale sur 32 bits (c'est-à-dire un taux de compression de 4). Ensuite, l'élagage du modèle est exploré, où les paramètres redondants peuvent être supprimés du réseau avec un impact minimal sur l'exactitude du modèle (Choudhary *et al.*, 2020a). Un élagage global non structuré basé sur la plus basse norme L1 a été utilisé à cinq taux d'élagage différents, de 10 à 30% par intervalles de 5%.

Pour entraîner nos modèles, nous utilisons le corpus LibriSpeech (Panayotov *et al.*, 2015) en tant qu'ensemble de données d'extraits vocaux d'énoncés intelligibles. Il est composé de 960 heures d'enregistrements de livres audio avec une fréquence d'échantillonnage de 16 kHz, dérivés du projet LibriVox.

0.3.3 Bruit Additif et Réverbération

Comme nous nous intéressons à comprendre l'impact des modèles de parole compressés dans des conditions en bordure, nous utilisons les signaux de bruit présents dans le jeu de données Deep Noise Suppression Challenge 4 (DNS4) (Dubey *et al.*, 2022) pour altérer les signaux de parole de test de l'ensemble de données LibriSpeech. Cet ensemble de bruits se compose de 180 heures de bruit, présent dans 62 000 extraits vocaux, couvrant 150 types de bruits différents distincts de la parole. Ces fichiers sont ajoutés aux échantillons de test à cinq niveaux de rapport signal sur bruit (SNR) différents, allant de 0 dB à 20 dB par intervalles de 5 dB. La réverbération est simulée en convoluant les signaux clairs avec une réponse impulsionnelle de pièce échantillonnée de manière uniforme (RIR). Les ensembles de données openSLR28 avec 248 RIR réelles et openSLR26 avec 60 000 RIR synthétiques sont utilisés (Ko *et al.*, 2017) pour simuler des pièces de petite et moyenne taille. Enfin, la réverbération et le bruit sont simulés en combinant les deux étapes précédentes. Dans tous les cas, si nécessaire, les formes d'onde sont rééchantillonnées à 16 kHz.

0.3.4 Résultats et Discussion

Tout d'abord, nous explorons l'impact de la quantification sur les versions originales et robustes de wav2vec 2.0 dans des conditions expérimentales contrôlées. Les deux modèles ont obtenu un WER de 3.2% avec des poids stockés en précision flottante 32 bits (taille totale du modèle de 1262 Mo). Après la quantification, le WER a augmenté à 3.3%, soit une légère augmentation pour une compression du modèle par 4 (taille totale du modèle de 354.5 Mo). Ensuite, nous explorons l'impact de la taille des poids des couches convolutionnelles et linéaires des modèles. On observe que la version robuste du modèle de traitement de la parole n'a montré qu'une légère augmentation du WER à 3.3% avec un taux de taille de 0.3, tandis que sa version d'origine est passée à 5.7%.

Ensuite, nous explorons l'impact des niveaux variables de bruit et de la réverbération. Les figures 3.1a et 3.1b montrent les graphiques WER des modèles originaux et quantifiés, en fonction du SNR et de la taille de la pièce, respectivement. Comme on peut le voir, la quantification sur 8 bits a montré un impact minimal sur les performances du modèle pour le cas bruyant et un faible impact en présence de réverbération (par exemple, le WER pour robust wav2vec est passé de 0.049 à 0.052 pour la condition de taille de pièce moyenne). Ensuite, nous effectuons une analyse approfondie du WER obtenu pour différents types de bruit. Le tableau 3.1 montre le WER moyen pour des fichiers audio contaminés par six types de bruit courants, à savoir les sons domestiques (par exemple, le bruit d'aspirateur), la voix humaine, la musique, le bruit des véhicules, le bruit du vent et d'autres sources diverses. Comme on peut le voir, la voix humaine, les sons domestiques et le bruit des véhicules ont montré la plus grande détérioration des performances. Ce sont des conditions dans lesquelles les applications de péripherie sont généralement rencontrées, comme les enceintes intelligentes et la reconnaissance vocale embarquée. Dans l'ensemble, le modèle robuste de wav2vec 2.0 surpasse le wav2vec 2.0 d'origine pour tous les types de bruit, avec la meilleure correspondance obtenue avec le bruit du vent.

Ensuite, nous explorons la robustesse des modèles à la taille. Les figures 3.2a et 3.2b montrent les graphiques WER en fonction du taux de taille et du SNR ou du taux de taille et de la taille de la pièce, respectivement. Comme on peut le voir, la taille des deux modèles a affecté les WER, surtout pour des SNR inférieurs à 15 dB, la version originale de wav2vec 2.0 montrant la plus grande détérioration, en particulier avec des taux de taille supérieurs à 20%. La réverbération, à son tour, a montré un impact minimal sur la version taille robuste, mais a eu un impact substantiel

sur wav2vec 2.0, en particulier pour les pièces de taille moyenne et des taux de taille supérieurs à 15%.

Enfin, nous évaluons la robustesse des deux modèles à la taille avec ajout de bruit et de réverbération combinés dans les signaux de test. La figure 3.3 montre le WER en fonction du taux de taille et de la taille de la pièce, où les niveaux de bruit ont été moyennés sur la plage de 0 à 20 dB. Là encore, le modèle robuste s'est révélé insensible aux taux de taille accrus, mais sensible aux dégradations elles-mêmes. Par exemple, à un SNR de 5 dB, le modèle robuste a obtenu un WER de 10.1% avec un taux de taille de 0.3. Cette erreur est passée à 25.5% à un SNR de 0 dB. Cependant, ce résultat est nettement meilleur que ce qui a été montré avec le modèle original de wav2vev 2.0 qui, dans les mêmes conditions de compression et de bruit, a obtenu un WER de 62.2%.

Dans l'ensemble, les expériences décrites ici suggèrent que les schémas de compression existants par quantification et taille semblent bien adaptés aux applications de reconnaissance vocale embarquée lorsqu'ils sont appliqués dans des conditions relativement propres. Dans de tels scénarios, des taux de compression aussi élevés que 4 pourraient être obtenus avec un impact minimal sur le WER. En revanche, si les applications de test impliquent des conditions bruyantes et/ou réverbérantes, des représentations vocales améliorées sont toujours nécessaires, au-delà de ce qui peut être atteint par le modèle robuste wav2vec considéré comme résilient. La distillation des connaissances sensible à l'environnement peut être une solution possible.

0.3.5 Conclusions

Dans ce chapitre, nous évaluons la robustesse de deux modèles vocaux ASR SOTA, à savoir wav2vec 2.0 et robust wav2vec 2.0, dans des conditions bruyantes et réverbérantes non vues lorsque les modèles sont compressés via des schémas de quantification et de taille. En particulier, la quantification sur 8 bits et la taille mondiale non structurée basée sur la norme L1 ont été explorées. On a constaté que, bien que la quantification et la taille aient un impact minimal sur le WER dans des conditions propres, le bruit et la réverbération entraînent une dégradation significative du WER, même avec des modèles construits de manière inhérente pour être robustes à de telles conditions. Les travaux futurs devraient explorer une compression plus robuste et des représentations auto-supervisées avant que les applications de reconnaissance vocale embarquée ne puissent être déployées "en condition réelle".

0.4 Distillation de Connaissances Sensibles à l'Environnement pour la Reconnaissance de la Parole en Conditions "Réelles"

0.4.1 Introduction

Des travaux récents ont commencé à proposer des solutions intégrant conjointement la compression et la robustesse environnementale (par exemple, (Guimarães *et al.*, 2023; Huang *et al.*, 2022a)). Cependant, ces solutions n'ont pas encore été explorées pour la reconnaissance automatique de la parole (ASR) et ont montré une certaine sensibilité aux conditions environnementales variables.

Dans ce chapitre, nous abordons cette problématique. Plus précisément, notre objectif global est triple : (1) adapter la représentation existante de DistilHuBERT (Chang *et al.*, 2022) pour la rendre mieux adaptée aux tâches de l'ASR, (2) l'exploiter la technique d'enrichissement du jeu de données pour accroître la robustesse des modèles compressés face à des conditions non vues, et (3) proposer une solution hiérarchique consciente de l'environnement où des modèles compressés optimisés pour différentes conditions environnementales sont choisis durant la phase d'inférence, rendant ainsi les modèles compressés plus robustes aux conditions environnementales variables généralement rencontrées dans des environnements spécifiques.

0.4.2 Modèles Proposés

Notre première innovation consiste à adapter le protocole DistilHuBERT pour optimiser les têtes de prédiction pour la tâche ASR. Le protocole DistilHuBERT originale utilisait comme têtes de prédiction les 4^{e}, 8^{e}, et 12^{e} couches du modèle enseignant HuBERT Base (Chang *et al.*, 2022). Bien que ces couches aient montré une précision améliorée pour différentes tâches, elles n'étaient pas nécessairement optimales pour l'ASR. D'après les instructions de SUPERB, après le pré-entraînement, les états cachés de différentes couches en amont sont pondérés, sommés, et transmis aux couches spécifiques à la tâche, les poids de chaque couche changeant en fonction de la tâche en aval. Un poids plus élevé indique une plus grande contribution de la couche correspondante. Nous avons analysé les poids de chaque couche du modèle HuBERT et constaté que les couches 8, 9 et 10 apportaient la plus grande contribution à la tâche ASR. Ainsi, notre méthode proposée s'appuiera sur ces trois

couches. La Figure 4.1 montre l'analyse des poids des couches du modèle HuBERT de base, affiné pour l'ASR.

Ensuite, nous adaptons le protocole RobustDistiller proposée par (Guimarães *et al.*, 2023), comme illustré dans la Figure 4.2. Au moment de l'entraînement, étant donné un lot d'extraits vocaux d'énoncés intelligibles, nous échantillonnons une action à appliquer à chaque énoncé dans le lot : (i) aucune modification n'est apportée à l'énoncé d'entraînement ; (ii) contamination de l'énoncé avec soit du bruit additif avec un rapport signal/bruit choisi de manière aléatoire entre $[0, 30]$ dB ou convolution de la forme d'onde de la parole avec une réponse impulsionnelle de salle sélectionnée de manière aléatoire. Les probabilités d'échantillonnage des scénarios (i) et (ii) sont de 30% et 70%, respectivement.

Enfin, nous proposons une troisième innovation, à savoir l'utilisation de la sensibilité à l'environnement. Plus précisément, différents modèles compressés sont obtenus, chacun optimisé pour une condition environnementale distincte. Lors de l'inférence, le meilleur modèle est sélectionné et utilisé pour l'ASR. Cette approche hiérarchique permet à chaque modèle compressé de jouer le rôle d'expert pour un scénario environnemental donné. Bien que l'augmentation du nombre de modèles utilisés réduise les gains de compression globaux obtenus avec la distillation, ici nous explorons l'utilisation de seulement deux modèles, réalisant ainsi toujours une certaine compression par rapport au modèle enseignant original. De plus, comme un seul modèle est utilisé lors de l'inférence, les gains de temps d'inférence ne sont pas affectés en utilisant deux modèles. Les figures 4.3a et 4.3b représentent deux modèles que nous explorons ici, l'un caractérisant les niveaux de bruit dans l'environnement et l'autre les niveaux de réverbération, respectivement.

Comme le montrent les figures, pour une inférence en temps réel, un sélecteur de niveau de bruit/réverbération est nécessaire. Comme nous supposons que l'accès à des signaux de référence propres n'est pas disponible, une mesure « aveugle » est nécessaire. Ici, nous explorons l'usage d'un estimateur de rapport signal/bruit (SNR) nommée Analyse de la Distribution d'Amplitude de Forme d'Onde (WADA) et décrit dans (Kim & Stern, 2008). Pour l'estimation du temps de réverbération (RT60), nous nous appuyons sur la métrique du rapport d'énergie de modulation parole-réverbération (SRMR), suivi d'une mise en correspondance de la métrique SRMR avec le temps de réverbération (RT60) via un régresseur à vecteur de support, comme décrit dans (Falk & Chan, 2010).

0.4.3 Configuration Expérimentale

Pour entraîner nos modèles, nous utilisons une fois de plus le corpus LibriSpeech (Panayotov *et al.*, 2015) comme ensemble de données d'extraits vocaux d'énoncés intelligibles. Comme nous sommes intéressés par la réalisation d'une enrichissement de données et l'évaluation de la sensibilite environnementale, nous utilisons des signaux sonores présents dans les ensembles de données MUSAN (Snyder *et al.*, 2015) et UrbanSound8K (Salamon *et al.*, 2014) pour altérer les signaux du corpus LibriSpeech lors de l'étape d'entraînement. Ces ensembles de données contiennent environ 15 heures d'enregistrements dans une grande variété de catégories. UrbanSound8K contient 8732 extraits sonores étiquetés de sons urbains provenant de 10 classes distinctes. De plus, nous utilisons la partie bruit de MUSAN, qui contient 929 fichiers de différents types de bruit. Tous les extraits vocaux sont échantillonnés à 16 kHz.

Un ensemble de données de réponses impulsionnelles de salle (RIR) est également utilisé, à savoir le Big Impulse Response Dataset (BIRD) (Grondin *et al.*, 2020), composé de réponses impulsionnelles de salle simulées correspondant à des salles de différentes tailles et coefficients d'absorption, avec des valeurs de RT60 allant de 140 ms à 1 seconde. L'ensemble d'entraînement utilisé pour la réverbération se compose d'environ 35 000 RIR simulées échantillonnées à partir de l'ensemble de données BIRD. La moitié de ces échantillons ont un temps de réverbération faible (RT60 inférieur à 500 ms) et l'autre moitié ont un temps de réverbération élevé (RT60 supérieur à 500 ms).

Durant la phase de test, en plus de l'ensemble de tests LibriSpeech, deux ensembles de données supplémentaires sont utilisés pour tester les performances du modèle dans des conditions inconnues. Le premier est la "Detection and Classification of Acoustic Scenes and Events" (DCASE) de 2020, soit DCASE2020 (Mesaros *et al.*, 2018). Le jeu de données se compose de 64 heures d'enregistrements audio dans 10 scènes acoustiques. Pour la réverbération, nous utilisons un sous-ensemble différent de l'ensemble de données BIRD composé d'environ 4 000 réponses impulsionnelles de salle simulées. Encore une fois, les signaux sont répartis également entre des valeurs de RT60 faibles et élevées.

Pour évaluer les avantages de notre méthodologie proposée, différentes expériences sont réalisées. Tout d'abord, nous avons sélectionné HuBERT Base comme notre modèle enseignant pour le processus de distillation. Par la suite, nous appliquons nos modèles proposés. Le modèle appelé *Robust DistilHuBERT* est une variation proposée du modèle DistilHuBERT avec seulement les deux

premières innovations mises en œuvre et sert à évaluer les avantages supplémentaires de la prise de conscience du bruit. Ce modèle est soit robuste au bruit de 0 à 30 dB, soit à la réverbération de 140 ms à 1 s. Pendant ce temps, notre modèle proposé avec les trois modifications mises en œuvre est appelé *Noise-Aware DistilHuBERT* ou *Reverb-Aware DistilHuBERT*. Ils sont composés du pipeline représenté dans les figures 4.3a et 4.3b, respectivement. De plus, HuBERT Large (Hsu *et al.*, 2021a), HuBERT Base (Hsu *et al.*, 2021a), Robust HuBERT (Huang *et al.*, 2022b), DistilHuBERT (Chang *et al.*, 2022), DistilRobustHuBERT et RobustDistiller (Guimarães *et al.*, 2023) sont utilisés comme modèles de référence.

Dans toutes les expériences, les modèles amont sont entraînés à l'aide d'un seul GPU NVidia A100. Notre méthode de distillation robuste et l'étape d'affinement du modèle pour la tâche aval ASR prennent environ 30 heures chacune pour être terminées, soit un total de 60 heures de temps d'entraînement. Nous utilisons l'optimiseur AdamW, avec un lot de 24 extraits vocaux, pour 200 000 itérations, tandis qu'après 14 000 mises à jour, le taux d'apprentissage décroît linéairement de 2×10^{-4} à zéro.

Nous utilisons le taux d'erreur de mots (WER) comme métrique pour évaluer la performance du modèle (Huang *et al.*, 2014). Le WER est défini par l'équation 4.1. Il est important de noter que, bien que le WER soit généralement présenté comme une valeur comprise entre 0 et 100, il ne représente pas un pourcentage réel. Alors qu'un WER de zéro signifie une estimation parfaite, cette métrique n'est pas limitée à une borne supérieure, et une séquence avec plus d'insertions que de mots corrects aura un WER supérieur à 100.

0.4.4 Résultats Expérimentaux et Discussion

Tout d'abord, nous devons valider si les estimateurs proposés de niveaux de bruit et de réverbération sont précis. La Figure 4.4 montre le graphique de dispersion entre les valeurs estimées et réelles du SNR. Une corrélation globale de 82.5% est obtenue. Cependant, il est important de noter que notre approche se base sur deux modèles compressés, l'un optimisé pour des niveaux de SNR faibles et l'autre pour des niveaux élevés. L'utilisation de l'algorithme WADA pour détecter les signaux dans ces deux classes donne une exactitude totale de 94.2%, suggérant que le modèle est suffisamment précis pour le déploiement. La Figure 4.5, quant à elle, montre le graphique de dispersion entre l'inverse du SRMR et le vrai RT60. Une corrélation totale de 82.2% a été obtenue. Encore une fois,

comme la solution proposée utilise des modèles optimisés pour des niveaux de réverbération faibles et élevés, seul un classificateur binaire est nécessaire. Des expériences utilisant un classificateur SVM standard avec un noyau à fonction radiale (RBF) ont abouti à une précision de classification de 88.9%, indiquant à nouveau que le modèle est suffisamment précis pour le déploiement.

Le tableau 4.1 compare les méthodes proposées aux cinq algorithmes de référence en termes de nombre de paramètres du modèle, de nombre d'opérations de multiplication–accumulation (MACs) et du WER obtenu sur les fichiers de test propres et bruyants. Comme on peut le voir, les deux premières innovations (ligne 'Robust DistilHuBERT') apportent déjà une amélioration substantielle par rapport au modèle DistilHubert original. Globalement, des gains relatifs de 33.04, 31.16 et 22.40% sont obtenus pour les types de bruit en intérieur, en extérieur et dans les transports, respectivement. En incorporant les trois innovations proposées (ligne 'Noise-aware DistilHuBERT'), une légère diminution du WER est obtenue. On peut également observer que les solutions proposées donnent des modèles compressés qui atteignent un WER similaire quel que soit le type de bruit, suggérant une robustesse améliorée aux facteurs ambiants et une applicabilité aux conditions de bord.

Pour explorer davantage les avantages de la solution proposée de sensibilite contextuelle, nous nous concentrons ensuite sur les conditions de SNR plus faibles, connues pour être les plus impactantes pour l'ASR. Le tableau 4.2 rapporte les précisions obtenues pour les références et les solutions proposées uniquement pour les fichiers bruyants contaminés par un bruit additif allant de 0 à 10 dB de SNR. Comme on peut le voir, les gains obtenus avec la solution proposée de prise en compte du bruit surpassent le modèle Robust DistilHuBERT de 6.19, 4.24 et 4.62%, pour les types de bruit intérieur, extérieur et de transport respectivement. Par rapport au DistilHuBERT original, ces gains relatifs sont de 48.42, 46.42 et 37.85%, respectivement. Il est important de souligner que bien que la solution dote de sensibilite contextuelle du bruit nécessite le stockage du double du nombre de paramètres par rapport à DistilHuBERT et RobustDistiller, le temps d'inférence et les exigences de calcul restent les mêmes, car un seul des deux modèles est utilisé à la fois. Ainsi, les gains obtenus peuvent toujours être utiles pour des applications de bord impliquant des conditions très bruyantes.

Enfin, le tableau 4.3 compare les WER obtenus pour les solutions de référence et proposées pour les signaux vocaux en conditions réverbérantes. Comme on peut le voir, la réverbération est une

distorsion plus difficile à traiter et le WER du modèle distillé de référence est gravement dégradé. La mise en œuvre des deux premières innovations (Robust DistilHuBERT) permet de réduire le WER de 66.52% au total et de 88.87 et 55.27% pour les conditions de RT60 élevé et faible, respectivement, par rapport à DistilHuBERT. La solution proposée sensible à l'environnement diminue encore le WER de 2.55, 2.76 et 2.87%, respectivement. Fait intéressant, les solutions proposées avec seulement 24 M ou 48 M de paramètres améliorent déjà de manière significative l'exactitude de l'ASR par rapport à HuBERT Large avec 300 M de paramètres. Pour les conditions de niveau de réverbération élevé, par exemple, la solution sensible a l'environnement proposée a un gain de 70.75% par rapport à HuBERT Large tout en nécessitant qu'approximativement un sixième du nombre de paramètres.

0.5 Conclusions et Travaux Futurs

Dans cette étude, nous présentons l'état de l'apprentissage automatique de la représentation de la parole auto-supervisée, spécifiquement dans le contexte de la reconnaissance automatique de la parole pour les appareils limités en ressources. Nous explorons ses capacités et limitations et proposons des innovations dans le but de réduire la taille du modèle et le nombre de paramètres, tout en augmentant sa robustesse aux conditions environnementales.

Nous explorons d'abord la robustesse de deux modèles de pointe et examinons l'impact des données additionnelles sur les performances des modèles compressés. Nous testons nos modèles avec deux méthodes de compression, à savoir la quantification et l'élagage des paramètres. Ces modèles compressés sont évalués avec deux niveaux contamination au bruit, spécifiquement, le bruit additif et la réverbération. Nos résultats montrent que bien que l'entraînement de décalage de domaine ait un impact limité sur les modèles compressés, le bruit et la réverbération ont toujours des effets significatifs sur les performances.

De plus, nous proposons trois innovations du protocole de distillation de DistilHuBERT. Nous menons des expériences approfondies et comparons avec six modèles de référence. Lorsqu'ils sont évalués dans des conditions bruyantes, les modèles proposés surpassent les références de taille comparable jusqu'à 33.04%. Les gains sont plus importants dans des conditions très bruyantes, où des gains allant jusqu'à 48.42% ont été observés par rapport à DistilHuBERT. À son tour, pour

les niveaux de réverbération élevés, le modèle proposé a montré qu'il surpassait même les modèles enseignants avec 2 à 6 fois plus de paramètres jusqu'à 89.19%.

Chapter 1

Introduction

1.1 Motivation

Large deep learning models have recently achieved great success on speech recognition tasks (Ao *et al.*, 2021; Babu *et al.*, 2021). These models, however, use a considerable amount of computational resources, which can be unfeasible for many edge applications. Edge applications focus on bringing computing as close to the source of data as possible in order to reduce latency and bandwidth use. It can be particularly important for speech recognition applications, where private and/or sensitive speaker data may need to be sent over the network to be processed remotely on large data processing clusters hosting very large and complex models. Bringing such large models to the edge can be challenging, as some edge devices may be resource-constrained with reduced storage and processing capacity. Moreover, edge applications, which typically involve in-the-wild scenarios, are corrupted by several environmental factors, such as ambient noise and/or room reverberation, which are known to be detrimental to speech-based applications. As such, a more detailed study on the impact of model compression and inference efficiency for large speech recognition models is needed. This MSc research aims to fill this gap.

Automatic speech recognition (ASR) aims to convert a continuous speech signal into a discrete text representation. While speech is one of the most efficient communication methods for humans (O'shaughnessy, 1987), text is an important representation for machines, and a large number

of techniques can be more easily applied to structure and understand the data. Recently, large deep learning models have achieved great success in ASR (Ao *et al.*, 2021; Babu *et al.*, 2021; Baevski *et al.*, 2020b; Hsu *et al.*, 2021a; Chen *et al.*, 2022; Radford *et al.*, 2023), with methods matching human speech recognition in a wide variety of acoustical environments (Spille *et al.*, 2018).

Self-supervised speech representation learning (S3RL) has established itself as a driving force behind these innovations. In this learning paradigm, the aim is to extract meaningful data representations through the utilization of large unlabeled datasets by exploiting the data modality. Subsequently, the model is fine-tuned with labeled data. Today, wav2vec 2.0 (Baevski *et al.*, 2020b), HuBERT (Hsu *et al.*, 2021a), and WavLM (Chen *et al.*, 2022) represent the most widely deployed universal speech representations, achieving state-of-the-art results not only for ASR but for other tasks such as speaker and emotion recognition (Feng *et al.*, 2023). While these models represent a breakthrough in terms of their performance, there is still a gap to bridge between ASR systems based on large models deployed on the cloud and ASR systems meant for edge deployment.

Existing universal S3RL methods, however, face two significant limitations in the context of edge applications: (i) their large size, which can be unfeasible for many edge applications; and (ii) their robustness, at inference time, against unseen environmental conditions, such as noise and reverberation. For example, the HuBERT model typically ranges from 95 million to 1 billion parameters, which is prohibitive on edge devices with limited storage and processing capacity. Model compression techniques, such as quantization, model pruning, and knowledge distillation, have been explored, with the former yielding the most promising outcome, as exemplified by the recent development of the "DistilHuBERT" model (Chang *et al.*, 2022).

On the environmental robustness side, previous studies have shown that using signal-based enhancement techniques is usually insufficient, as the distortions introduced by these algorithms can also degrade model performance (e.g., Kshirsagar *et al.* (2023)). Unseen noise and reverberation levels, for example, are known to drastically reduce the accuracy of even state-of-the-art ASR systems (Zhang *et al.*, 2023) and can be highly sensitive to different environmental conditions (Pimentel *et al.*, 2023a; Li *et al.*, 2022). While domain adaptation techniques, such as those proposed in "Robust HuBERT" (Huang *et al.*, 2022b) and "deHuBERT" (Ng *et al.*, 2023), can alleviate this problem, the methods are not directly applicable for compression.

In this work, we present two contributions to the field of S3RL. First, we explore the effects that train/test mismatch conditions have on speech recognition accuracy based on compressed self-supervised speech models. In particular, we report on the effects that parameter quantization and model pruning have on speech recognition accuracy based on the so-called robust wav2vec 2.0 model under noisy, reverberant, and noise-plus-reverberation conditions. Our results show that training with more data diversity significantly improves the robustness of speech recognition models not only against noise and reverberation, but against model compression as well. Second, we propose three innovations on top of the existing DistilHuBERT distillation recipe: optimize the prediction heads, employ a targeted data augmentation method for different environmental scenarios, and employ a real-time environment estimator to choose between compressed models for inference. Experiments with the LibriSpeech dataset, corrupted with varying noise types and reverberation levels, show the proposed method outperforming several benchmark methods, both original and compressed, by as much as 48.4% and 89.2% in the word error reduction rate in extremely noisy and reverberant conditions, respectively, while reducing by 50% the number of parameters. Thus, the proposed method is well suited for resource-constrained edge speech recognition applications.

1.2 Outline

This thesis builds upon the research presented in two manuscripts, namely :*On the Impact of Quantization and Pruning of Self-Supervised Speech Models for Downstream Speech Recognition Tasks "In-the-Wild"* (Pimentel *et al.*, 2023a) and *Environment-Aware Knowledge Distillation for Improved Resource-Constrained Edge Speech Recognition* (Pimentel *et al.*, 2023b).

This thesis is organized as follows. Chapter 2 presents the theory and background of the methods used in this study. Specifically, we describe the three methods of compression applied to our models, the concept of self-supervised speech representation learning and the speech processing universal performance benchmark. We also introduce the state-of-the-art (SOTA) open source speech recognition models that were used in this work as well as the recent adaptations to these models. Finally, we describe how in-the-wild scenarios, in particular, noise and reverberation affect the performance of speech recognition systems and how we can use blind signal-to-noise ratio (SNR) and reverberation time (RT60) estimators to improve the context awareness of speech recognition models. In Chapter 3, we evaluate the effects that pruning and quantization have on the Wav2Vec

2.0 models under noisy and reverberant ambient conditions. In Chapter 4 we use the knowledge distillation compression method and propose two context-aware compressed models, based on the HuBERT architecture that is robust to environmental conditions, while having fewer parameters. We compare the results of these models to multiple benchmark models of different sizes and show that our models possess grater robustness to detrimental environmental conditions. Finally, Chapter 5 presents our conclusions, summarizes our findings, and discusses the necessary steps to develop our work further.

Chapter 2

Background

In this chapter, we provide a comprehensive overview of the background material needed.

2.1 Compression

Model compression involves transforming a large, resource-intensive model into a compact version suitable for storage on constrained mobile devices. Additionally, it can involve optimizing the model for faster execution with minimal latency or achieving a balance between these objectives (Zhu *et al.*, 2023). Here, we present the compression methods used in this study.

2.1.1 Quantization

The idea of quantizing neural networks was introduced in Fiesler *et al.* (1990). It consists of compressing the original network by reducing the number of bits required to represent each weight (Cheng *et al.*, 2018). This idea can also be further extended to represent gradient and activation in the quantized form. The weight in a neural network are typically stored as 32-bit floating-point numbers. These weights can be quantized to 16-bit, 8-bit, 4-bit or even with 1-bit (which is a particular case of quantization, in which weights are represented with binary values only, known as weight binarization). Quantization can be performed during training, referred to as quantization-aware training (QAT) (Nagel *et al.*, 2022), or after training (post-training quantization (Liu *et al.*, 2021)).

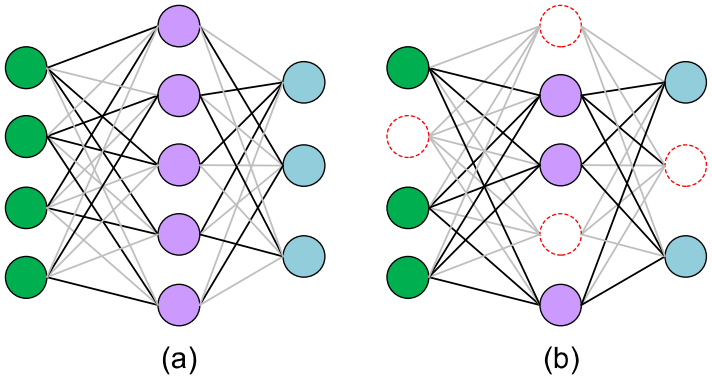


Figure 2.1: Diagram of (a) weight pruning and (b) neuron pruning. Grey lines represent pruned weights and dashed circles represent pruned neurons. Image taken from (Deng *et al.*, 2020).

2.1.2 Pruning

Network pruning, introduced in LeCun *et al.* (1989), is different from reducing the bitwidth of operands in quantization, in that pruning attempts to reduce the number of operands. Although it cannot simplify the arithmetic itself as quantization does, it is able to reduce the number of memory accesses and computational operations, thus obtaining acceleration (Deng *et al.*, 2020). In deep neural networks, many parameters are redundant and do not contribute much during training to lower the error and generalize the network. So, after training, such parameters can be removed from the network, and the removal of these parameters will have the least effect on the accuracy of the network. The primary motive of pruning is to reduce the storage requirement of the deep learning model and make it storage-friendly (Choudhary *et al.*, 2020b). The most common forms of pruning are weight pruning and neuron pruning. While the former reduces the number of edges, the latter reduces the number of nodes in a fully connected neural network. Figure 2.1 show a diagram of these two methods of pruning.

2.1.3 Knowledge Distillation

Knowledge Distillation (KD) is a widely used technique to transfer knowledge from a large model (teacher) to a smaller one (student) to achieve all types of efficiency. KD usually requires designing a loss function to minimize the distance of the output or intermediate features between the student and the teacher (Xu & McAuley, 2023). KD was originally proposed by Buciluǎ *et al.* (2006) and later expanded and popularized by Hinton *et al.* (2015).

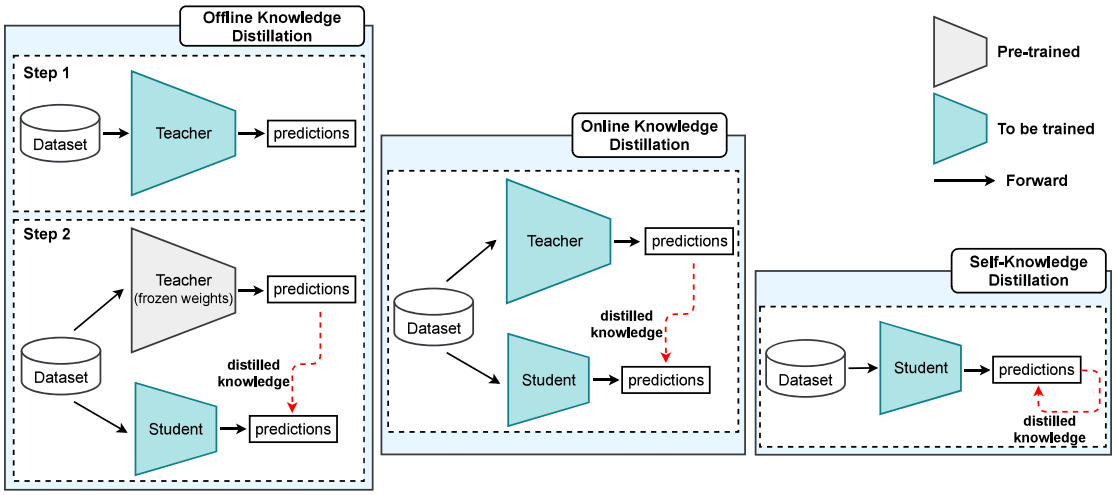


Figure 2.2: Comparison between different knowledge distillation mechanisms. Image taken from (Ghimire *et al.*, 2022).

The KD process can be offline, online, or through self-distillation. Offline distillation distills the knowledge from the pre-trained teacher model, whereas online distillation distills while the teacher and student models are being trained. In self-knowledge distillation, the student network is trained progressively using its own knowledge without requiring a pre-trained teacher model (Ghimire *et al.*, 2022). Figure 2.2 shows diagrams of the different types of KD processes.

In this work, we use 8-bit quantization, weight pruning and offline knowledge distillation as our compression methods due to the efficiency and ease of implementation of these methods.

2.2 Self-supervised speech representation learning (S3RL)

S3RL enables deep neural network models to capture meaningful and disentangled factors from raw speech waveforms (Guimarães *et al.*, 2023). S3RL can be seen as a special case of unsupervised learning as both schemes learn without annotations. Although conventional unsupervised methods rely on reconstruction or density estimation objectives, S3RL approaches rely on pretext tasks that exploit knowledge about the data variability used for training. Although supervised learning methods tend to learn stronger features than unsupervised learning approaches, they require costly and time-consuming work from human annotators to generate the required labels. S3RL techniques aim for the best of both worlds: training a powerful feature extractor using discriminative learning

without the need for manual annotation of training examples (Ericsson *et al.*, 2022). The learned universal speech representations can then be used across numerous tasks (Mohamed *et al.*, 2022).

In the context of S3RL, 'upstream models' are designed for pre-training and learning general representations from unlabeled data; in essence, they serve as feature generators for downstream tasks. The goal is to capture high-level features and patterns in the input speech data. These models are trained on a large dataset without task-specific labels. In turn, the 'downstream models' are task-specific models fine-tuned on labeled data for a particular speech-related task, such as speech recognition or emotion classification. Downstream models are trained on a smaller dataset with task-specific labels. The pre-trained upstream model is thus fine-tuned on this data to adapt its learned representations to the specific requirements of the target task.

2.2.1 Speech processing universal performance benchmark

The Speech processing Universal PERformance Benchmark (SUPERB) benchmark was proposed in Yang *et al.* (2021) with goal to offer the community a standard and comprehensive testbed for evaluating the generalizability of pretrained models on various tasks covering all aspects of speech. Figure 2.3 shows the interface used for the SUPERB benchmark.

General speech processing can be categorized into discriminative and generative tasks. Discriminative tasks involve learning the boundary between different classes in the input data, while generative tasks focus on capturing the underlying probability distribution of the entire dataset to generate new samples. The initial release of SUPERB focuses on the former, where ten tasks are collected from five domains. The tasks for each of the domains are:

- **Recognition** - Phoneme Recognition (PR) and Automatic Speech Recognition (ASR)
- **Detection** - Keyword Spotting (KS) and Query by Example Spoken Term Detection (QbE)
- **Semantics** - Intent Classification (IC) and Slot Filling (SF)
- **Speaker** - Speaker Identification (SID), Automatic Speaker Verification (ASV) and Speaker Diarization (SD)
- **Paralinguistics** - Emotion Recognition (ER)

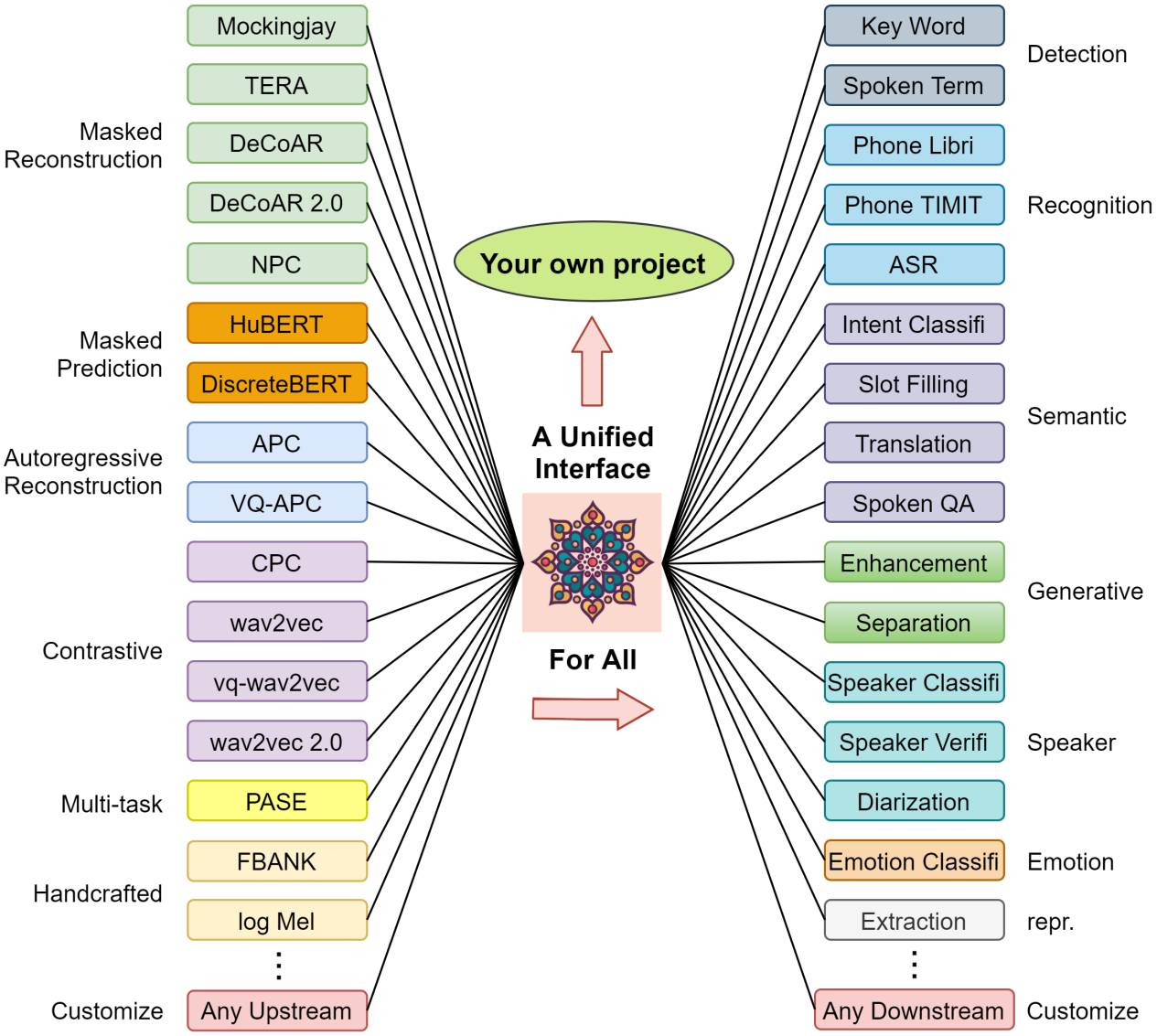


Figure 2.3: The SUPERB interface. An upstream model is used to acquire universal representations of speech, which are fed to a downstream model, used to perform specific speech related tasks. Image taken from `https://github.com/s3prl/s3prl`.

Since its original release, the SUPERB benchmark has added Speech Translation (ST) as a task in the semantics domain and has added "Generation" as an additional domain with Speech Enhancement (SE) and Speech Separation (SS) as new tasks. Meanwhile, the SUPERB organizers maintain a leaderboard, which takes in submissions from the public and ranks them based on their performance on all of the tasks compared against other SOTA models.

2.2.2 Speech Recognition Models

Wav2vec 2.0

The *wav2vec 2.0* (Baevski *et al.*, 2020a) model learns basic speech units used to tackle a self-supervised task. The architecture consists of a multi-layer convolutional feature encoder which takes as input raw audio and outputs latent speech representations at each time step, which are then fed to a context network. These representation carry information about the data and capture relevant features or patterns without explicit labeling. The encoder consists of several blocks containing a temporal convolution followed by layer normalization (Ba *et al.*, 2016) and a GELU (Hendrycks & Gimpel, 2016) activation function, while the context network is comprised of stacked Transformer layers, creating contextualised representations from the entire sequence.

The encoder network encoder layer consists of six stacked convolutional layers. It performs a mapping f from the raw waveform sample $x_i \in \mathcal{X}$ to an intermediate representation $z_i \in \mathcal{Z}$, in the form $f : \mathcal{X} \mapsto \mathcal{Z}$. The next important component is the context network, a mapping $g : \mathcal{Z} \mapsto \mathcal{C}$. The key idea of this network is to map different features (z_t ... z_T) into a single context vector $c_i = g(z_t...z_T)$, for T time-steps. This network is composed of a 12-layer Transformer network for the base model and a 24-layer network for the large model. The representations from $\mathcal{C}$ are the input features for downstream tasks later. The contrastive characteristic takes place in the loss function of the method. The idea is to compare samples (hence the name contrastive), and distinguish between the sample z_{t+k} that is k steps into the future from a set of distractors sampled from a uniform distribution p_n. Using the representations from $\mathcal{C}$, it is possible to train models for downstream tasks using less labeled data while keeping good performance. Furthermore, the loss function in this model includes both the contrastive loss and also a component related to the diversity loss of the learned codebook of the quantization module. This is done using the entropy maximization of the Gumbel-Softmax distribution, which encourages the model to use as many as possible different codes from the codebook instead of collapsing into a small subgroup. Figure 2.4 shows a diagram of the wav2vec 2.0 model.

The *robust wav2vec 2.0* Hsu *et al.* (2021b) model is a recent variant developed to provide improved robustness against domain shifts (e.g., due to noise, varying datasets, or other factors) at test time. Using the same architecture as its predecessor, robust wav2vec utilizes target domain

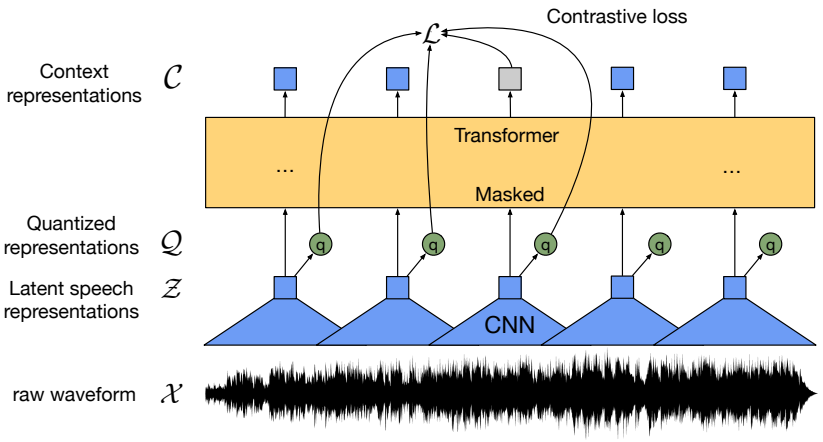


Figure 2.4: Diagram of the wav2vec 2.0 model. Image taken from (Baevski *et al.*, 2020a)

data during pre-training, thus leading to significant performance improvements in out-of-domain ASR.

HuBERT and Variants

The HuBERT model, as introduced in Hsu *et al.* (2021a), represents a speech pre-training technique designed to learn effective speech representations by means of masked feature reconstruction. This architecture is comprised of two major components: the encoder and the context network. The encoder network, denoted as f, operates as a mapping function from the raw waveform sample $x_i \in \mathcal{X}$ to an intermediate feature representation $z_i \in \mathcal{Z}$. It is structured as a 1D convolutional network, encompassing seven layers with 512 feature maps. The kernel and stride sizes for each layer vary accordingly. Following this convolutional network, a GELU non-linear activation function is applied. The context network, represented by the mapping function $g : \mathcal{Z} \mapsto \mathcal{C}$, serves a crucial role in the HuBERT architecture. Its fundamental principle revolves around mapping diverse features into a unified context vector through multiple Transformer encoders, thereby capturing long-range information. The interested reader can refer to Vaswani *et al.* (2017) for more details on the Transformer architecture.

Different versions of the HuBERT model have been developed, including the HuBERT *Base*, *Large*, and *X-Large* versions. The main difference between them is the size of their respective Transformer networks, comprising 12, 24, and 48 layers, respectively. Furthermore, it's worth noting that the HuBERT Base model was trained on a dataset encompassing 960 hours of the

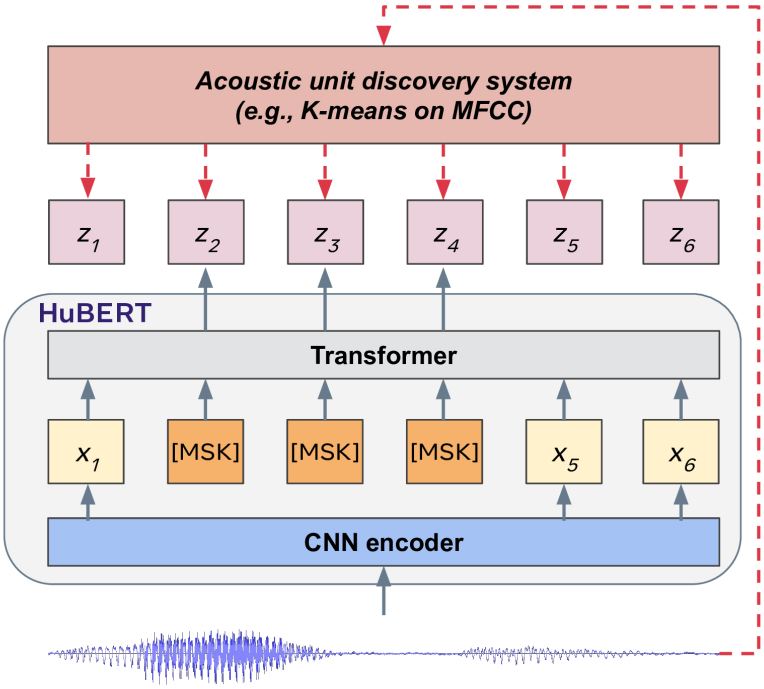


Figure 2.5: Diagram of the HuBERT model. Image taken from (Hsu *et al.*, 2021a)

LibriSpeech dataset (Panayotov *et al.*, 2015), while the Large and X-Large versions were trained on a substantially larger dataset consisting of 60,000 hours from the LibriLight dataset (Kahn *et al.*, 2020). Figure 2.5 shows a diagram of the HuBERT model.

Making HuBERT Robust to Noise

While larger versions of HuBERT have shown improved ASR accuracy in noisy conditions (Guimarães *et al.*, 2023), for edge applications such large footprint models can be problematic and smaller models can be more sensitive to noise. To this end, the authors in Huang *et al.* (2022b) proposed the so-called *Robust HuBERT* method, where domain adversarial training was used to make the system more robust to environmental factors. More specifically, a domain discriminator is responsible for classifying the source of the distortions applied to the utterance. The HuBERT Base model was continually trained on more diverse data, with a probability of receiving speech data corrupted by varying types and levels of noise, including Gaussian noise and recorded noises taken from the MUSAN database (Snyder *et al.*, 2015). Moreover, another technique to increase robustness is that of noise disentangling, as showed in Ng *et al.* (2023). The authors propose *deHuBERT*, a novel self-

supervised training framework that encourages noise invariance in HuBERT's embedded contextual representations by introducing a second embedding from different noise-augmented signals, using a shared convolutional neural network (CNN) encoder and a new pair of auxiliary loss functions.

Distilled versions of HuBERT

As mentioned above, one emerging topic in speech processing is that of edge ASR. For these applications, smaller models are needed, thus knowledge distillation has been explored recently. The recent work in Chang *et al.* (2022), for example, proposed the so-called *DistilHuBERT* model, where hidden representations from a HuBERT model were directly distilled to reduce model size and inference time.

More specifically, DistilHuBERT proposes a novel teacher-student framework for speech representation learning by multi-task knowledge distillation. The model consists of a CNN feature extractor and a small Transformer encoder. The idea is to learn to generate multiple teacher's hidden representations from shared representations. This is done by predicting the teacher's hidden representations with separate prediction heads. This objective is a multi-task learning paradigm and encourages the Transformer encoder to produce compact representations for multiple prediction heads. DistilHuBERT then takes a frozen pre-trained HuBERT Base model and uses only three prediction heads to respectively predict the 4^{th}, 8^{th} and 12^{th} HuBERT hidden layers' output. Before pre-training, the student is initialized with the teacher's parameters. Then, the student's prediction heads learn to generate the teacher's hidden representations by minimizing the loss function $\mathcal{L}^{(l)} = \mathcal{L}^{(4)} + \mathcal{L}^{(8)} + \mathcal{L}^{(12)}$. After training, the heads are removed, as the multi-task learning paradigm forces the DistilHuBERT model to learn representations containing rich information. These layers, however, are chosen to maximize performance across various speech related tasks and may not be optimal for ASR. In our experiments, we explore how different prediction heads can improve model performance.

One other method proposed recently for distilled universal representations is the one described in Guimarães *et al.* (2023) and called *RobustDistiller*. This is a recipe which combines data augmentation and multi-task denoising learning which are incorporated into the knowledge distillation process. In the data augmentation step, an online contamination of the data is performed during the distillation process. In particular, the student model receives the noisy data as input, but the

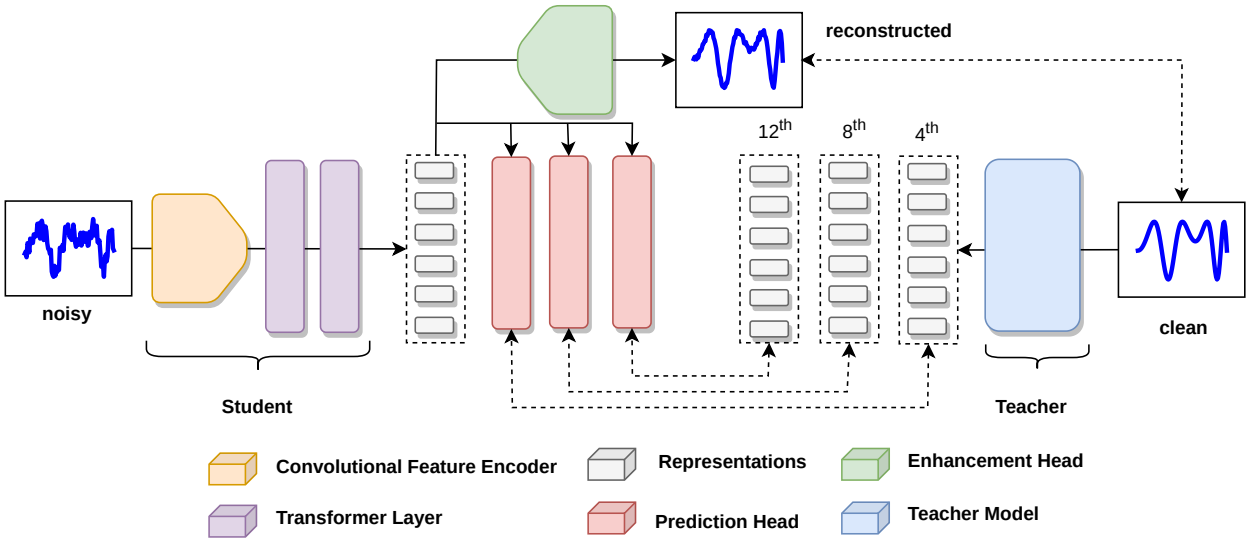


Figure 2.6: Diagram of the RobustDistiller recipe. (a) is the noisy; (b) is the clean; and (c) the reconstructed signal. Image taken from (Guimarães *et al.*, 2023).

network's target is to reconstruct the clean representations of the teacher model. Thus, in the multi-task denoising learning step, beyond learning to reconstruct the teacher's representations, an additional enhancement head is responsible for rebuilding the clean speech waveform from the learned representations. In contrast to usual enhancement techniques, the objective is to enforce the upstream model to carry enough information about the speech itself and not the noise components. While RobustDistiller was shown to outperform DistilHuBERT on keyword spotting, intent classification, and emotion recognition tasks, with noise and/or reverberation, it still showed some sensitivity to different environmental conditions. Figure 2.6 shows a diagram of the RobustDistiller recipe.

Lastly, we explore if the distillation of a teacher model built to be robust to environmental conditions also results in an environment-robust compressed student model. This is an important step to decide if robustness needs to be implemented at the teacher level or during the distillation process, as proposed herein. To this end, we apply the same distillation process used in DistilHuBERT, but applied it to the *Robust HuBERT* teacher described in Section 2.2.2. Henceforth, we term this model *DistilRobustHuBERT*.

2.3 In-the-Wild scenarios

2.3.1 Noise and Reverberation

Noise, in the context of acoustics, is any undesired sound, either one that is intrinsically objectionable or one that interferes with other sounds that are being listened to. In electronics and information theory, noise refers to random, unpredictable, and undesirable signals, or changes in signals, that mask the desired information content. One relevant type of noise is 'white noise'. Defined as a complex signal or sound that covers the entire range of audible frequencies, all of which possess equal intensity (Britannica, 2013). White noise is typically added to recordings of clean speech signals, along with other environmental noises, to simulate how this signal would be acquired in the real world. Environmental noise refers to unwanted sounds caused by human activity, such as road noise, aircraft, rail traffic, construction noise, building systems (e.g., ventilation or cooling), or neighbourhood noise (e.g., loud music, sports fields) (NCCEH, 2022).

Reverberation is the persistence of sound after its source has stopped due to multiple reflections from surfaces within a closed surface. These reflections build up with each reflection and decay gradually as they are absorbed by the surfaces of objects in the enclosed space. The reverberation time is an important parameter for characterizing the quality of an auditory space. Sounds in reverberant environments are subject to spectral distortions. This affects speech intelligibility and sound localization. Historically, the reverberation time has been referred to as the RT60, which is the time taken for the sound to decay to 60 dB below its value at cessation (Ratnam *et al.*, 2003).

2.3.2 Noise and Reverberation Estimators

Noise estimator

In this work, we utilize the Waveform Amplitude Distribution Analysis (WADA) SNR estimator, an algorithm initially proposed by Kim & Stern (2008). This algorithm provides a direct mapping from a signal to an estimated SNR value. It assumes that the amplitude distribution of clean speech can be approximated by the Gamma distribution with a shaping parameter of 0.4, and that the signal is corrupted with additive noise with Gaussian distribution. Even with these assumptions, experiments from the original authors show that the algorithm performs well when tested for three

different types of noise: additive white Gaussian noise, musical segments from the DARPA HUB 4 Broadcast News database, and noise from a single interfering speaker. The noise signals were artificially added to the speech signals at different SNRs ranging from -10dB to 30 dB. We chose to use this algorithm due to its computational efficiency and ease of use.

Reverberation time estimator

In order to estimate the reverberation time of a signal without knowing the properties of the room in which it was recorded, we follow the method described by Falk & Chan (2010) and Falk *et al.* (2007). The authors investigate the use of temporal dynamics information for blind measurement of room acoustical parameters. Long-term dynamics information is obtained by means of spectral analysis of temporal envelopes of speech, a process commonly termed modulation spectrum processing. In our work, we rely on the speech-to-reverberation modulation energy ratio metric (SRMR) described in Falk & Chan (2010) with source code available at `https://github.com/jfsantos/SRMRpy`. SRMR is inversely proportional to RT60, with lower values indicating higher reverberation levels. Additionally, we use a support vector machine (SVM) algorithm to provide a mapping between the inverse of the SRMR and the RT60.

2.4 Conclusions

In this chapter, we established the foundational concepts and models essential for the following discussions. This chapter provides the necessary framework to delve into the in-depth analysis and applications presented in the subsequent chapters.

Chapter 3

Quantization and Pruning of Models for Speech Recognition Tasks "In-the-Wild"

3.1 Introduction

In this chapter our overarching goal is two-fold: (1) understand how well SOTA speech recognition models behave under different model compression schemes, and (2) how well the compressed model accuracy remains under varying noise and reverberation conditions. We hope that the results from this study will shed light on the performance gaps that may exist before "edge speech recognition" is implemented in practice. Experiments with the latest (robust) wav2vec 2.0 model are conducted under two different compression schemes (quantization and model pruning) and five noisy conditions (SNR = 0, 5, 10, 15, and 20 dB) and two reverberations conditions (small room and medium room).

In its original proposal, Robust wav2vec 2.0 (Hsu *et al.*, 2021b) evaluated model performance in different out-of-domain conditions. However, varying noise levels – an important condition in edge applications – was not explored comprehensively. Here, we aim to fill this gap, as well as gauge the impact that model compression may have. In our experiments, pre-trained and fine-tuned models from the *Hugging Face* platform were used. More specifically, the wav2vec 2.0 model is pre-trained and fine-tuned on 960 hours of the LibriSpeech dataset, while the robust wav2vec 2.0 is pre-trained

on the Libri-Light, CommonVoice, Switchboard and Fisher datasets and fine-tuned on 960 hours of the LibriSpeech dataset.

3.2 Methods and Materials

In this section, we describe the methods and materials used in the study.

3.2.1 Model Compression Techniques

Two classic model compression techniques are explored to gauge the potential of edge speech recognition applications. The first is quantization, where the number of bits required to store each weight is reduced, thus substantially shrinking the model size, saving memory and accelerating computation. The method can be further extended to represent gradient and activation in the quantized form (Deng *et al.*, 2020). Here, we explore the impact of 8-bit quantization on all linear layers of the speech models and compare against its original 32-bit version (i.e., a compression ratio of 4). Next, model pruning is explored where redundant parameters can be removed from the network with minimal effect on model accuracy (Choudhary *et al.*, 2020a). Global unstructured pruning based on the lowest L1-norm was used at five different pruning rates, from 10-30% at 5% intervals.

3.2.2 Datasets

To train our models, we use the LibriSpeech corpus (Panayotov *et al.*, 2015) as the dataset of clean speech utterances. It consists of 960 hours of audiobook recordings with a 16 kHz sampling rate derived from the LibriVox project.

Libri-light (Kahn *et al.*, 2020) is a benchmark for the training of ASR systems with limited or no supervision. It contains a large dataset of 60,000 hours of unlabelled speech from audiobooks in English and a small labelled dataset (10h, 1h, and 10 min) plus metrics, trainable baseline models, and pre-trained models that use these datasets. It is also derived from open-source audio books from the LibriVox project. It contains over 60,000 hours of audio. The audio has been segmented using voice activity detection and is tagged with SNR, speaker ID and genre descriptions.

The CommonVoice (Ardila *et al.*, 2019) corpus is a massively-multilingual collection of transcribed speech intended for speech technology research and development. It is designed for ASR purposes but can be useful in other domains (e.g. language identification). To achieve scale and sustainability, the Common Voice project employs crowdsourcing for both data collection and data validation. The speech data is read by users on the Common Voice website, and is based upon text from a number of public domain sources like user submitted blog posts, old books, movies, and other public speech corpora. Over 50,000 individuals have participated so far, resulting in over 2,500 hours of collected audio.

The Switchboard Telephone Speech Corpus (Godfrey & Holliman, 1993) consists of approximately 260 hours of speech data. Switchboard is a collection of about 2,400 two-sided telephone conversations among 543 speakers (302 male, 241 female) from many areas of the United States. A computer-driven robot operator system handled the calls, giving the caller appropriate recorded prompts, selecting and dialing another person (the callee) to take part in a conversation, introducing a topic for discussion and recording the speech from the two subjects into separate channels until the conversation was finished. About 70 topics were provided, of which about 50 were used frequently. Selection of topics and callees was constrained so that: (1) no two speakers would converse together more than once and (2) no one spoke more than once on a given topic.

Similarly, the Fisher English Training Speech Transcripts (Cieri *et al.*, 2004) was developed by the Linguistic Data Consortium (LDC) and contains 984 hours of time-aligned transcript data for 5,850 telephone conversations in English. In addition to the transcriptions, there is a complete set of tables describing the speakers, the properties of the telephone calls, and the set of topics that were used to initiate the conversations.

3.2.3 Additive Noise and Reverberation

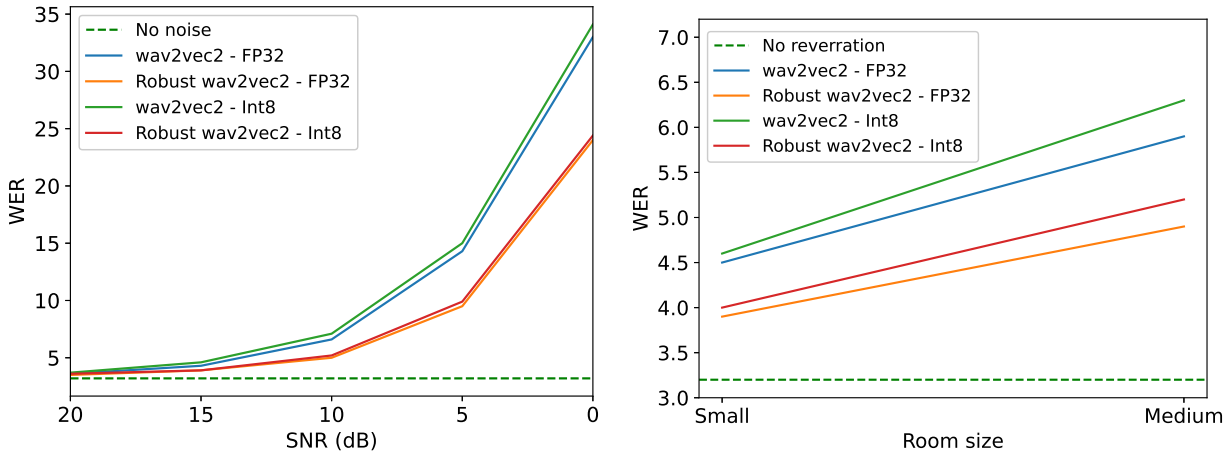
As we are interested in understanding the impact of compressed speech models in edge conditions, we use the noise signals present in the Deep Noise Suppression Challenge 4 (DNS4) dataset (Dubey *et al.*, 2022) to corrupt the test speech signals of the LibriSpeech dataset. This noise dataset consists of 180 hours of noise, present across 62,000 utterances, covering 150 different non-speech-like noise types. These files are added to the test samples at five varying SNR levels, ranging from 0 dB to 20 dB at 5 dB intervals. The audio files are sampled in order to make each class with at least 500 clips

and remove classes related to speech content. Reverberation, in turn, is simulated by convolving the clean signals with a uniformly sampled room impulse response (RIR). The openSLR28 dataset with 248 real RIRs and the openSLR26 with 60,000 synthetics RIRs are used (Ko *et al.*, 2017). The method to simulate the RIRs consist of sampling the room parameters and receiver position in the room and then using the image method (Allen & Berkley, 1979) to randomly generate a number of RIRs according to different speaker positions. The room parameters include the room dimensions (width, length and height) and the absorption coefficient. The set of rooms labeled as "small rooms" consists of rooms of length and width uniformly sampled between 1 m and 10 m. While the same parameters of "medium rooms" range between 10 m and 30 m. The height and absorption coefficient of all rooms are sampled uniformly from 2 m to 5 m and from 0.2 to 0.8, respectively. Lastly, the reverberation plus noise condition is simulated by combining the two previous steps. In all cases, if necessary, waveforms are resampled to 16 kHz.

3.3 Results and Discussion

First, we explore the impact of quantization on both original and robust versions of wav2vec 2.0 in clean matched conditions. Both models achieved a WER of 3.2% with weights stored in 32-bit floating point precision (total model size of 1262 Mb). After quantization, WER increased to 3.3%, thus a slight increase for a 4-fold model compression (total model size 354.5 Mb). Next, we explore the impact of pruning the weights of the convolutional and linear layers of the models. It is observed that the robust version of the speech model showed only a slight increase in WER to 3.3% at a prune rate of 0.3, while its original version increased to 5.7%.

Next, we explore the impact of varying noise levels and reverberation. Figures 3.1a and 3.1b show WER plots of original and quantized models, as a function of SNR and room size, respectively. As can be seen, 8-bit quantization showed minimal effect on model performance for the noisy case and a small impact when reverberation was present (e.g., WER for robust wav2vec went from 0.049 to 0.052 for the medium room size condition). Next, we perform an in-depth analysis of the WER achieved for different noise types. Table 3.1 shows the mean WER for audio files corrupted with six common noise types, namely domestic sounds (e.g., vacuuming), human voice, music, vehicle noise, wind noise, and other miscellaneous sources. As can be seen, human voice, domestic sounds, and vehicle noise showed the greatest performance deterioration. These are conditions in which



(a) (b)

Figure 3.1: WER for original (FP32) and quantized (Int8) models, with (a) noise and (b) reverberation.

Table 3.1: Mean WER for different noise types.

Noise Type	Mean WER	
	wav2vec 2.0	**Robust wav2vec 2.0**
Domestic sounds	13.6	10.6
Human voice	12.0	9.3
Miscellaneous sources	6.3	3.3
Music	11.0	7.6
Vehicle	11.7	8.8
Wind	5.8	5.4

edge applications would typically be seen, such as smart speakers and in-vehicle speech recognition. Overall, the robust wav2vec 2.0 model outperformed the original wav2vec 2.0 across all noise types, with the closest match achieved with wind noise.

Next, we explore the robustness of the models to pruning. Figures 3.2a and 3.2b show WER plots as a function of pruning rate and SNR or pruning rate and room size, respectively. As can be seen, pruning of the two models affected WERs, especially for SNRs lower than 15 dB, with the original wav2vec 2.0 model showing the greatest deterioration, particularly with pruning rates above 20%. Reverberation, in turn, showed minimal effect on the pruned robust version, but had a substantial impact on wav2vec 2.0, especially for medium sized rooms and pruning rates greater than 15%.

Lastly, we evaluate the robustness of the two models to pruning with combined additive noise and reverberation present in the test signals. Figure 3.3 shows the WER as a function of prune

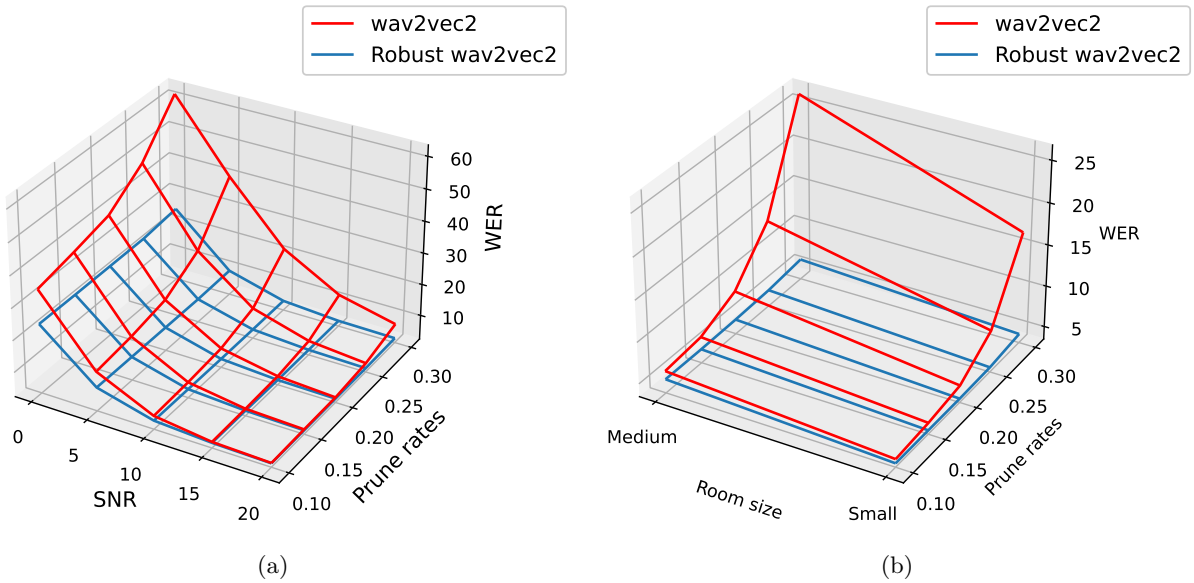


Figure 3.2: WER as a function of the pruning rate and (a) additive noise or (b) reverberation levels.

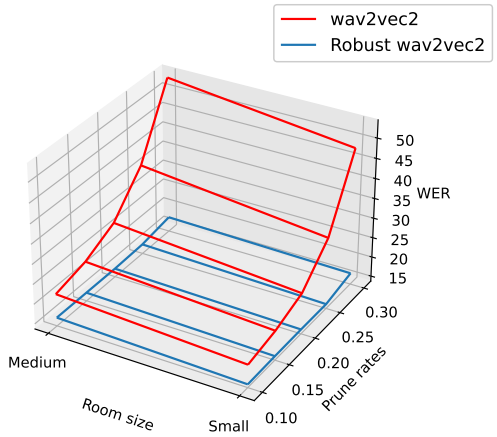


Figure 3.3: WER as a function of room size for signals with added noise between 0 and 20 dB.

rate and room size where noise levels have been averaged across the 0-20 dB range. Again the robust model showed to be insensitive to increased pruning rates, but sensitive to the degradations themselves. For example, at an SNR of 5 dB, the robust model achieved a WER of 10.1% at a pruning rate of 0.3. This error increased to 25.5% at an SNR of 0 dB. Notwithstanding, this is substantially better than what was shown with the original wav2vev 2.0 model that, under the same compression and noise conditions, achieved a WER of 62.2%.

Overall, the experiments described herein suggest that existing quantization and pruning compression schemes seem to be well-suited for edge speech recognition applications when applied in somewhat clean conditions. In such scenarios, compression rates as high as 4 could be achieved with

minimal impact on WER. On the other hand, if test applications involve noisy and/or reverberant conditions, improved speech representations are still needed, beyond what can be achieved with the so-called robust wav2vec model. Environment-aware knowledge distillation may be one possible solution.

3.4 Conclusions

In this chapter, we evaluate the robustness of two SOTA ASR speech models, namely wav2vec 2.0 and robust wav2vec 2.0, to unseen noisy and reverberant conditions when the models are compressed via quantization and pruning schemes. In particular, 8-bit quantization and L1-norm based global unstructured pruning were explored. It was found that while quantization and pruning have minimal impact on WER in clean conditions, noise and reverberation cause a significant WER degradation, even with models built inherently to be robust to such conditions. More robust compression and self-supervised representations are needed before edge speech recognition applications can be deployed "in the wild".

Chapter 4

Environment-Aware Knowledge Distillation for Speech Recognition in the Wild

4.1 Introduction

Recent works have started to propose solutions that tackle compression and environmental robustness jointly (e.g., Guimarães *et al.* (2023); Huang *et al.* (2022a)). These solutions, however, have yet to be explored for ASR and have shown some sensitivity to varying environmental conditions. While several universal representations exist (e.g., wav2vec, wav2vec 2.0, wavLM, and HuBERT), here we will focus on the HuBERT representation, as several variants have been proposed in the literature recently to make it more robust to environment noise, or to build more robust compressed versions for edge applications. This allows for several benchmark methods to be used for comparisons with the proposed method. It is important to emphasize that while the results reported herein will be based on a HuBERT representation, the proposed environment-aware method is applicable to any speech representation.

In this chapter, our overarching goal is three-fold: (1) adapt the existing DistilHuBERT representation (Chang *et al.*, 2022) to make it better suited for ASR tasks, (2) utilize data augmentation to increase the robustness of compressed models to unseen conditions, and (3) propose a hierar-

chical environment-aware solution where compressed models optimized for different environmental conditions are chosen during inference time, thus making the compressed models more robust to varying environmental conditions typically seen in edge conditions.

The remainder of this chapter is organized as follows: Section 4.2 presents the proposed methodology, while Section 4.3 introduces the study's experimental setup and Section 4.4 reports and discusses the obtained results. Lastly, Section 4.5 presents the conclusions.

4.2 Proposed Model

As mentioned previously, our proposed model incorporates three innovations to tackle issues highlighted above with existing systems. In the sections to follow, these three innovations are described.

4.2.1 Innovation #1: Modifying Prediction Heads

In an attempt to maximize the performance across the different tasks of the SUPERB (Yang *et al.*, 2021) benchmark, the original DistilHuBERT recipe utilized as the prediction heads the 4^{th}, 8^{th} and 12^{th} Transformer layers of the HuBERT Base teacher model (Chang *et al.*, 2022). These layers were shown to achieve improved accuracy across different tasks, but were not necessarily optimal for ASR. As such, our first innovation is to adapt the DistilHuBERT recipe to optimize the prediction heads for the ASR task.

According to the SUPERB instructions, after pre-training, the hidden states of different upstream layers are weighted, summed and fed to the task-specific layers, with the weights from each layer changing depending of the downstream task. A larger weight indicates a greater contribution of the corresponding layer. We analyzed the weights from each layer of the HuBERT model and found that layers 8, 9 and 10 provided the greatest contribution for the ASR task, thus our proposed method will rely on these three layers instead. Figure 4.1 shows the weight analysis of the HuBERT base model, finetuned for ASR.

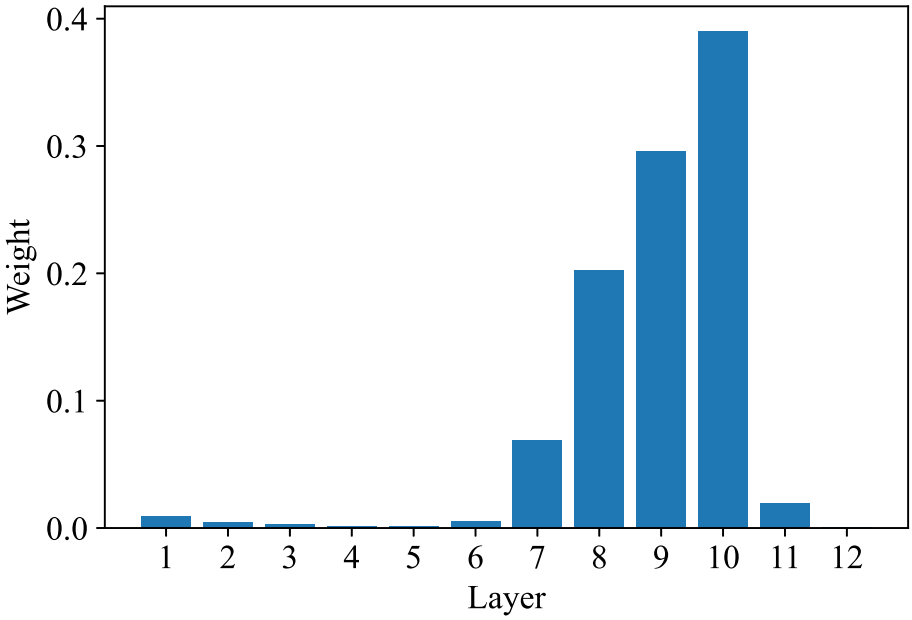


Figure 4.1: Weight analysis of the HuBERT model. The x-axis corresponds to each of the Transformer layers of the context network.

4.2.2 Innovation #2: Data Augmentation

Earlier S3RL models commonly relied on learning features from large unlabeled audiobook data, such as LibriSpeech (Panayotov *et al.*, 2015) or LibriLight (Kahn *et al.*, 2020). However, even though these models can learn fundamental characteristics from speech signals, real-world deployment data often involves diverse channel conditions and environmental noises that harm system performance, a problem known as domain shift. To tackle this issue, data augmentation has been proposed as a technique to improve data diversity and reduce model bias (Chen *et al.*, 2022; Huang *et al.*, 2022b; Hsu *et al.*, 2021b). Here, we adapt the RobustDistiller recipe proposed by Guimarães *et al.* (2023), as shown in Figure 4.2. In particular, aiming to reduce computational requirements, we do not utilize the speech enhancement head present in the original work. At training time, given a batch of clean speech utterances, we sample one action to be applied to each utterance in the batch: (i) no changes are made to the training utterance; (ii) contaminate the utterance with either additive noise with signal-to-noise ratio randomly chosen from $[0, 30]$ dB or convolve the speech waveform with a randomly selected room impulse response. The sampling probabilities of scenarios (i) and (ii) are 30% and 70%, respectively.

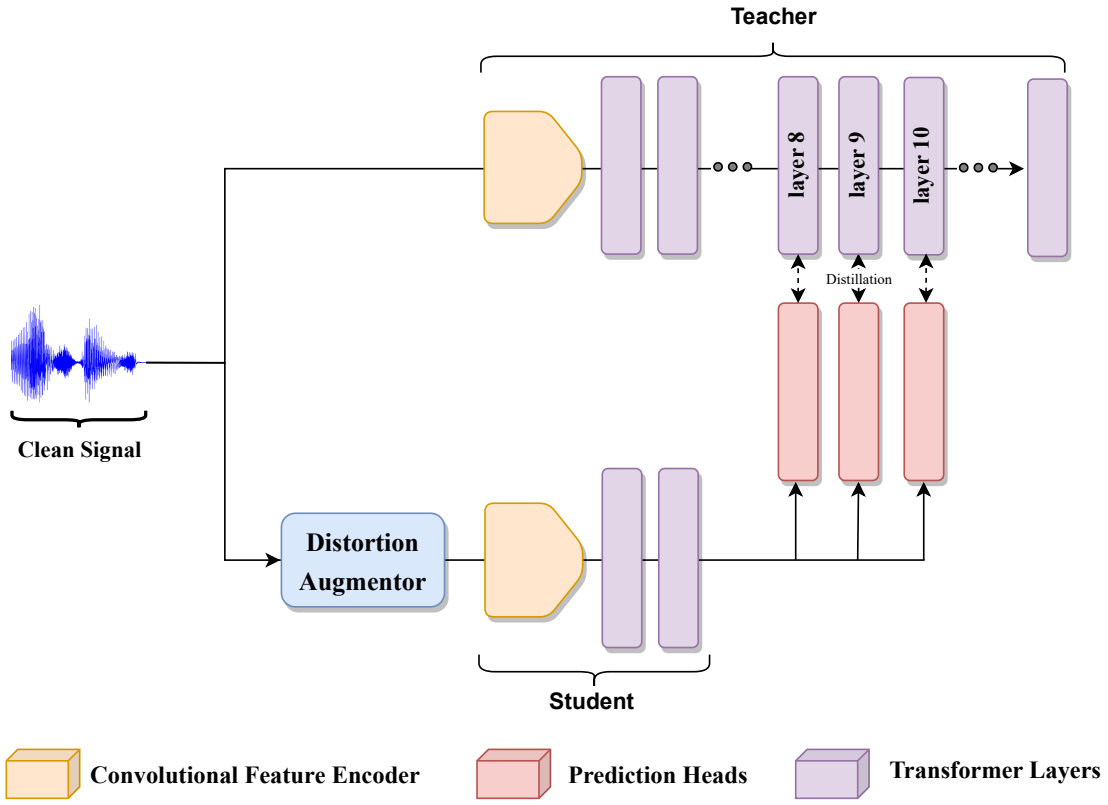


Figure 4.2: Block diagram of the adapted RobustDistiller pipeline (Figure 2.6) without the enhancement head.

4.2.3 Innovation #3: Environment Awareness

Recent work has shown that data augmentation during the knowledge distillation process can improve the performance of compressed models in mismatched domain scenarios, without compromising the model's size (Huang *et al.*, 2022a). However, there are still limitations on how much robustness a model can acquire from the distillation process. The latest system described in Guimarães *et al.* (2023), for example, showed some sensitivity to varying environmental conditions. Here, we propose a third innovation to overcome this issue, namely the use of environment-awareness.

More specifically, different compressed models are obtained, each optimized for a distinct environmental condition (e.g., high signal to noise or highly reverberant room). During inference, the best model is selected and used for ASR. This hierarchical approach allows for each compressed model to act as an expert for a given environment scenario. While increasing the number of models used reduces the overall compression gains seen with distillation, here we explore the use of only two models, thus still achieving some compression relative to the original teacher model. Moreover,

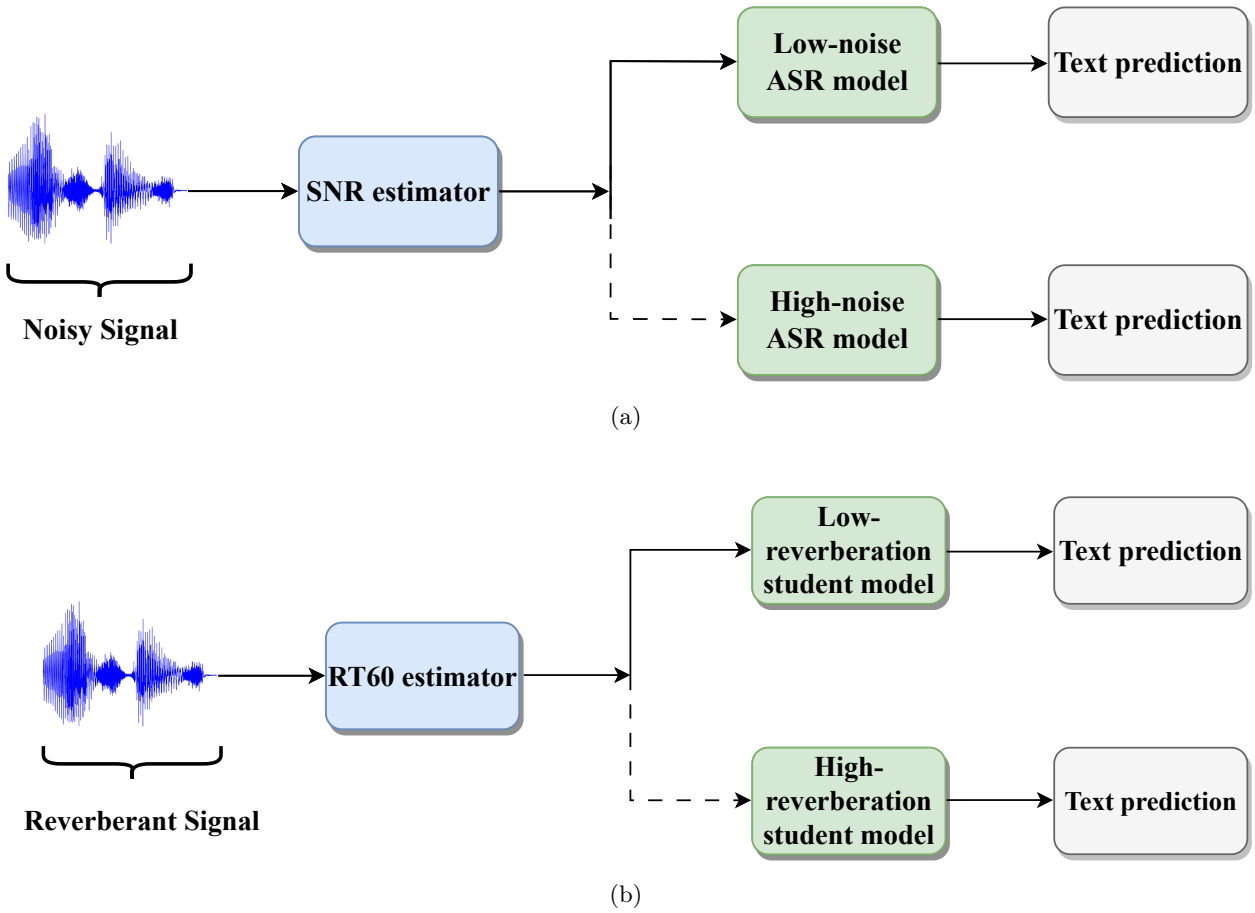


Figure 4.3: Diagram of the proposed Environment-Aware DistilHuBERT pipelines for (a) noisy and (b) reverberant environments.

as only one model is used during inference, the gains in inference time are not affected by relying on two models. Figures 4.3a and 4.3b depict two models we explore here, one that characterizes the noise levels in the environment and another the reverberation levels, respectively.

As seen from the figures, for real-time inference a noise/reverberation level selector is needed. As we assume that access to clean reference signals is not available, a "blind" measure is needed. Here, we explore with a signal-to-noise ratio (SNR) estimator called Waveform Amplitude Distribution Analysis (WADA) described in Kim & Stern (2008). For reverberation time (RT60) estimation, we rely on the speech-to-reverberation modulation energy ratio (SRMR) metric, followed by a mapping from the SRMR metric to reverberation time (RT60) via a support vector regressor, as described in Falk & Chan (2010).

4.3 Experimental Setup

In this Section we describe the databases used, the benchmark methods explored, figures-of-merit, as well as the blind noise and reverberation time estimators used.

4.3.1 Datasets

To train our models, we use once more the LibriSpeech corpus (Panayotov *et al.*, 2015) as the dataset of clean speech utterances. As we are interested in performing data augmentation and evaluating environmental awareness, we use noise signals present in the MUSAN (Snyder *et al.*, 2015) and UrbanSound8K (Salamon *et al.*, 2014) datasets to corrupt the signals from the LibriSpeech dataset during the training stage. These datasets contain approximately 15 hours of recordings in a wide variety of categories. UrbanSound8K contains 8732 labeled sound excerpts of urban sounds from 10 distinct classes, such as engine idling, car horn, and siren. We removed the children playing and street music categories to focus on non-speech like noise sources in this analysis. Moreover, we use the noise portion of MUSAN, which contains 929 files of assorted noise types, including office-like noises, as well as ambient sounds, such as car idling, thunder, wind, footsteps, and rain. This portion of the dataset does not include recordings with intelligible speech. However, some recordings include crowd and babble noises with indistinct voices. All the utterances are resampled to 16 kHz.

A room impulse response (RIR) dataset is also used, namely the Big Impulse Response Dataset (BIRD) (Grondin *et al.*, 2020), comprised of simulated room impulse responses corresponding to rooms of various sizes and absorption coefficients, with RT60 values ranging from 140 ms to 1 second. The train set used for reverberation consists of approximately 35,000 simulated RIRs sampled from the BIRD dataset. Half of these samples have low reverberation time (RT60 smaller than 500ms) and the other half have high reverberation time (RT60 greater than 500ms).

At test time, combined with the LibriSpeech test set, two additional datasets are used to test the model performance under unseen conditions. The first is the Acoustic Scene Classification from the Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge of 2020, namely DCASE2020, addapted from the original challenge proposed in 2018 (Mesaros *et al.*, 2018). The dataset consists of 64 hours of audio recordings in 10 acoustic scenes, recorded with four different

recording devices in 12 different cities. For reverberation, we use a different subset of the BIRD dataset comprised of approximately 4,000 simulated room impulse responses. Again, the signals are equally split between low and high RT60 values.

4.3.2 Pre-training

To gauge the benefits of our proposed methodology, different experiments are performed. First, we selected the HuBERT Base as our teacher model for the distillation process. Then, we implement our proposed models. The model referred to as *Robust DistilHuBERT* is a proposed variation of the DistilHuBERT model with only the first two innovations being implemented and is used to gauge the added benefits of noise awareness. This model is either robust to noise from 0 to 30 dB or to reverberation from 140 ms to 1 s. Meanwhile, our proposed model with all three modifications implemented is referred to as *Noise-Aware DistilHuBERT* or *Reverb-Aware DistilHuBERT*. They are composed of the pipeline depicted in Figures 4.3a and 4.3b, respectively. Additionally, HuBERT Large (Hsu *et al.*, 2021a), HuBERT Base (Hsu *et al.*, 2021a), Robust HuBERT (Huang *et al.*, 2022b), DistilHuBERT (Chang *et al.*, 2022), DistilRobustHuBERT and RobustDistiller (Guimarães *et al.*, 2023) are used as benchmark models.

In all experiments, the upstream models are trained using a single NVidia A100 GPU. Our robust distillation method and the fine-tuning step for the ASR downstream task take approximately 30 hours each to be completed, for a total of 60 hours of training time. We use the AdamW optimizer, with a batch of 24 utterances, for 200k iterations, whereas after 14k updates, the learning rate linearly decays from 2×10^{-4} to zero.

4.3.3 Evaluation Metric

In speech recognition, word error rate (WER) is a common metric for evaluating model performance (Huang *et al.*, 2014). The WER is defined as

$$WER = \frac{S + D + I}{N} \times 100 = \frac{S + D + I}{S + D + C} \times 100\,, \tag{4.1}$$

where S, D, I and C are the number of substitutions, deletions, insertions and correct words in the estimated sequence and N is the number of words in the true sequence ($N = S+D+C$). The WER

is based on the Levenshtein distance and a smaller value signifies a closer approximation between the estimated word sequence and the ground truth transcription (von Neumann *et al.*, 2023). It is important to notice that, while the WER is usually presented as value typically between 0 and 100, it does not represent a true percentage. While a WER of zero means perfect estimation, this metric is not constrained to an upper bound, and a sequence with more insertions than correct words will have a WER greater than 100.

4.4 Experimental Results and Discussion

In this Section, we describe the obtained results and discuss them in light of existing literature.

4.4.1 Accuracy of SNR and RT60 estimators

First, we need to validate if the proposed noise and reverberation level estimators are accurate. To this end, clean speech samples taken from the dev-clean set of the LibriSpeech dataset were corrupted with additive noise randomly sampled from the DCASE dataset. The SNR is uniformly sampled from 0 to 30 dB. A total of 2800 noisy test files are used for this experiment. Figure 4.4 shows the scatterplot between estimated and true SNR values. An overall correlation of 82.5% is achieved. It is important to remember, however, that as mentioned previously, here we are employing two compressed models, one optimized on low SNR levels ranging from 0-10 dB and another for high SNR levels (10-30 dB). Using the WADA algorithm to detect signals within these two classes results in an overall accuracy of 94.2%, suggesting the model is accurate enough for deployment.

Figure 4.5, in turn, shows the scatterplot between the inverse of the SRMR and the true RT60. The signals used in this ablation study are speech samples taken again from the dev-clean set of the LibriSpeech dataset convolved with RIR signals sampled from the BIRD dataset. As can be seen, an overall correlation of 82.2% was achieved. Again, as the proposed solution utilizes models optimized on low reverberation (RT60 smaller than 500 ms) and high reverberation (RT60 grater than 500 ms) levels, only a binary classifier is needed. We do a 90/10% train/test split on the data and use an SVM to perform the classification. Experiments using a vanilla SVM with a radial basis

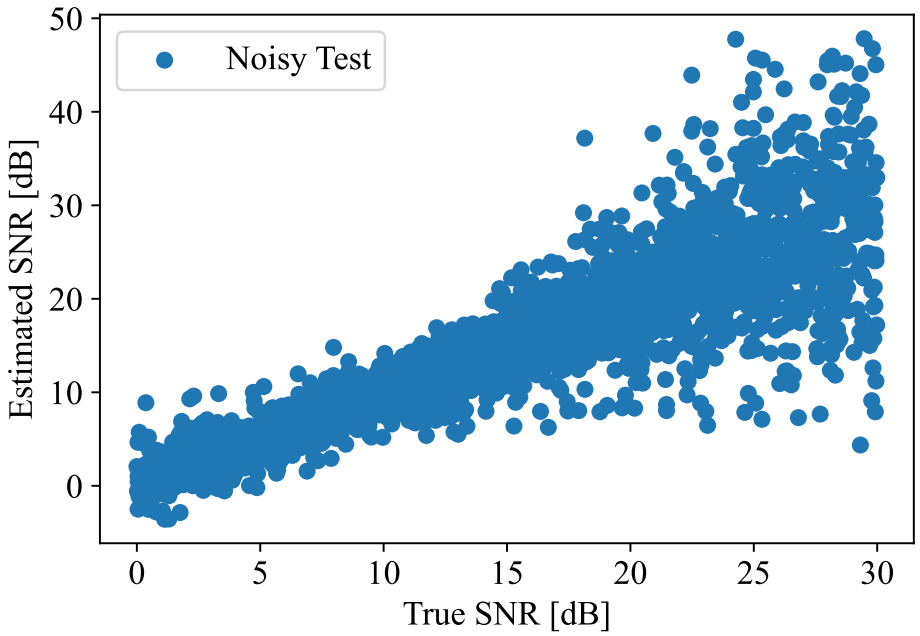


Figure 4.4: Scatterplot of estimated and true SNRs for noisy test signals using the WADA algorithm.

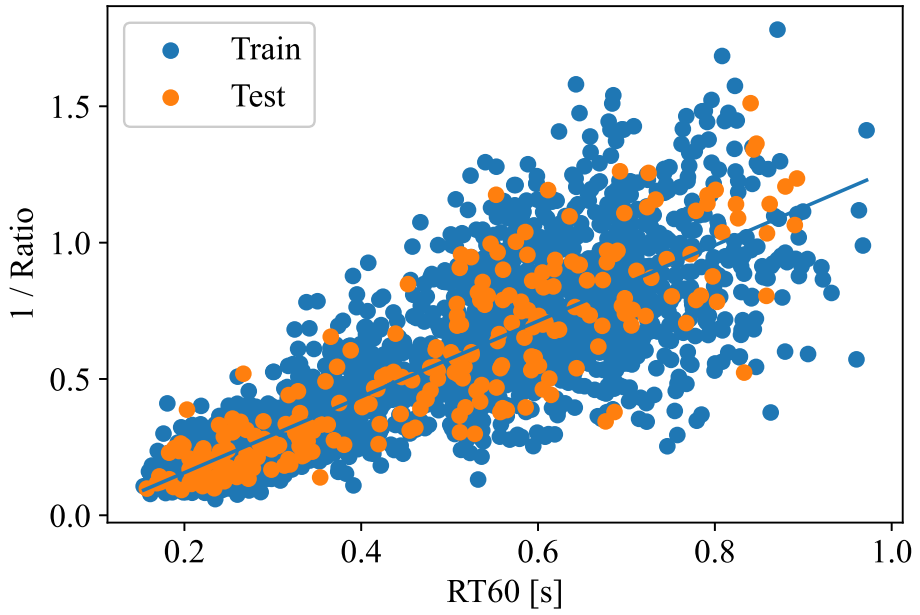


Figure 4.5: Scatterplot of inverse SRMR and true RT60.

function (RBF) kernel classifier resulted in an 88.9% classification accuracy, indicating again that the model is accurate enough for deployment.

4.4.2 Proposed System Performance

Table 4.1 compares the proposed methods to the five benchmark algorithms in terms of number of model parameters, number of multiply–accumulate operations (MACs), and the WER achieved on the clean and noisy test files. The noisy signals have been split into three categories: indoor, outdoor and transportation noise types, where the SNR was randomly sampled between 0 and 30 dB. As can be seen, the first two innovations (row 'Robust DistilHuBERT') already provide substantial improvement relative to the original DistilHubert model. Overall, relative gains of 33.04, 31.16 and 22.40% are obtained for indoor, outdoor, and transportation noise types, respectively. By incorporating all three proposed innovations (row 'Noise-aware DistilHuBERT'), a slight decrease in WER is achieved. It can also be observed that the proposed solutions result in compressed models that achieve similar WER across noise type conditions, suggesting improved robustness to ambient factors and applicability to edge conditions.

To further explore the benefits of the proposed context-awareness solution, we next focus on the lower SNR conditions, known to be the most impactful for ASR. Table 4.2 further reports the achieved accuracy for the benchmarks and proposed solutions for only the noisy files corrupted by additive noise ranging from 0-10 dB SNR. As can be seen, the gains achieved with the proposed noise-aware solution outperform the Robust DistilHuBERT model by 6.19, 4.24 and 4.62%, for indoor, outdoor, and transportation noise types, respectively. Relative to the original DistilHuBERT, these relative gains are of 48.42, 46.42 and 37.85%, respectively. It is important to emphasize that while the noise-aware solution does require storage of double the number of parameters relative to DistilHuBERT and RobustDistiller, inference time and computation requirements remain the same, as only one of the two models is used at a time. As such, the achieved gains can still be useful for edge applications involving very noisy conditions.

Lastly, Table 4.3 compares the WERs achieved for the benchmark and proposed solutions for speech signals under reverberant conditions. As can be seen, reverberation is a more challenging distortion to tackle and the WER of the benchmark distilled model is severely degraded. Implementing the first two innovations (Robust DistilHuBERT) allows the WER to be reduced by 66.52% overall and by 88.87 and 55.27% for the high RT60 and low RT60 conditions, respectively, relative to DistilHuBERT. The proposed environment-aware solution further decreases WER by an extra 2.55, 2.76 and 2.87%, respectively. Interestingly, the proposed solutions with only 24M or 48M

Table 4.1: Performance comparison across different clean and noisy conditions with SNR between 0-30 dB.

Model	#params (M)	MACs($\times 10^9$)	Clean	Noise Type (WER)		
				Indoor	Outdoor	Transport
HuBERT Large (Hsu *et al.*, 2021a)	300	4324	3.62	8.64	7.58	5.23
HuBERT Base (Hsu *et al.*, 2021a)	95	1669	6.43	11.80	10.82	8.89
Robust HuBERT (Huang *et al.*, 2022b)	95	1669	6.75	9.38	8.85	8.04
DistilHuBERT (Chang *et al.*, 2022)	24	785	13.29	26.21	23.94	19.64
DistilRobustHuBERT	24	785	12.70	24.41	21.92	18.36
RobustDistiller (Guimarães *et al.*, 2023)	24	785	14.03	19.93	18.53	17.22
Robust DistilHuBERT	24	785	12.74	17.55	16.48	15.24
Noise-Aware DistilHuBERT	48	785	12.44	17.33	16.39	15.09

Table 4.2: Performance comparison across different noisy conditions with SNR between 0-10 dB.

Model	Noise Type (WER)		
	Indoor	Outdoor	Transport
HuBERT Large (Hsu *et al.*, 2021a)	17.76	14.96	8.26
HuBERT Base (Hsu *et al.*, 2021a)	20.77	18.36	12.95
Robust HuBERT (Huang *et al.*, 2022b)	14.64	12.37	9.40
DistilHuBERT (Chang *et al.*, 2022)	45.56	39.64	29.59
DistilRobustHuBERT	41.80	36.35	27.04
RobustDistiller (Guimarães *et al.*, 2023)	28.86	24.99	20.99
Robust DistilHuBERT	25.05	22.18	19.28
Noise-Aware DistilHuBERT	23.50	21.24	18.39

Table 4.3: Performance comparison across different clean and reverberant conditions with RT60 from 140 ms to 1 s

Model	Clean	Reverberation time (WER)		
		All	High	Low
HuBERT Large (Hsu *et al.*, 2021a)	3.62	79.99	239.42	22.46
HuBERT Base (Hsu *et al.*, 2021a)	6.43	77.15	145.59	36.02
Robust HuBERT (Huang *et al.*, 2022b)	6.75	58.36	89.13	30.29
DistilHuBERT (Chang *et al.*, 2022)	13.29	156.98	648.02	74.04
DistilRobustHuBERT	12.70	77.14	97.48	59.61
RobustDistiller (Guimarães *et al.*, 2023)	14.03	67.54	90.86	43.61
Robust DistilHuBERT	13.78	52.56	72.02	33.12
Reverb-Aware DistilHuBERT	13.21	51.22	70.03	32.17

parameters already significantly improve ASR accuracy relative to HuBERT Large with 300M parameters. For the high reverberation level conditions, for example, the proposed environment-aware solution improves on HuBERT Large by 70.75% while requiring roughly one-sixth the number of parameters.

The results reported above have relied on predicted SNR/RT60 estimates. We have also explored the use of an 'oracle' system in which the true SNR/RT60 values are used (i.e., assuming perfect

classification), but this did not result in any significant improvement, suggesting that the few errors made by the classifiers had minimal impact on overall recognition accuracy. Overall, the obtained findings show the benefits of the proposed solutions for both noisy and reverberant settings, with the greatest gains seen in extremely low SNR conditions and highly reverberant settings, thus making the proposed models ideal for edge applications.

4.4.3 To Distill or Not to Distill (A Robust Model)

Several works have proposed to make large speech models more robust to environmental factors via adversarial training (Huang *et al.*, 2022b), disentanglement (Ng *et al.*, 2023), or data augmentation (Chen *et al.*, 2022), to name a few methods. The results reported herein suggest, however, that applying conventional distillation methods to robust teacher models does not guarantee that the resulting compressed model will retain robustness to its fullest. Results from Table 4.1, for example, show the WER achieved from the DistilRobustHuBERT student model being roughly 2.5 times higher than that achieved by the original Robust HuBERT teacher. In the high-noise conditions, from Table 4.2, WER went up almost three times with the compressed student versions. While distilling from a noise-robust teacher (DistilRobustHuBERT) showed some improvement over distilling from an original teacher (DistilHuBERT), the obtained results are still far from those achieved with the proposed method. These findings suggest that adding robustness to the distillation process, as proposed herein, is more important than finding a robust latent representation in which distillation can be applied to.

4.5 Conclusions

In this chapter, environmental awareness is proposed as a method of improving the robustness of compressed universal representations to be used for speech recognition. In particular, we propose three innovations on top of the existing DistilHuBERT distillation recipe: (1) optimize the prediction heads, (2) employ a targeted data augmentation method for different environmental scenarios, and (3) employ a real-time SNR or RT60 estimator to choose the best compressed model for inference. Extensive experiments show the proposed method outperforming several benchmarks, especially when signals are very noisy (SNR between 0 and 10 dB) or reverberant (RT60 greater than 500

ms). Overall, the findings suggest the proposed method can be better suited for edge speech applications under varying environmental conditions.

Chapter 5

Conclusions and Future Work

In this work, we explore the use of self-supervised speech representation learning, specifically in the field of ASR for resource constrained devices. We explore its capabilities and limitations and proposed innovations with the goal of reducing model size and number of parameters, while increasing its robustness to environmental conditions.

We first explore the robustness of two SOTA models, wav2vec 2.0 and Robust wav2vec 2.0 and examine how training with additional out-of-domain data domain affects the performance of compressed models. We test our models with two compression methods, namely 8-bit weight quantization and parameter pruning, with pruning rates ranging from 10% to 30%. These compressed models are evaluated under two corrupted conditions, specifically, additive noise using an SNR of 20 to 0 dB and reverberations of two simulated rooms of different sizes. Our results show that while domain shift training has a small impact on compressed models, especially the ones compressed through quantization, noise and reverberation still have significant effects on performance.

Additionally, we propose three innovations on top of the existing DistilHuBERT distillation recipe: optimize the prediction heads, employ a targeted data augmentation method for different environmental scenarios, and employ a real-time environment estimator to choose between compressed models for inference. We perform extensive experiments and compare against six benchmark models. When evaluated under noisy conditions (SNR between 0 and 30 dB), the proposed models outperform the benchmarks of comparable size (i.e., DistilHuBERT and RobustDistiller) by as much as 33.04%. The gains are more substantial in very noisy conditions (SNR between 0

and 10 dB), where gains of up to 48.42% were seen relative to DistilHuBERT. In turn, for high reverberation levels (RT60 greater than 500 ms), the proposed model showed to outperform even the teacher models with 2-6 times greater number of parameters (i.e., HuBERT base and large models) by as much as 89.19%.

5.0.1 Study limitations and future work

The proposed study is not without limitations. While the wav2vec 2.0 and HuBERT models were used because of the robust and compressed variants that have been published in the literature recently (thus can be used as benchmarks), more recent larger models, such as WavLM (with up to 300M parameters) (Chen *et al.*, 2022) or Whisper (up to 1.5B parameters) (Radford *et al.*, 2023) have been proposed and shown to be more robust to environmental noise. As such, the findings here are to be considered a lower bound on what could be achieved with environment-awareness.

Future work should explore quantization-aware and pruning-aware training, as opposed to the post-training methods used, as well as the proposed distillation recipe with these emerging models. Moreover, most of our analysis evaluated the performance of the models on noise-only and reverberation-only conditions. In realistic settings, their combined effects may be present. In such scenarios, it may be possible to combine the compression methods and utilize the SNR and RT60 predictors together and see which distortion condition is most severe.

Moreover, the present work has focused on edge ASR, an emerging application with the burgeoning of voice-based assistants in smart devices (e.g., speakers, cars, phones, and watches). Nevertheless, other important applications are starting to emerge, such as speaker verification to ensure users are secure when using applications, or emotion recognition for more intelligent and affective user interfaces. As such, future work should explore the impact of the proposed environment-awareness solution for other applications beyond ASR.

Bibliography

Allen J & Berkley D (1979). Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65:943–950. DOI:10.1121/1.382599.

Ao J, Wang R, Zhou L, Wang C, Ren S, Wu Y, Liu S, Ko T, Li Q, Zhang Y, Wei Z, Qian Y, Li J & Wei F (2021). Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing. DOI:10.48550/ARXIV.2110.07205.

Ardila R, Branson M, Davis K, Henretty M, Kohler M, Meyer J, Morais R, Saunders L, Tyers FM & Weber G (2019). Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*. `https://commonvoice.mozilla.org/en/datasets`.

Ba JL, Kiros JR & Hinton GE (2016). Layer normalization. *arXiv:1607.06450*. DOI:10.48550/ARXIV.1607.06450.

Babu A, Wang C, Tjandra A, Lakhotia K, Xu Q, Goyal N, Singh K, von Platen P, Saraf Y, Pino J, Baevski A, Conneau A & Auli M (2021). Xls-r: Self-supervised cross-lingual speech representation learning at scale. DOI:10.48550/ARXIV.2111.09296.

Baevski A, Zhou H, Mohamed A & Auli M (2020a). wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv:2006.11477*.

Baevski A, Zhou Y, Mohamed A & Auli M (2020b). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.

Britannica TEoE (2013). *noise*. `http://www.britannica.com/science/noise-acoustics`. Accessed 20 November 2023.

Buciluǎ C, Caruana R & Niculescu-Mizil A (2006). Model compression. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 535–541.

Chang HJ, Yang Sw & Lee Hy (2022). Distilhubert: Speech representation learning by layer-wise distillation of hidden-unit bert. *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pages 7087–7091.

Chen S, Wang C, Chen Z, Wu Y, Liu S, Chen Z, Li J, Kanda N, Yoshioka T, Xiao X *et al.* (2022). Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*.

Cheng Y, Wang D, Zhou P & Zhang T (2018). Model compression and acceleration for deep neural networks: The principles, progress, and challenges. *IEEE Signal Processing Magazine*, 35(1):126–136. DOI:10.1109/MSP.2017.2765695.

Choudhary T, Mishra V *et al.* (2020a). A comprehensive survey on model compression and acceleration. *Artificial Intelligence Review*, 53(7):5113–5155. DOI:10.1007/s10462-020-09816-7.

Choudhary T, Mishra V, Goswami A & Sarangapani J (2020b). A comprehensive survey on model compression and acceleration. *Artificial Intelligence Review*, 53(7):5113–5155. DOI:10.1007/s10462-020-09816-7.

Cieri C *et al.* (2004). *Fisher English Training Speech Part 1 Transcripts*. `https://catalog.ldc.upenn.edu/LDC2004T19`. Accessed 29 November 2023.

Deng L, Li G, Han S, Shi L & Xie Y (2020). Model compression and hardware acceleration for neural networks: A comprehensive survey. *Proceedings of the IEEE*, 108(4):485–532. DOI:10.1109/JPROC.2020.2976475.

Dubey H, Gopal V *et al.* (2022). Icassp 2022 deep noise suppression challenge. *Proc. ICASSP*.

Ericsson L, Gouk H, Loy CC & Hospedales TM (2022). Self-supervised representation learning: Introduction, advances, and challenges. *IEEE Signal Processing Magazine*, 39(3):42–62. DOI:10.1109/MSP.2021.3134634.

Falk T, Yuan H & Chan WY (2007). Spectro-temporal processing for blind estimation of reverberation time and single-ended quality measurement of reverberant speech. *Proc. Interspeech 2007*, volume 2, pages 514–517.

Falk TH & Chan WY (2010). Temporal dynamics for blind measurement of room acoustical parameters. *IEEE Transactions on Instrumentation and Measurement*, 59(4):978–989. DOI:10.1109/TIM.2009.2024697.

Feng Th, Dong A, Yeh CF, Yang Sw, Lin TQ, Shi J, Chang KW, Huang Z, Wu H, Chang X, Watanabe S, Mohamed A, Li SW & Lee Hy (2023). Superb @ slt 2022: Challenge on generalization and efficiency of self-supervised speech representation learning. *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 1096–1103.

Fiesler E, Choudry A & Caulfield HJ (1990). Weight discretization paradigm for optical neural networks. *Optical Interconnections and Networks*. Bartelt H, éditeur, International Society for Optics and Photonics, SPIE, volume 1281, pages 164 – 173.

Ghimire D, Kil D & Kim Sh (2022). A survey on efficient convolutional neural networks and hardware acceleration. *Electronics*, 11(6). DOI:10.3390/electronics11060945.

Godfrey JJ & Holliman E (1993). *Switchboard-1 Release 2 LDC97S62*. `https://catalog.ldc.upenn.edu/LDC97S62`. Accessed 29 November 2023.

Grondin F, Lauzon JS, Michaud S, Ravanelli M & Michaud F (2020). Bird: Big impulse response dataset.

Guimarães HR, Pimentel A, Avila AR, Rezagholizadeh M, Chen B & Falk TH (2023). Robustdistiller: Compressing universal speech representations for enhanced environment robustness. *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Hendrycks D & Gimpel K (2016). Gaussian error linear units (gelus). *arXiv:1606.08415v4*. DOI:10.48550/ARXIV.1606.08415.

Hinton G, Vinyals O & Dean J (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Hsu WN, Bolte B, Tsai YHH, Lakhotia K, Salakhutdinov R & Mohamed A (2021a). HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Hsu WN, Sriram A, Baevski A, Likhomanenko T, Xu Q, Pratap V, Kahn J, Lee A, Collobert R, Synnaeve G & Auli M (2021b). Robust wav2vec 2.0: Analyzing Domain Shift in Self-Supervised Pre-Training. *Proc. Interspeech 2021*, pages 721–725.

Huang KP, Fu YK, Hsu TY, Gutierrez FR, Wang FL, Tseng LH, Zhang Y & Lee Hy (2022a). Improving generalizability of distilled self-supervised speech processing models under distorted settings. *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 1112–1119.

Huang KP, Fu YK, Zhang Y & yi Lee H (2022b). Improving Distortion Robustness of Self-supervised Speech Processing Tasks with Domain Adaptation. *Proc. Interspeech 2022*, pages 2193–2197.

Huang X, Baker J & Reddy R (2014). A historical perspective of speech recognition. *Communications of the ACM*, 57:94–103. DOI:10.1145/2500887.

Kahn J, Rivière M *et al.* (2020). Libri-light: A benchmark for ASR with limited or no supervision. *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7669–7673. `https://github.com/facebookresearch/libri-light`.

Kim C & Stern R (2008). Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis. pages 2598–2601.

Ko T, Peddinti V, Povey D, Seltzer ML & Khudanpur S (2017). A study on data augmentation of reverberant speech for robust speech recognition. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pages 5220–5224.

Kshirsagar S, Pendyala A & Falk TH (2023). Task-specific speech enhancement and data augmentation for improved multimodal emotion recognition under noisy conditions. *Frontiers in Computer Science*, 5:1039261.

LeCun Y, Denker J & Solla S (1989). Optimal brain damage. *Advances in Neural Information Processing Systems*. Touretzky D, éditeur, Morgan-Kaufmann, volume 2.

Li S, Yerebakan MO, Luo Y, Amaba B, Swope W & Hu B (2022). The effect of different occupational background noises on voice recognition accuracy. *Journal of Computing and Information Science in Engineering*, 22(5):050905. DOI:10.1115/1.4053521.

Liu Z, Wang Y, Han K, Zhang W, Ma S & Gao W (2021). Post-training quantization for vision transformer. *Advances in Neural Information Processing Systems*. Ranzato M, Beygelzimer A, Dauphin Y, Liang P & Vaughan JW (éditeurs), Curran Associates, Inc., volume 34, pages 28092–28103.

Mesaros A, Heittola T & Virtanen T (2018). A multi-device dataset for urban acoustic scene classification. *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, pages 9–13.

Mohamed A, Lee Hy, Borgholt L, Havtorn JD, Edin J, Igel C, Kirchhoff K, Li SW, Livescu K, Maaløe L, Sainath TN & Watanabe S (2022). Self-supervised speech representation learning: A review. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1179–1210. DOI:10.1109/JSTSP.2022.3207050.

Nagel M, Fournarakis M, Bondarenko Y & Blankevoort T (2022). Overcoming oscillations in quantization-aware training. *Proceedings of the 39th International Conference on Machine Learning*. Chaudhuri K, Jegelka S, Song L, Szepesvari C, Niu G & Sabato S (éditeurs), PMLR, volume 162 de *Proceedings of Machine Learning Research*, pages 16318–16330.

NCCEH TEo (2022). *Environmental noise*. `https://ncceh.ca/resources/subject-guides/environmental-noise`. Accessed 20 November 2023.

Ng D, Zhang R, Yip JQ, Yang Z, Ni J, Zhang C, Ma Y, Ni C, Chng ES & Ma B (2023). De'hubert: Disentangling noise in a self-supervised model for robust speech recognition. *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

O'shaughnessy D (1987). *Speech communications: Human and machine*. IEEE Press.

Panayotov V, Chen G, Povey D & Khudanpur S (2015). Librispeech: an asr corpus based on public domain audio books. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pages 5206–5210.

Pimentel A, Guimarães H, Avila AR, Rezagholizadeh M & Falk TH (2023a). On the impact of quantization and pruning of self-supervised speech models for downstream speech recognition tasks "in-the-wild". *arXiv preprint arXiv:2309.14462*.

Pimentel A, Guimarães HR, Avila A & Falk TH (2023b). Environment-aware knowledge distillation for improved resource-constrained edge speech recognition. *Applied Sciences*, 13(23). DOI:10.3390/app132312571.

Radford A, Kim JW, Xu T, Brockman G, McLeavey C & Sutskever I (2023). Robust speech recognition via large-scale weak supervision. *Proceedings of the 40th International Conference on Machine Learning*, JMLR.org, ICML'23.

Ratnam R, Jones DL, Wheeler BC, O'Brien, William D. J, Lansing CR & Feng AS (2003). Blind estimation of reverberation time. *The Journal of the Acoustical Society of America*, 114(5):2877–2892. DOI:10.1121/1.1616578.

Salamon J, Jacoby C & Bello JP (2014). A dataset and taxonomy for urban sound research. *22nd ACM International Conference on Multimedia (ACM-MM'14)*, pages 1041–1044, Orlando, FL, USA.

Snyder D, Chen G & Povey D (2015). MUSAN: A Music, Speech, and Noise Corpus. *arXiv preprint arXiv:1510.08484v1*.

Spille C, Kollmeier B & Meyer BT (2018). Comparing human and automatic speech recognition in simple and complex acoustic scenes. *Computer Speech and Language*, 52:123–140. DOI:https://doi.org/10.1016/j.csl.2018.04.003.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł & Polosukhin I (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

von Neumann T, Boeddeker C, Kinoshita K, Delcroix M & Haeb-Umbach R (2023). On word error rate definitions and their efficient computation for multi-speaker speech recognition systems. *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Xu C & McAuley J (2023). A survey on model compression and acceleration for pretrained language models. AAAI Press, AAAI'23/IAAI'23/EAAI'23.

Yang Sw, Chi PH, Chuang YS, Lai CIJ, Lakhotia K, Lin YY, Liu AT, Shi J, Chang X, Lin GT *et al.* (2021). SUPERB: Speech Processing Universal PERformance Benchmark. *Proc. Interspeech 2021*, pages 1194–1198.

Zhang P, Huang Y, Yang C & Jiang W (2023). Estimate the noise effect on automatic speech recognition accuracy for mandarin by an approach associating articulation index. *Applied Acoustics*, 203:109217. DOI:https://doi.org/10.1016/j.apacoust.2023.109217.

Zhu X, Li J, Liu Y, Ma C & Wang W (2023). A survey on model compression for large language models. *arXiv preprint arXiv:2308.07633*.