UNIVERSITÉ DE SHERBROOKE
Faculté de génie
Département de génie électrique et de génie informatique

SYSTÈME PORTABLE DE REHAUSSEMENT DE LA PAROLE ATTENTIF AU BRUIT

Mémoire de maitrise
Spécialité : génie électrique

Thomas BERNARD

Sherbrooke (Québec) Canada

Avril 2022

MEMBRES DU JURY

François GRONDIN

Directeur

François MICHAUD

Évaluateur

François FERLAND

Évaluateur

RÉSUMÉ

La reconnaissance vocale est une interface intéressante pour un système intelligent en environnement industriel, permettant une interaction naturelle sans interruption des activités. Il y a actuellement un besoin entre les applications traditionnelles à lexique étroit (de type *commande et contrôle*) robustes au bruit et les assistants vocaux modernes. Ceux-ci, utilisés à travers des appareils intelligents, souffrent grandement en présence de bruit et de réverbération. Le projet présenté par ce mémoire vise à combler cette lacune en fournissant un système portable de rehaussement de la voix adapté aux milieux industriels, de manière à pouvoir y communiquer aisément avec un système commercial de reconnaissance de la parole infonuagique.

Ce mémoire présente la méthode KISS (*Keep It Super Simple* – « Garde ça super simple ») pour effectuer le pré-traitement (*front-end*) avec filtre spatial (*beamformer*) et ainsi rehausser un signal de parole multicanal capté par une matrice de microphones portable. KISS applique une approche intuitive de traitement de signal à l'estimation des matrices de covariance représentant la cible et l'interférence du système de rehaussement, une problématique qui est couramment solutionnée par des réseaux de neurones plus coûteux en calculs. De plus, puisque KISS ne requiert aucun entraînement avec un ensemble de données, la méthode est robuste face à des bruits inconnus et ne dépend pas d'une connaissance préalable de son environnement d'utilisation. La simplicité de calcul et l'intuitivité distinguant KISS d'autres méthodes à l'état de l'art sont à l'origine de son nom.

Combinée à une méthode de filtrage spatial établie comme celle de décomposition en valeurs propres généralisée (*Generalized Eigenvalue* – GEV), le système proposé surpasse largement l'approche populaire de filtrage Délai-et-Somme (DS), tout en se basant uniquement sur la même information de direction d'arrivée de la cible. En effet, les résultats de l'évaluation de KISS-GEV en environnement simulé montrent une amélioration du rapport signal-à-distorsion (*Signal to Distortion Ratio* – SDR) de 5,7 dB et de 4,9 dB comparativement au signal original et à celui rehaussé par DS, respectivement.

De plus, la viabilité de la solution a également été démontrée en environnement acoustique industriel réel, en utilisant une matrice de microphones embarquée fixée à un gilet de sécurité porté par un mannequin. L'évaluation de ce système a été réalisée à plusieurs emplacements d'un complexe industriel en simulant plusieurs locuteurs féminins et masculins. Les performances de rehaussement ont été évaluées avec un moteur de reconnaissance de la parole infonuagique, et le taux de reconnaissance moyen des mots (*Word Recognition Rate* – WRR) s'est avéré supérieur à 85 % avec la méthode KISS-GEV. De surcroît, dans les conditions les plus adverses, KISS-GEV va jusqu'à doubler le WRR comparativement au signal original ou à un rehaussement DS, ce qui transforme un résultat virtuellement inutilisable en une proportion suffisante de mots reconnus pour qu'une commande soit traitée par un assistant vocal.

Mots-clés : rehaussement de la parole, environnement industriel, assistant vocal, interface humain-machine, bruit inconnu, filtrage spatial (*beamforming*)

À ma mère, Chantal.

REMERCIEMENTS

Tout d'abord, je tiens à remercier mon directeur, François Grondin. Sa disponibilité et son soutien tant technique qu'humain ont été d'une valeur inestimable pour ma réussite et m'ont encouragé à persévérer. François est pour moi un mentor exceptionnel et j'espère faire honneur, dans les travaux présentés dans ce mémoire ainsi que dans nos collaborations futures, à la confiance et à l'encadrement qu'il m'a accordés tout au long de ma maîtrise.

J'aimerais également remercier le programme de formation CoRoM et son équipe pour leur contribution financière significative à mon projet. Par extension, je remercie également le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG) et Mitacs pour leur financement connexe à CoRoM.

Je tiens à remercier Jacob Kealey, pour son adaptation de la méthode KISS vers une implémentation en-ligne au sein de PyODAS, qui a fait partie intégrante des tests présentés au chapitre 4. Nos interactions ont toujours été agréables et, de mon côté, j'ai pu bénéficier de l'expertise de Jacob pour bonifier mes connaissances en Python.

Je remercie Stratégie Agilean inc. pour ses contributions financières, pour ses contributions à l'élaboration du projet et, surtout, pour m'avoir donné l'opportunité de rencontrer François. Merci également à leur client AWI inc. pour m'avoir permis de réaliser des enregistrements au sein de leurs installations industrielles à plusieurs reprises durant mon projet.

Finalement, je tiens à remercier mes parents, Chantal et Louis, mes amis proches, ma famille et Myriam. Sans leur écoute et leur soutien moral tout au long de mon parcours académique, je ne serais assurément pas rendu où je suis aujourd'hui et je leur suis profondément reconnaissant pour l'aide qu'ils m'ont offerte pour traverser les embûches que j'ai rencontrées au fil du temps.

TABLE DES MATIÈRES

1 INTRODUCTION **1**

2 MÉTHODES DE RECONNAISSANCE ET DE REHAUSSEMENT DE LA PAROLE **3**

2.1 Reconnaissance automatique de la parole 3
2.2 Conditions acoustiques 4
2.2.1 Bruit 6
2.2.2 Réverbération 7
2.2.3 Mesures de niveau sonore 9
2.3 Rehaussement de la parole 10
2.3.1 Canal unique 10
2.3.2 Canaux multiples 12

3 REHAUSSEMENT DE LA PAROLE PAR KISS-GEV **18**

3.1 Abstract 20
3.2 Introduction 20
3.3 KISS-GEV 21
3.4 Results 24
3.5 Conclusion 26

4 PERFORMANCES D'ASR DANS UN ENVIRONNEMENT INDUSTRIEL **29**

4.1 Méthodologie 29
4.1.1 Évaluation des performances 32
4.2 Résultats 33

5 CONCLUSION **38**

A RÉSULTATS DE L'ANALYSE DU WRR EN ENVIRONNEMENT INDUSTRIEL **40**

LISTE DES RÉFÉRENCES **43**

LISTE DES FIGURES

2.1 Spectrogrammes de l'énonciation « un, deux, trois, quatre, cinq » (a) signal propre ; (b) avec bruit blanc, sans réverbération ; (c) sans bruit, avec réverbération ; (d) avec bruit blanc, avec réverbération 5
2.2 Exemples de classification de plusieurs types de bruit 6
2.3 Exemple d'ordres de réflexions 7
2.4 Réponse impulsionnelle d'une pièce 8
2.5 Courbe de pondération A 9
2.6 Schémas-blocs comparatifs de systèmes de traitement de la parole à canal unique et à canaux multiples, où $x_m(t)$ est le signal brut en sortie du microphone m et $y(t)$ est le signal audio rehaussé 10

3.1 Overview of the KISS-GEV processing pipeline 21
3.2 Two-filter filterbank, with first filter ($b = 1$) in red, and second filter ($b = 2$) in blue. 22
3.3 Time-frequency masks (the color black stands for a value of 1, and white for a value of 0) 26
3.4 Spectrograms of the enhanced mixture, noise interference. Note the variation in the sharpness of upper harmonics (1.0 - 2.5 kHz) around 1 s mark and of fricative (4.0 - 8.0 kHz) at 2.5 s. 27
3.5 Spectrograms of the enhanced mixture, speech interference. Note the attenuation of voiced interference (0.0 - 4.0 kHz) segment before target utterance starts, around 0.6 s, and at 4.8 sec. 28

4.1 ReSpeaker Core v2.0 (microphones encerclés en rouge) 29
4.2 Boîtier pour gilet de sécurité imprimé en 3D 30
4.3 Mannequin 31
4.4 WRR par algorithme en fonction de l'emplacement 34
4.5 WRR par algorithme en fonction du locuteur 35
4.6 Spectrogrammes des 10 premières secondes d'une élocution du lecteur *masculin-complexe*, dans l'environnement *Génératrice*. Noter le très faible niveau de bruit ambiant dans (a), la déréverbération visible en région fricative (4 à 8 kHz) dans (b) et (c), et la coupure agressive de KISS-GEV en-ligne (c) entre 0 et 1 kHz. 37

A.1 WRR par algorithme pour chaque enregistrement 42

LISTE DES TABLEAUX

2.1 Exemples de matrices de microphones commerciales 14
2.2 Comparaison des méthodes de filtrage spatial 16

3.1 SDR (in dB) in various noise conditions 25

4.1 Caractéristiques des emplacements sélectionnés 31

A.1 WRR par algorithme pour chaque enregistrement 41

LISTE DES ACRONYMES

Acronyme	**Définition**
ASR	Reconnaissance automatique de la parole (*Automatic Speech Recognition*)
CSR	Taux de succès des commandes (*Command Success Rate*)
CTC	Classification temporelle connectionniste (*Connectionist Temporal Classification*)
dBA	Décibels pondérés en gamme A
DNN	Réseau de neurones profonds (*Deep Neural Network*)
DoA	Direction d'arrivée (*Direction of Arrival*)
DS	Délai-et-somme (*Delay-and-Sum*)
GCC	Corrélation croisée généralisée (*Generalized Cross-Correlation*)
GEV	Décomposition en valeurs propres généralisée (*Generalized Eigenvalue*)
GMM	Modèle de mélange gaussien (*Gaussian Mixture Model*)
HMM	Modèle de Markov caché (*Hidden Markov Model*)
KISS	« Garde ça super simple » (*Keep It Super Simple*)
LSTM	Mémoire à court et long terme (*Long Short-Term Memory*)
LTI	Système linéaire invariant dans le temps (*Linear Time-Invariant*)
MCRA	Moyenne récursive contrôlée par minima (*Minima Controlled Recursive Averaging*)

LISTE DES ACRONYMES

Acronyme	**Définition**
MFCC	Coefficients cepstraux sur l'échelle de fréquences de Mel (*Mel-Frequency Cepstral Coefficients*)
MVDR	Réponse sans distorsion à variance minimale (*Minimum Variance Distortionless Response*)
NMF	Factorisation de matrices non-négatives (*Nonnegative Matrix Factorization*)
PHAT	Transformation de phase (*PHAse Transform*)
RIR	Réponse impulsionnelle de pièce (*Room Impulse Response*)
RNN	Réseau de neurones récurrents (*Recurrent Neural Network*)
SCM	Matrice de covariance spatiale (*Spatial Covariance Matrix*)
SDR	Rapport signal à distorsion (*Signal to Distortion Ratio*)
SIR	Rapport signal à interférence (*Signal to Interference Ratio*)
SNR	Rapport signal à bruit (*Signal to Noise Ratio*)
SoC	Système embarqué sur une puce (*System on a Chip*)
SPL	Niveau de pression sonore (*Sound Pressure Level*)
SRP-PHAT	Puissance en réponse dirigée avec transformation de phase (*Steered-Response Power PHAse Transform*)
STFT	Transformée de Fourier en temps court (*Short Time Fourier Transform*)
STT	Parole à texte (*Speech-To-Text*)

LISTE DES ACRONYMES

Acronyme	**Définition**
SVD	Décomposition en valeurs singulières (*Singular Value Decomposition*)
TDoA	Délai temporel d'arrivée (*Time Delay of Arrival*)
VAD	Détecteur d'activité vocale (*Voice Activity Detector*)
WER	Taux d'erreur des mots (*Word Error Rate*)
WMMSE	Erreur quadratique moyenne minimale pondérée (*Weighted Minimum Mean Square Error*)
WRR	Taux de reconnaissance des mots (*Word Recognition Rate*)

CHAPITRE 1

INTRODUCTION

En milieu industriel, les interfaces humain-machine traditionnelles (p.ex., ordinateurs, écrans tactiles, claviers) sont complexes à déployer et sont gênées par certains équipements de protection (p.ex., gants). La voix s'avère une interface alternative intéressante pour interagir naturellement avec une machine, un robot ou encore un agent conversationnel. Les agents conversationnels ont récemment grandement gagné en popularité à travers les assistants virtuels offerts par les géants des technologies (Google Assistant, Apple Siri, Amazon Alexa, Microsoft Cortana, etc.). Ils sont maintenant très faciles d'accès à partir d'appareils intelligents d'usage général (téléphones, tablettes, ordinateurs, etc.) et de haut-parleurs intelligents disponibles à faible coût. Les enregistrements de parole en environnements résidentiels fournis par les utilisateurs de ces services constituent des ensembles de données de taille colossale. Ceux-ci, combinés avec la disponibilité d'une grande puissance de calcul pour l'entraînement, permettent à des modèles d'apprentissage machine complexes d'offrir d'excellentes performances de reconnaissance vocale. Néanmoins, ces performances se dégradent rapidement lorsque les conditions acoustiques s'écartent des données d'entraînement. Dans une usine, par exemple, des sources de bruit variées et parfois très puissantes sont présentes dans des environnements très réverbérants.

Avec l'avènement de l'industrie 4.0, plusieurs cas d'utilisation d'un système de reconnaissance automatique de la parole (*Automatic Speech Recognition* – ASR) avec agent conversationnel sont d'actualité. Parmi ces cas d'utilisation, il y a actuellement un besoin de la part des fournisseurs de logiciels de gestion industrielle infonuagiques d'employer un tel système comme interface entre leur logiciel et les travailleurs, cette interface étant plus polyvalente que les interfaces tactiles et pouvant être portée par un travailleur. Toutefois, les conditions acoustiques d'une usine sont très différentes de celles des scénarios d'utilisation résidentielle qui composent les ensembles de données d'entraînement des systèmes d'ASR commerciaux. Cette différence cause un appauvrissement des performances [52]. Un système de rehaussement adapté à l'environnement, employé en amont, permettrait de réduire l'écart avec les données d'entraînement et, conséquemment, de bonifier les performances d'ASR dans le milieu visé.

L'objectif principal de ce projet est de concevoir un système de rehaussement qui améliorera les performances d'ASR par rapport aux solutions commerciales offertes actuellement

(i.e., haut-parleurs et téléphones intelligents) en environnements industriels dont les conditions acoustiques sont inconnues. La question de recherche qui en découle est :

Est-ce possible de rehausser suffisamment un signal de parole en environnement acoustique industriel inconnu, par un système embarqué portable, pour interagir efficacement avec un agent conversationnel infonuagique ?

Ce mémoire présente la méthode KISS (*Keep It Super Simple* - « Garde ça super simple ») pour l'estimation des matrices de covariance nécessaires au calcul d'un vecteur de filtrage spatial (*beamformer*) sophistiqué, comme celui de décomposition en valeurs propres généralisée (*Generalized Eigenvalue* – GEV) ou celui de réponse sans distorsion à variance minimale (*Minimum Variance Distortionless Response* – MVDR). KISS constitue une solution simple, entièrement basée sur une approche de traitement de signal, à un problème complexe qui est couramment solutionné par des réseaux de neurones. Sa faible complexité de calcul en fait une solution idéale pour être exécutée par des systèmes embarqués à faible coût tout en conservant une latence très faible, tant pour une utilisation hors-ligne qu'en-ligne. De plus, puisque la méthode ne nécessite aucun entraînement, elle demeure performante face à des caractéristiques de bruit inconnues.

Le mémoire est structuré comme suit. D'abord, le chapitre 2 présente une revue de littérature menant à l'état de l'art en reconnaissance et en rehaussement de la parole. Un cadre théorique est également présenté sur les conditions acoustiques observables en situation réelle et les impacts de celles-ci sur le traitement du signal audio. Ensuite, le chapitre 3 présente, sous la forme d'un article soumis à la conférence *IEEE International Conference on Acoustics, Speech and Signal Processing*, le développement théorique de la méthode KISS-GEV appliquée au rehaussement hors-ligne dans des environnements acoustiques simulés. Finalement, le chapitre 4 présente l'étude de l'application des méthodes KISS-GEV et KISS-MVDR, en-ligne et hors-ligne, au rehaussement en environnement industriel réel, préalable au traitement par un système d'ASR commercial.

CHAPITRE 2

MÉTHODES DE RECONNAISSANCE ET DE REHAUSSEMENT DE LA PAROLE

La problématique de la reconnaissance automatique de la parole est loin d'être nouvelle [44]. Néanmoins, elle soulève certains défis dont la complexité est encore importante de nos jours. En effet, l'environnement acoustique d'une usine, par exemple, peut être particulièrement hostile aux systèmes d'ASR lorsqu'il comporte plusieurs sources de bruit distinctes et variées, et/ou un milieu très réverbérant. Les prochaines sous-sections abordent l'état de l'art et le cadre théorique concernant les méthodes de reconnaissance automatique de la parole, les conditions acoustiques pertinentes au contexte de rehaussement, puis les méthodes de rehaussement de la parole.

2.1 Reconnaissance automatique de la parole

Le processus d'ASR vise à extraire une séquence de caractères, de mots ou de phrases à partir d'un signal audio et à représenter cette séquence sous forme de texte. Celui-ci peut constituer un produit fini (p.ex., transcription d'un dialogue à des fins d'archivage) ou peut être soumis à un autre système de traitement, comme un agent conversationnel.

Les premières solutions d'ASR ayant démontré du succès utilisaient des modèles de Markov cachés (*Hidden Markov Models* – HMM), et des modèles de mélanges gaussiens (*Gaussian Mixture Models* – GMM), entraînés sur des caractéristiques (*features*) extraites manuellement, soit les coefficients cepstraux sur l'échelle de fréquences de Mel (*Mel-Frequency Cepstral Coefficients* – MFCC) [42].

Le taux d'erreur de mots (*Word Error Rate* – WER) est une métrique couramment utilisée pour l'évaluation et la comparaison des performances de systèmes d'ASR [39]. Il est défini par l'équation 2.1 [27] :

$$WER = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C}, \tag{2.1}$$

où

- S est le nombre de substitutions de mots ;
- D est le nombre de suppressions de mots ;

– I est le nombre d'insertions de mots ;
– C est le nombre de mots correctement détectés ;
– N est le nombre total de mots attendus ($N = S + D + C$).

Plus récemment, des approches hybrides combinant HMM et réseaux neuronaux profonds (*Deep Neural Networks* – DNN), ont offert une réduction notable du WER. En effet, l'expérience réalisée par Seltzer et al. [48] a relevé un WER moyen de 23,0 pour l'approche GMM-HMM et de 15,3 pour DNN-HMM, toutes deux opérant sur MFCC. Lorsque le bruit environnant est connu, l'entraînement du modèle sur le signal bruité permet à cette méthode de se démarquer davantage par sa robustesse aux conditions défavorables.

Alternativement, les approches de reconnaissance de la parole bout-en-bout par des réseaux de neurones utilisant la Classification Temporelle Connectionniste (CTC) [31] offrent également des performances très intéressantes, et sont entre autres utilisées par Google pour son service infonuagique de parole à texte (*Speech-To-Text* – STT) [45].

Ceci dit, même les méthodes d'ASR les plus avancées voient leurs performances se dégrader lorsqu'il y a une disparité entre les ensembles de données d'entraînement et les cas d'utilisation réels. Il est donc toujours difficile d'offrir des performances adéquates en présence de conditions inconnues [52]. C'est d'ailleurs pourquoi les systèmes commerciaux les plus performants, dont le STT de Google, permettent de choisir un modèle entraîné sur des conditions très spécifiques, telles que des bandes vidéo, des appels téléphoniques ou encore des commandes courtes [45].

2.2 Conditions acoustiques

Dans une situation idéale, le son voyage sur une très courte distance en ligne droite à partir d'une source vers un récepteur dans un environnement parfaitement anéchoïque et sans bruit : le signal est donc perçu par le récepteur exactement comme il est produit par la source. Ces conditions sont rares, voire presque impossibles à recréer parfaitement en pratique lorsque le microphone s'éloigne de la source. À l'exception d'une chambre anéchoïque, tout environnement comporte habituellement un certain niveau de bruit et de réverbération qui peuvent réduire l'intelligibilité du signal original, comme l'illustre la figure 2.1. La contamination d'un phonème à l'autre est occasionnée par la réverbération (particulièrement visible sur le troisième formant de « deux » (2 à 3 kHz), entre (a) et (c)). De plus, il y a l'effacement presque total d'une grande partie des informations spectrales par le bruit :

– il est uniquement possible de distinguer les formants les plus forts en périodes voisées ;
– les arrêts (*stops*) des consonnes /D/, /T/, /Q/ sont perdus ;

– la fricative /S/ est grandement atténuée.

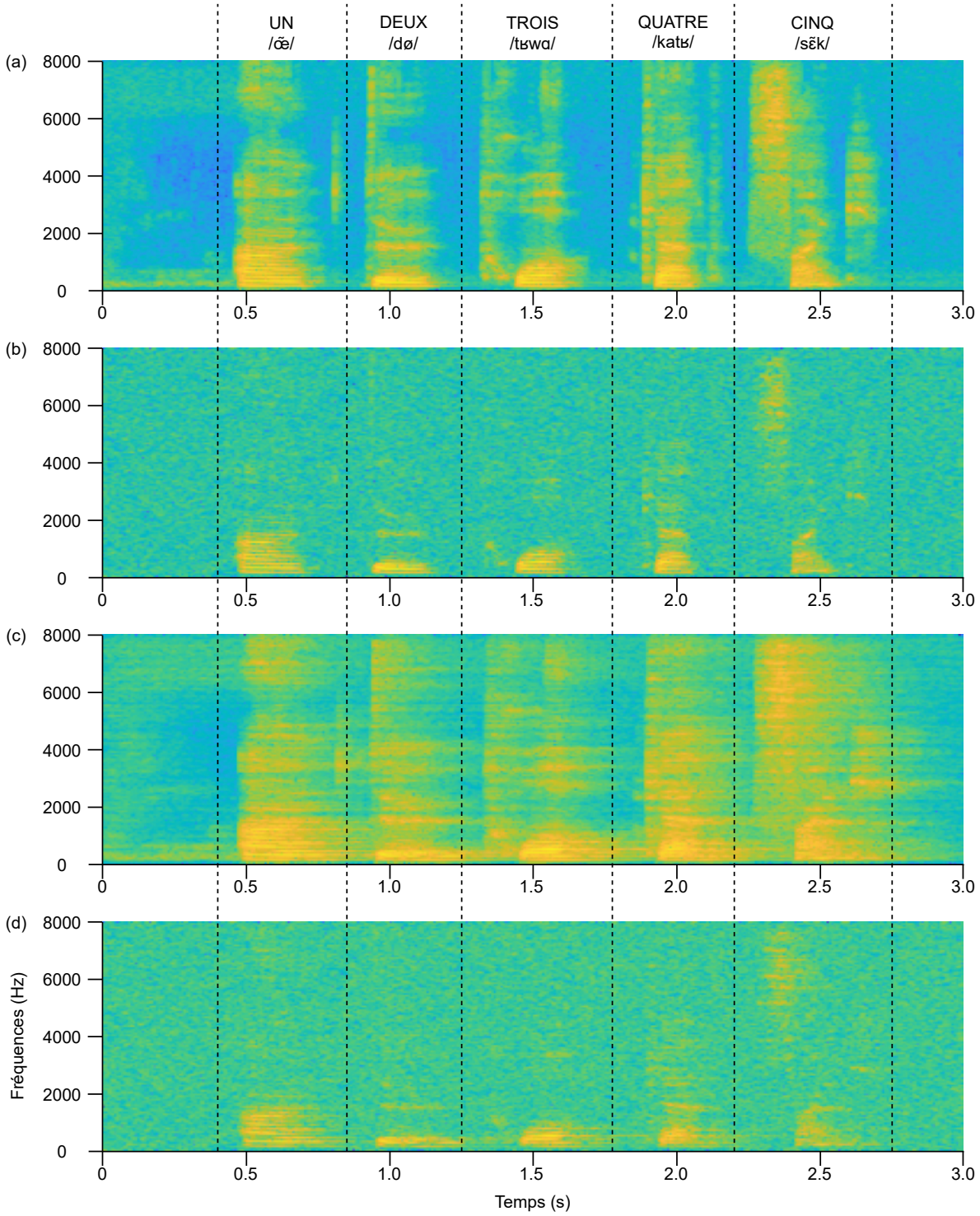


FIGURE 2.1 Spectrogrammes de l'énonciation « un, deux, trois, quatre, cinq »
(a) signal propre ; (b) avec bruit blanc, sans réverbération ;
(c) sans bruit, avec réverbération ; (d) avec bruit blanc, avec réverbération

2.2.1 Bruit

La présence de bruit, quel qu'il soit, réduit l'intelligibilité d'un signal sonore, qu'il soit perçu par un humain ou une machine. Dans un contexte d'ASR, la présence de bruit se traduit par une augmentation significative et indésirable du WER. Le bruit observé se distingue selon certaines caractéristiques, ce qui permet de mieux évaluer l'approche de rehaussement appropriée à un contexte donné.

- Un bruit peut être stationnaire ou non. Les caractéristiques spectrales d'un bruit stationnaire sont constantes, tandis que celles d'un bruit non-stationnaire varient dans le temps. Incidemment, les bruits non-stationnaires sont plus complexes à estimer et sont de nature beaucoup plus variée.
- Un bruit peut provenir d'une position spécifique (i.e., bruit cohérent) ou être distribué uniformément ou non dans l'espace (i.e., bruit diffus).

En pratique, il est fréquent d'observer une combinaison de plusieurs profils de bruits différents dont la nature, l'intensité et la position peuvent varier dans le temps. La figure 2.2 donne quelques exemples de profils différents.

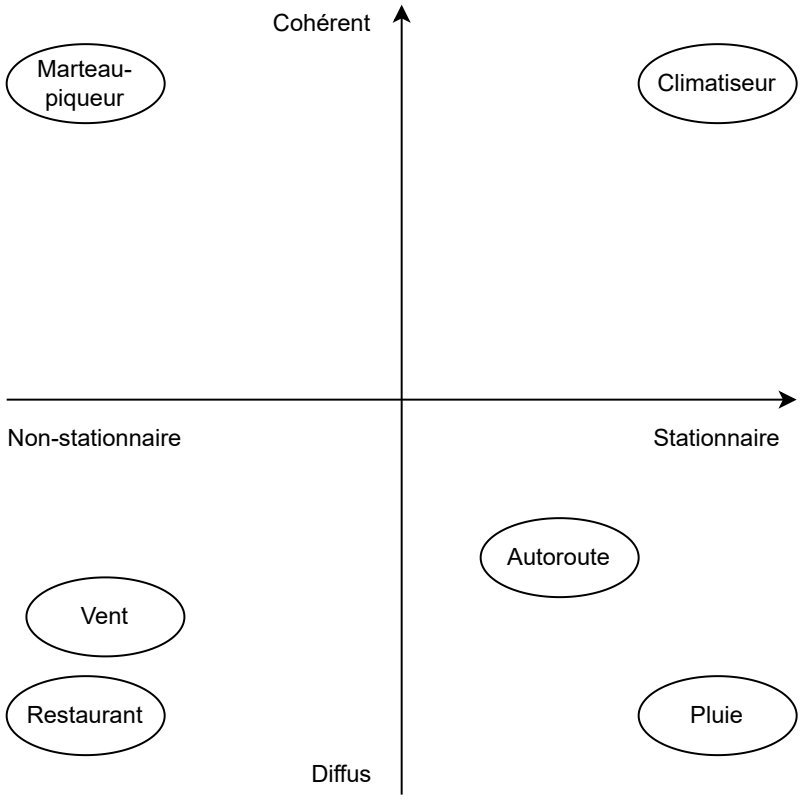


FIGURE 2.2 Exemples de classification de plusieurs types de bruit

2.2.2 Réverbération

Dans une situation acoustique réelle, un son émis par une source omnidirectionnelle est reflété par les différentes surfaces et objets qui l'entourent. Dans ce cas, le signal observé par un récepteur sera différent du signal produit par la source. Le signal observé peut généralement se décomposer en trois parties distinctes relatives au signal original, soit le chemin direct, les premières réflexions (*early reflections*), et la réverbération tardive (*late reverb*) [47], tel qu'illustre la figure 2.3.

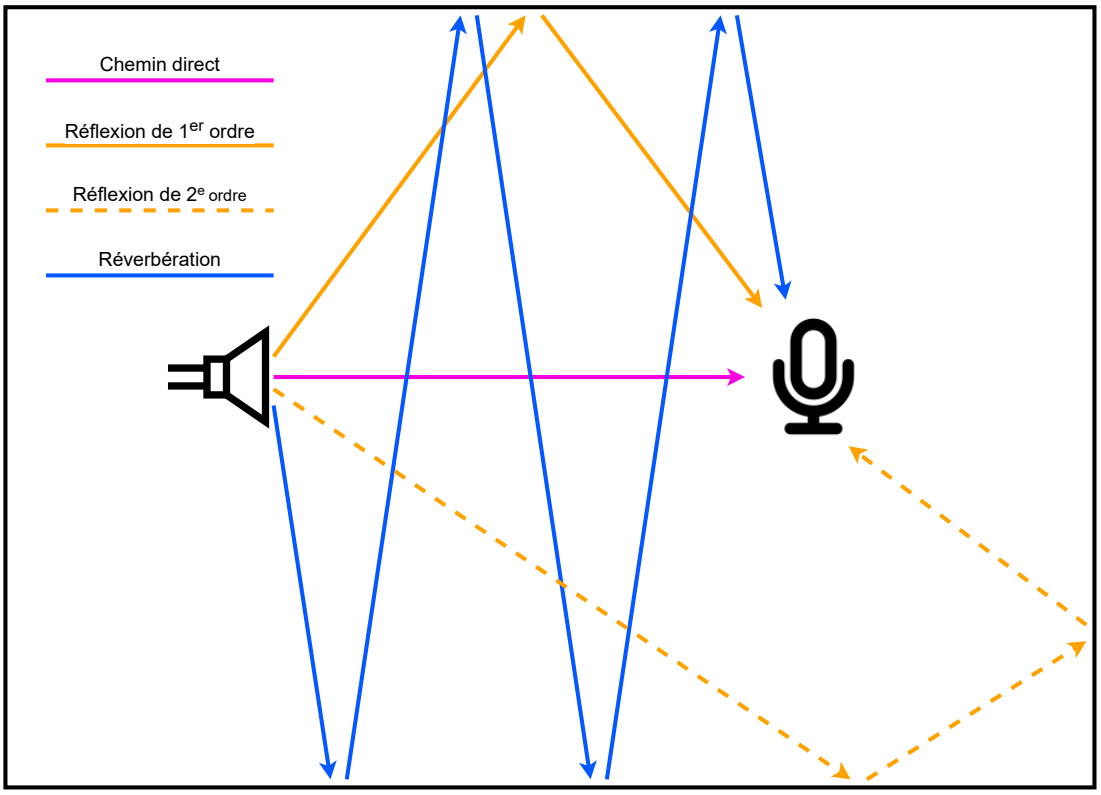


FIGURE 2.3 Exemple d'ordres de réflexions

En effet, la réponse en fréquences d'une surface ou d'un obstacle est très rarement uniforme. Chaque surface rencontrée par une trajectoire donnée occasionne une réflexion qui transforme le signal. En généralisant ce concept à une propagation omnidirectionnelle et donc à une infinité de directions d'émission, le signal reçu comprend :

- le chemin direct de la source à l'émetteur (peu ou pas atténué) ;
- les quelques premiers ordres de réflexion (légèrement atténués mais encore semblables au signal original) ;
- la réverbération plus avancée qui est essentiellement une superposition de tous les ordres de réflexion plus élevés (significativement affectée par les obstacles, donne l'allure de la réponse en fréquences de l'environnement).

Comme tout autre système linéaire invariant dans le temps (*Linear Time-Invariant* – LTI), une pièce avec une configuration source-récepteur spécifique se caractérise par sa réponse impulsionnelle, nommée réponse impulsionnelle de pièce (*Room Impulse Response* – RIR) [51]. Une telle représentation, dont un exemple est donné à la figure 2.4, permet de distinguer les trois composantes précédemment mentionnées d'un signal soumis à un environnement réverbérant.

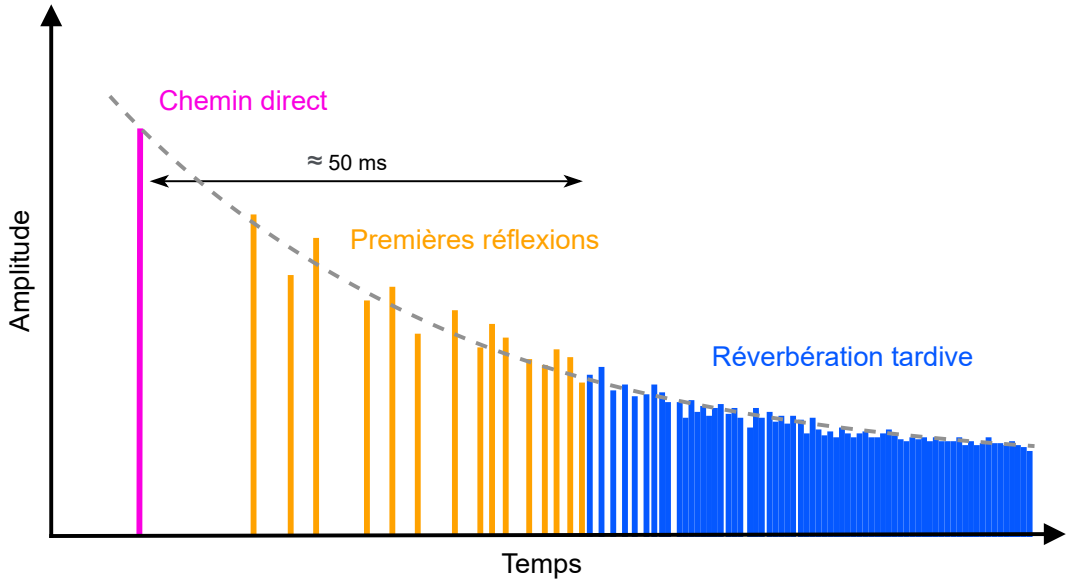


FIGURE 2.4 Réponse impulsionnelle d'une pièce

Pour revenir à la problématique du rehaussement de la parole, le phénomène de réflexion/réverbération nuit grandement à la performance des systèmes d'ASR. Il est particulièrement complexe à corriger étant donné sa relation avec le signal d'intérêt ainsi que sa variabilité lorsque l'environnement est changeant et/ou que la configuration source-récepteur n'est pas stationnaire.

Les phénomènes de bruit et de réverbération s'expriment à l'aide de l'équation 2.2 :

$$y(t) = h(t) * s(t) + n(t), \tag{2.2}$$

où $h(t)$ est la RIR spécifique à la configuration source/microphone utilisée, $s(t)$ est le signal de parole original, $n(t)$ le bruit observé et $*$ l'opérateur de convolution linéaire [32].

2.2.3 Mesures de niveau sonore

Le niveau sonore absolu observable dans un environnement est couramment défini par le niveau de pression sonore (*Sound Pressure Level* – SPL). Le SPL d'une onde sonore (en dB) est défini comme suit :

$$SPL = 10 \log \left(\frac{P^2}{P_0^2} \right) = 20 \log \left(\frac{P}{P_0} \right), \tag{2.3}$$

où $P_0 = 2 \times 10^{-5} Pa$, ce qui correspond à l'intensité acoustique $I_0 = 10^{-12} W/m^2$ à température ambiante et pression atmosphérique. Bien que chaque oreille soit unique, cette valeur de I_0 est couramment utilisée comme valeur de référence pour le seuil de l'audibilité humaine [42].

Puisque l'oreille humaine n'a pas une perception fréquentielle linéaire, le SPL est parfois pondéré pour représenter plus adéquatement certaines situations. La pondération la plus populaire est la gamme A, présentée à la figure 2.5. Bien que celle-ci soit critiquée pour sa sous-estimation de l'importance des basses fréquences dans la perception de l'intensité sonore et de l'inconfort causé par un son trop puissant [35], les décibels pondérés en gamme A (dBA) sont toujours couramment utilisés, entre autres pour définir des normes de sécurité telles que les limites d'exposition au bruit établies par le gouvernement du Canada [8].

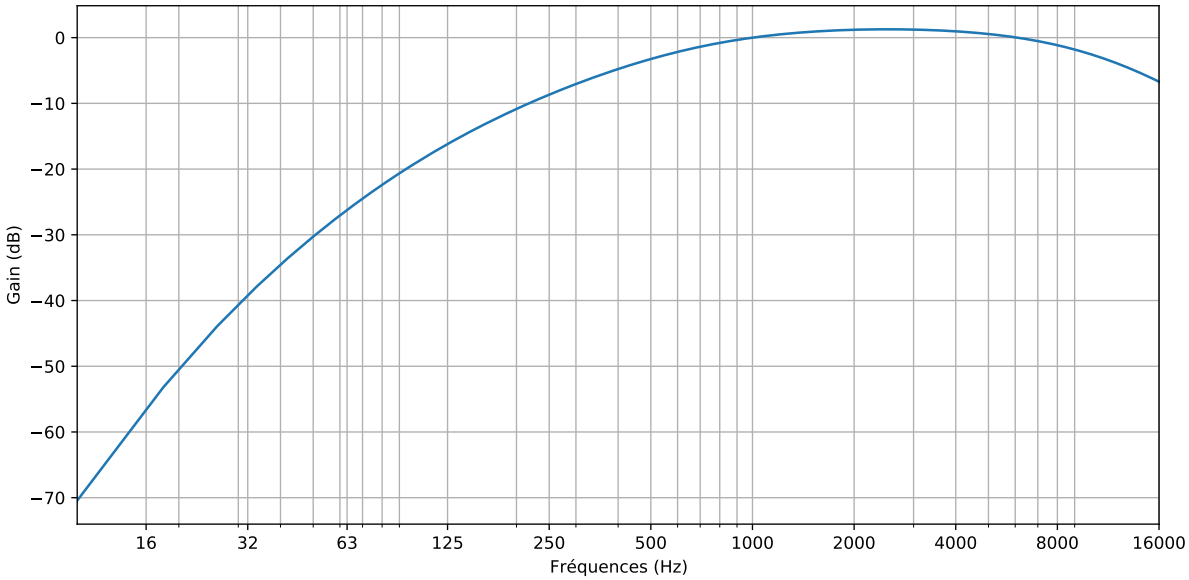
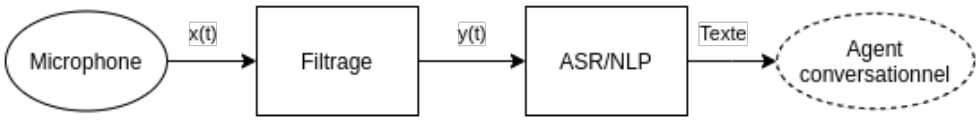


FIGURE 2.5 Courbe de pondération A

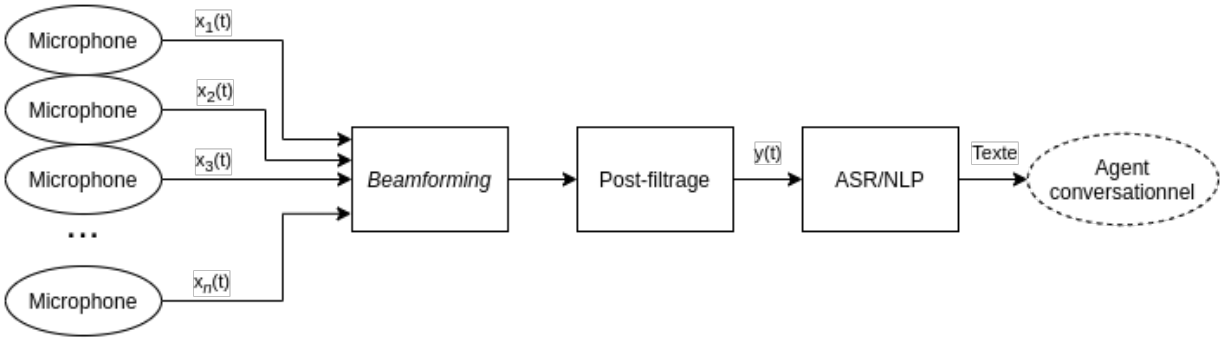
2.3 Rehaussement de la parole

Heureusement, il est possible de mettre à profit la performance élevée des systèmes d'ASR infonuagiques commerciaux en employant un système de rehaussement de la parole en amont, de manière à leur fournir un signal en entrée le plus près possible de leur ensemble de données d'entraînement.

Traditionnellement, le rehaussement est effectué sur un canal unique. Bien que le rehaussement du signal d'un seul microphone puisse rendre des résultats satisfaisants dans plusieurs situations, la disponibilité de canaux multiples, qu'il s'agisse d'un ensemble stéréo ou d'une matrice dotée d'un plus grand nombre de microphones, permet l'emploi de méthodes de rehaussement additionnelles, notamment l'estimation de la direction d'arrivée et le filtrage spatial (*beamforming*). Le rehaussement à canaux multiples est donc généralement plus performant face à des conditions exigeantes. La figure 2.6 présente un exemple à haut niveau de ces deux configurations.



(a) Canal unique



(b) Canaux multiples

FIGURE 2.6 Schémas-blocs comparatifs de systèmes de traitement de la parole à canal unique et à canaux multiples, où $x_m(t)$ est le signal brut en sortie du microphone m et $y(t)$ est le signal audio rehaussé

2.3.1 Canal unique

L'approche la plus simple à la problématique du débruitage est la soustraction spectrale. Tel que son nom le suggère, il s'agit de soustraire au spectre en fréquences du signal bruité une estimation du spectre du bruit, puis d'obtenir le signal rehaussé par transformée inverse [6]. Par conséquent, il faut disposer d'une estimation du bruit, ce qui constitue en soit une autre problématique. Avec un détecteur d'activité vocale idéal (*Voice Activity Detector* – VAD), il est possible de relever un échantillon lors de périodes sans voix et de

l'utiliser comme représentation du bruit. Bien que la soustraction spectrale puisse sembler intéressante en regardant le rapport signal à bruit (*Signal to Noise Ratio* – SNR) du signal rehaussé, l'approche est critiquée par l'insertion de bruit musical dans le signal rehaussé. Ce bruit est introduit lorsque l'amplitude du spectre de l'interférence estimée est supérieure à celle du signal cible, dans quel cas le résultat de la soustraction spectrale est négatif. Puisque l'amplitude du spectre ne peut pas être négative, elle est fixée à zéro, ce qui cause un saut abrupt dans le groupe de fréquences affecté et introduit une tonalité arbitraire à bande étroite dans le signal en sortie. [16] Ce type de tonalité est désagréable à l'oreille et nuit significativement à un processus d'ASR. Une approche moins radicale en réduction de bruit mais respectant davantage l'intégrité du signal est le filtre de Wiener, qui estime le signal rehaussé à partir du modèle spectral du bruit en minimisant l'erreur quadratique moyenne entre le signal estimé et le processus idéal [3].

Dans tous les cas, l'estimation du bruit dans une situation réelle est une tâche complexe [13]. Les approches reposant sur un VAD sont relativement limitées, étant donné les performances variables d'un VAD en présence de conditions défavorables et la supposition d'un bruit stationnaire. Une solution plus générale d'estimation de bruit est l'algorithme de moyenne récursive contrôlée par minima (*Minima Controlled Recursive Averaging* – MCRA), qui consiste essentiellement en une moyenne des échantillons passés de puissance spectrale avec un paramètre de lissage dépendant de la probabilité de présence de parole dans chaque sous-bande [12]. L'approche MCRA est intéressante pour sa faible complexité de calcul et est encore utile de nos jours dans certains scénarios [33]. Son utilisation avec un filtre de Wiener a d'ailleurs été étudiée en profondeur [15, 3]. Toutefois, l'algorithme MCRA rend une performance limitée lors de la présence de bruit non-stationnaire.

Les approches d'estimation spectrale et de filtrage mentionnées jusqu'à présent partagent les mêmes faiblesses face à du bruit non-stationnaire. Une approche plus adaptée au rehaussement contre ce type de bruit est la séparation des sources. En effet, la décomposition d'un signal en plusieurs composantes, une ou plusieurs constituant de la parole et les autres constituant du bruit, permet de considérer des modèles de bruit plus complexes qui sont plus adaptés aux bruits non-stationnaires. Les défis CHiME s'intéressent à cette problématique depuis 2006 [5, 4, 25, 10] et ont mené à la proposition de plusieurs solutions innovantes.

- La factorisation de matrices non-négatives (*Nonnegative Matrix Factorization* – NMF) est une méthode d'apprentissage machine à complexité de calcul modérée qui, sommairement, sépare un signal mixte [55] en plusieurs composantes qui sont ensuite classifiées entre bruit et parole à partir de dictionnaires obtenus dans une

phase préalable d'entraînement supervisé [46]. La faible complexité de calcul de cette méthode permet l'entraînement par des ressources matérielles limitées, ce qui laisse entrevoir la possibilité de mises à jour automatiques et non-supervisées du dictionnaire de bruit.

– L'utilisation d'un DNN, initialisé par NMF, a également mené à une amélioration du SNR, du rapport signal à distorsion (*Signal to Distortion Ratio* – SDR), et du rapport signal à interférence (*Signal to Interference Ratio* – SIR) par rapport à la séparation réalisée avec NMF seulement [17]. Toutefois, ce gain en performance est possible au coût d'une complexité de calcul élevée durant la phase d'entraînement, ce qui implique un entraînement hors-ligne sur du bruit connu.

2.3.2 Canaux multiples

Plusieurs configurations géométriques de matrices de microphones ont été étudiées, soit linéaire, circulaire, sphérique et arbitraire.

– La disposition **linéaire** a l'inconvénient de comporter une ambigüité dans la localisation des sources, c'est-à-dire que toute source n'étant pas directement sur l'axe de la matrice peut se situer à plusieurs endroits [9]. Cette limitation n'est pas nécessairement problématique, si la localisation est nécessaire sur un seul axe, par exemple.
– En contraste, en plus de ne pas être sujette à cette dernière particularité, une matrice **circulaire** requiert moins d'espace pour offrir une performance équivalente en filtrage spatial [2]. Elle permet la localisation en trois dimensions dans la demi-sphère ayant comme base le côté opposé de l'ouverture des microphones.
– Les matrices **sphériques**, quant à elles, emploient généralement un nombre beaucoup plus élevé de microphones disposés en une configuration équidistante, ce qui donne lieu à des simplifications mathématiques intéressantes pour du filtrage spatial très précis et omnidirectionnel en trois dimensions [40]. Néanmoins, sa géométrie, sa taille et son coût élevé réduisent la polyvalence de ce type de matrices et en restreignent généralement l'application à des études acoustiques en milieu très contrôlé comme une chambre anéchoïque.
– Dans un autre ordre d'idées, si l'espace n'est pas très restreint, comme sur un robot mobile par exemple, il peut être intéressant d'employer une matrice à **géométrie arbitraire**. Contrairement aux autres types de matrices à géométrie fixe qui combinent microphones et contrôleur sur un même circuit, la carte 16SoundsUSB [28] supporte jusqu'à 16 microphones reliés par des câbles, ce qui permet une flexibilité beaucoup plus grande dans la disposition des microphones. Cela permet notamment

de se conformer à un support de forme arbitraire et de maximiser l'écart entre les microphones, ce qui peut améliorer les performances du filtrage spatial [43].

Le tableau 2.1 compare quelques matrices commerciales. Les microphones y sont encerclés en rouge. Par souci de concision, seulement des matrices circulaires sont présentées puisque cette configuration est la plus adaptée aux requis de ce projet, par son efficacité sans ambigüité dans une solution compacte. Dans le contexte d'une matrice de microphones, l'ouverture (*aperture*) se définit par la distance la plus grande entre une paire de microphones.

Un des avantages de disposer de canaux multiples est la possibilité d'en extraire des informations sur la position des sources. Dans les explications qui suivent, une source est considérée en champ lointain (*far-field*) à partir d'une distance entre la source et la matrice environ cinq fois plus grande que l'ouverture de celle-ci [30]. Un signal émis par une source, dont la position est fixe et en champ lointain, est perçu par chaque canal avec une forme d'onde très semblable mais décalée dans le temps d'un délai proportionnel à la distance entre les microphones. Une estimation du délai temporel d'arrivée (*Time Delay of Arrival* – TDoA) permet par la suite d'estimer le vecteur de direction d'arrivée (*Direction of Arrival* – DoA) [54].

Une approche intuitive à l'estimation du TDoA sur une paire de microphones est d'effectuer une corrélation croisée entre les deux canaux, ce qui produira une valeur maximale au délai d'intérêt pour une source cohérente en environnement propre [42]. Cette approche est cependant vulnérable à la présence de réflexions ou de composantes périodiques indésirables dans le signal. La solution la plus populaire à ce problème est la corrélation croisée généralisée (*Generalized Cross-Correlation* – GCC), qui est obtenue par le fenêtrage, ou préfiltrage, du spectre croisé dans le domaine des fréquences (*cross-power spectrum*) [34]. Plusieurs fonctions de préfiltrage ont été étudiées [29], dont la plus populaire est la transformation de phase (*PHAse Transform* – PHAT). La combinaison GCC-PHAT est d'ailleurs à la base de plusieurs algorithmes de localisation des sources pour un plus grand nombre de canaux, notamment celui de puissance en réponse dirigée avec transformation de phase (*Steered-Response Power PHAse Transform* – SRP-PHAT) [53]. La méthode SRP-PHAT est très performante, même en environnement difficiles, et est considérée comme l'état de l'art en terme de localisation. Elle est néanmoins critiquée pour sa complexité de calcul. La méthode nommée SVD-PHAT, qui consiste en la décomposition en valeurs singulières (*Singular Value Decomposition* – SVD), de la matrice de projection SRP-PHAT, a démontré des performances équivalentes à SRP-PHAT avec un coût de calcul significativement réduit [18]. Une méthode basée sur l'erreur quadratique moyenne minimale pondérée (*Weighted*

Tableau 2.1 Exemples de matrices de microphones commerciales

Nom	Image	Microphones	Ouverture	Géométrie
XMOS [1]		7 MEMS	90 mm	Circulaire + centre
ReSpeaker [2]		4 MEMS	64,8 mm	Circulaire
Matrix creator [3]		8 MEMS	104 mm	Circulaire
MiniDSP [4]		7 MEMS	90 mm	Circulaire + centre

2. https://www.xmos.ai/download/xCORE-Microphone-Array-hardware-manual(2V0).pdf
3. https://wiki.seeedstudio.com/ReSpeaker-USB-Mic-Array/
4. https://matrix-io.github.io/matrix-documentation/matrix-creator/resources/microphone/
5. https://www.minidsp.com/products/usb-audio-interface/uma-8-microphone-array

Minimum Mean Square Error – WMMSE) a également été proposée, qui n'était pas tout à fait aussi robuste mais grandement moins complexe que SRP-PHAT [57].

Un autre avantage du traitement de canaux multiples est la possibilité de combiner les canaux de manière à créer sélectivement de l'interférence, qui sera constructive vers une composante d'intérêt et destructive dans les autres directions. Cette pratique est appelée filtrage spatial.

Si la DoA est connue pour la source d'intérêt, un très simple système délai-et-somme (*delay-and-sum* – DS) peut être employé. Il s'agit alors d'appliquer un délai distinct à chaque canal d'entrée pour que la somme de ceux-ci maximisent l'interférence constructive sur la source d'intérêt [21]. Un système DS est par définition un algorithme de filtrage spatial fixe, c'est-à-dire que ses paramètres sont indépendants du signal et ne varient pas à moins de changer les valeurs de DoA/TDoA. Il est donc très dépendant d'une estimation exacte de ces valeurs.

Par opposition, le filtrage spatial MVDR est dépendant du signal : son vecteur varie selon les statistiques des signaux en entrée. En effet, quand un modèle statistique du bruit est connu, représenté sous la forme d'une matrice de covariance spatiale (*Spatial Covariance Matrix* – SCM), un système MVDR vise à minimiser la puissance des composantes indésirables du signal en entrée tout en s'assurant de ne pas introduire de distorsion dans le signal en sortie [14]. Le filtrage MVDR peut être implémenté de deux façons. La première utilise la matrice de covariance de l'interférence, notée Φ_{NN}, ainsi que la fonction de transfert acoustique v, et est représentée par l'équation 2.4 :

$$W_{MVDR}(f) = \frac{\Phi_{NN}^{-1}(f)v(f)}{v^H(f)\Phi_{NN}^{-1}(f)v(f)}, \tag{2.4}$$

où ^{H} représente la conjuguée hermitienne. Dans cette implémentation, v est souvent le vecteur de propagation anéchoïque (*anechoic steering vector*), le même vecteur utilisé pour calculer un filtrage DS. La seconde implémentation emploie Φ_{NN} ainsi que Φ_{XX}, la matrice de covariance de la cible :

$$W_{MVDR}(f) = \frac{\Phi_{NN}^{-1}(f)\Phi_{XX}(f)}{tr(\Phi_{NN}^{-1}(f)\Phi_{XX}(f))}u, \tag{2.5}$$

où tr représente la trace de la matrice et u est un vecteur binaire *one-hot* représentant le canal de référence.

Lorsqu'un algorithme de filtrage spatial utilise le vecteur de propagation anéchoïque, sa performance se dégrade en présence de réflexions produites par des environnements réverbérants. De la même façon que l'implémentation de MVDR présentée à l'équation 2.5, cette limitation est palliée par le filtrage spatial GEV, qui ignore la DoA et opère uniquement à partir de Φ_{NN} et de Φ_{XX}, comme suit [20] :

$$\mathbf{F}_{\text{GEV}}(f) = \mathcal{P}\{\mathbf{\Phi}_{\mathbf{NN}}^{-1}(f)\mathbf{\Phi}_{\mathbf{XX}}(f)\}. \tag{2.6}$$

où $\mathcal{P}$ représente la décomposition en valeurs propres pour obtenir le vecteur propre lié à la composante principale.

TABLEAU 2.2 Comparaison des méthodes de filtrage spatial

Méthode	DoA	Matrice de covariance	Décomposition en valeurs propres	Complexité de calcul
DS	Oui	Non	Non	faible
MVDR	Oui	Oui	Non	élevée
GEV	Non	Oui	Oui	élevée

Le tableau 2.2 compare sommairement les méthodes de filtrage spatial présentées. Le rehaussement serait trivial si les SCMs exactes étaient connues à l'avance (i.e., calculées à partir des signaux originaux). Puisque seul un signal mixte est disponible lors d'une réelle situation de rehaussement, un masque temps-fréquence représentant la cible est appliqué à la représentation de Fourier en temps court (*Short Time Fourier Transform* – STFT) du signal mixte pour en simuler la séparation, idem pour l'interférence. Ce sont les deux signaux masqués qui sont ensuite utilisés pour estimer les SCMs. Dans ce type d'implémentation, l'estimation des masques constitue le facteur déterminant de la qualité du rehaussement.

Récemment, les réseaux de neurones à mémoire à court et long terme (*Long Short-Term Memory* – LSTM) ont démontré des résultats particulièrement prometteurs pour l'estimation de ces masques, pour l'utilisation avec le filtrage GEV [25]. Contrairement aux réseaux standards à propagation avant (*feedforward*), les réseaux LSTM, un sous-ensemble des réseaux de neurones récurrents (*Recurrent Neural Network* – RNN), ont des liens de propagation arrière (*feedback*) qui leur donnent un contexte à long terme permettant l'analyse de longues séquences de données telles qu'un signal audio. Néanmoins, cette méthode implique une phase d'entraînement supervisé d'une complexité de calcul importante sur un grand ensemble de données de bruit connu. Un réentraînement adaptatif automatique sur le terrain n'est donc pas viable par un système embarqué, ce qui rend l'utilisation d'un

rehaussement par réseau LSTM moins intéressant s'il y a nécessité de rehausser contre des bruits inconnus.

En somme, les conditions acoustiques de bruit et de réverbération dégradent grandement les performances d'un système de reconnaissance de la parole. Dans un environnement industriel où ces deux composantes sont très présentes, il est indispensable d'intégrer un système de rehaussement adapté en amont du système de reconnaissance, pour lui fournir un signal plus près d'un environnement acoustique idéal. En captant le signal audio par une matrice de microphones, il peut être rehaussé par *beamforming*, une approche dont la performance est supérieure aux approches à canal unique. Parmi les approches de *beamforming*, les algorithmes MVDR et GEV, basées sur les SCM du signal cible et du signal d'interférence, sont performants dans les environnements très bruités et réverbérants tels les milieux industriels. Ces méthodes requièrent une estimation préalable des SCM, opération typiquement réalisée par une approche par apprentissage machine.

CHAPITRE 3

REHAUSSEMENT DE LA PAROLE PAR KISS-GEV

Avant-propos

Auteurs et affiliation :

Thomas Bernard : étudiant à la maîtrise, Université de Sherbrooke, Faculté de génie, Département de génie électrique et de génie informatique.
François Grondin : professeur, Université de Sherbrooke, Faculté de génie, Département de génie électrique et de génie informatique.

Date de soumission : 6 octobre 2021

Conférence :

IEEE International Conference on Acoustics, Speech and Signal Processing

Titre français :

Rehaussement de la parole à faible coût en environnements bruités et réverbérants par filtrage spatial KISS-GEV

Contribution au document :

Cet article détaille l'algorithme de rehaussement KISS-GEV, qui combine la contribution originale du pré-traitement KISS estimant les matrices de covariance représentant la cible et l'interférence, avec le filtrage spatial GEV dont les performances sont déjà démontrées dans le domaine du rehaussement de la parole. L'article met en contexte la problématique que KISS-GEV vise à combler, ainsi que le raisonnement qui a mené à choisir une approche de traitement de signal plutôt que d'employer un réseau de neurones comme les approches plus répandues. Il fait également ressortir le caractère intuitif de la solution proposée, comparativement aux approches par DNN qui s'apparentent davantage à une « boîte noire ». Les résultats démontrent que la méthode KISS-GEV rend une amélioration du SDR significativement supérieure au populaire filtrage DS sur trois types d'interférence inconnues différents, soit du bruit stationnaire et non stationnaire, de la musique et de la parole, tout en gardant une faible complexité de calcul.

Résumé français :

Cet article introduit une nouvelle méthode de filtrage spatial nommée KISS-GEV (pour *Keep It Super Simple Generalized EigenValue* en anglais, ou « décomposition en valeurs propres gardée super simple »). Alors que le filtrage spatial GEV se fie généralement sur des réseaux de neurones pour estimer les masques temps-fréquences de la cible et de l'interférence, cette méthode utilise plutôt une approche de traitement de signal basée sur la direction d'arrivée (*Direction of Arrival* – DoA) de la cible. Cela réduit considérablement la complexité des calculs impliqués, et offre un bon rendement de rehaussement de la parole en conditions inconnues puisque la méthode ne dépend pas d'un réseau de neurones entraîné sur de la parole bruitée. L'approche proposée peut également être utilisée pour séparer la parole d'un mélange dans la mesure où les sources proviennent de directions distinctes. Les résultats obtenus démontrent que la méthode proposée surpasse le filtrage spatial traditionnel délai-et-somme, bien qu'elle se base sur la même supposition minimale de DoA.

Note :

À la suite des corrections demandées par les membres du jury, le contenu de cet article diffère de celui qui a été soumis.

3.1 Abstract

This paper introduces a new method referred to as KISS-GEV (Keep It Super Simple Generalized EigenValue) beamforming. While GEV beamforming usually relies on deep neural network for estimating target and noise time-frequency masks, this method uses a signal processing approach based on the direction of arrival (DoA) of the target. This considerably reduces the amount of computations involved at test time, and works for speech enhancement in unseen conditions as there is no need to train a neural network with noisy speech. The proposed method can also be used to separate speech from a mixture, provided the speech sources come from different directions. Results also show that the proposed method uses the same minimal DoA assumption as Delay-and-Sum beamforming, yet outperforms this traditional approach.

Keywords : Beamforming, Masking, Direction of Arrival

3.2 Introduction

Distant speech processing is a challenging task as the target speech is usually corrupted by reverberation and interfering noise sources from the environment [52]. Training an automatic speech recognition (ASR) system in the same environmental conditions is desirable, yet remains a challenging task when the additive noise and reverberation to be observed at test time is unknown [49]. Enhancing speech with a microphone array prior to the ASR task stands as a simple solution to cope with the training and testing conditions mismatch [10].

Delay-and-Sum (DS) and Minimum Variance Distortionless Response (MVDR) beamformers use the direction of arrival (DoA) of the target sound source to enhance speech [11, 23, 14, 56]. However, they rely on the anechoic assumption, i.e., free field propagation of sound, which is an inaccurate approximation for reverberant environments. Alternatively, generalized eigenvalue (GEV) beamforming can enhance speech using only estimations of the speech and noise spatial covariance matrices (SCMs), based on an estimated time-frequency mask to distinguish speech from background noise. A neural network usually estimates this time-frequency mask, which implies the background interference type has to be known in advance at training time [25, 26, 24, 20]. While this can yield satisfactory performance when trained and evaluated on large synthetic datasets [19], this approach remains challenging in real environments with unknown interference.

The proposed method, referred to as KISS-GEV (Keep It Super Simple GEV), assumes the DoA of the target signal is obtained using a sound localization method [18] or audio-visual fusion [22]. The idea behind KISS-GEV is to provide a mask estimation method that

is computationally much lighter than state-of-the-art machine learning-based approaches, while providing adequate enhancement in real-world scenarios with unseen noise, making it viable to run on low-cost embedded hardware. Results demonstrate that this method works without training a neural network to predict a time-frequency mask to differentiate speech from noise, which makes it ideal for enhancement in unseen conditions. The proposed method can also be used to separate speech from a mixture, provided the speech sources come from different directions. Results also suggest that KISS-GEV uses the same minimal DoA assumption as DS beamforming, yet outperforms this traditional approach.

3.3 KISS-GEV

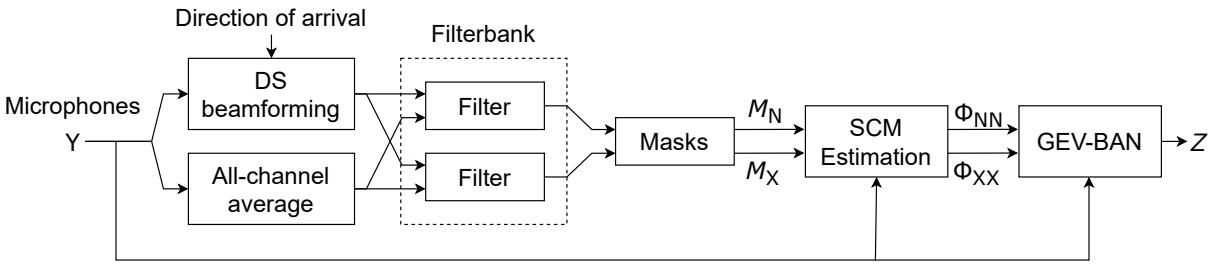


FIGURE 3.1 Overview of the KISS-GEV processing pipeline

Figure 3.1 illustrates the KISS-GEV pipeline. KISS-GEV compares the power of the signal obtained with DS beamforming pointing in the direction of the target against the average power from all channels. A two-filter bank then computes the power in low and high frequency regions, and generates a coarse binary time-frequency mask representing the target signal. These two broadband filters are suitable for reverberant environments as they can cope with early reflections, which usually significantly impact the spectral shape of the DS beamformed signal. This idea is similar to the Generalized Cross-Correlation with Phase Transform (GCC-PHAT) approach, where analyzing the normalized broadband spectrum makes time difference of arrival (TDoA) estimation more robust to early reflections [7]. With $\mathbf{X}$ representing the target and $\mathbf{N}$, the interference, the masks obtained with the filterbank ($M_{\mathbf{X}}$ and $M_{\mathbf{N}}$) then provide an estimation of the target and interference spatial covariance matrices, denoted as $\Phi_{\mathbf{XX}}$ and $\Phi_{\mathbf{NN}}$, respectively. The two SCMs are finally used to perform GEV beamforming.

The Short-Time Fourier Transform (STFT) is computed for the signal captured by each microphone $d \in \{1, 2, \ldots, D\}$ (where D denotes the number of microphones) to generate the time-frequency frames $Y_d(t, f) \in \mathbb{C}$, with frames of $N \in 2\mathbb{N}$ samples in the time-domain, where $t \in \mathbb{N}$ stands for the frame index and $f \in \{0, 1, \ldots, N/2\}$ the frequency

bin index. The proposed approach then defines a binary filterbank made of $B \in \mathbb{N}$ filters, where $b \in \{1, 2, \ldots, B\}$ is the filter index :

$$H_b(f) = \begin{cases} 1 & l_b \leq f \leq u_b \\ 0 & \text{otherwise} \end{cases}, \tag{3.1}$$

where $l_b, u_b \in \{0, 1, \ldots, N/2\}$ represent the lower and the upper bounds of the filter b, respectively. In this paper, we assume there is no overlap between adjacent filters ($u_b < l_{b+1}$), and the spectrum is fully covered by the filters in the filterbank ($l_1 = 0$, $l_{b+1} = u_b + 1$ and $u_B = N/2$). More specifically, we restrict the number of filters to $B = 2$ (to intuitively capture either low-frequency voiced phonemes or high-frequency fricatives), which implies the only parameter of the filterbank is the separator position $\gamma = l_2 = u_1 + 1$, as shown in Figure 3.2.

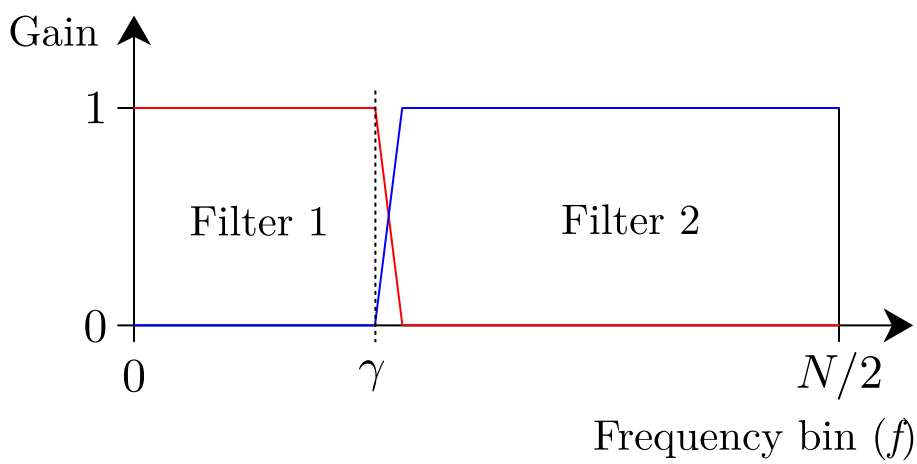


FIGURE 3.2 Two-filter filterbank, with first filter ($b = 1$) in red, and second filter ($b = 2$) in blue.

The delay-and-sum beamformer is computed using the anechoic steering vector $W_d(f) \in \mathbb{C}$ towards the target :

$$W_d(f) = \exp\left(j2\pi f \tau_d / N\right), \tag{3.2}$$

where $\tau_d \in \mathbb{R}$ is the Time Delay of Arrival in samples, obtained from the known array geometry and target DoA.

The power of the beamformed signal corresponds to $|\sum_d W_d(f) Y_d(t, f)|^2$ for each frequency f, while the sum of the power of each channel is $\sum_d |Y_d(t, f)|^2$. It is then possible to compute the total beamformed power, as well as the average power, for each filter b in the filterbank, by multiplying by the expression $H_b(f)$ and summing across all frequencies. The ratio $R_b(t)$ of the beamformed and average power for each filterbank is hence defined

as :

$$R_b(t) = \frac{\sum_f H_b(f) |\sum_d W_d(f) Y_d(t,f)|^2}{D \sum_f H_b(f) \sum_d |Y_d(t,f)|^2}. \tag{3.3}$$

where the constant D is introduced for normalization purpose.

In other words, when the frequency region covered by the filter b is dominated by the target signal, we expect the ratio to get closer to a value of 1. Similarly, when there is little power from the target signal in the region covered by filter b, the ratio gets closer to a value of 0. The complete time-frequency ratio for frame t and bin f is then obtained by summing every R_b with the filterbank :

$$R(t,f) = \sum_b R_b(t) H_b(f). \tag{3.4}$$

Then, only both extremities of $R(t,f)$, with a width defined by parameter α, are kept to define binary masks in order to capture only the most significant spectral features. Thresholds $T_\nu(f)$ are defined as follows :

$$T_\nu(f) = \begin{cases} P_{100-\alpha}(R(f)) & \text{if } \nu = \mathbf{X} \\ P_\alpha(R(f)) & \text{if } \nu = \mathbf{N} \end{cases}, \tag{3.5}$$

where P_i is the i-th percentile of R for bin f across all t.

Using these thresholds, the binary time-frequency masks $M_\nu(t,f)$ are defined as :

$$M_\nu(t,f) = \begin{cases} 1 & \text{if } R(t,f) > T_{\mathbf{X}}(f), \nu = \mathbf{X} \\ 1 & \text{if } R(t,f) < T_{\mathbf{N}}(f), \nu = \mathbf{N} \\ 0 & \text{otherwise} \end{cases}, \tag{3.6}$$

With the masks defined, both SCMs can be estimated as follows :

$$\mathbf{\Phi}_{\nu\nu}(f) = \sum_t M_\nu(t,f) \mathbf{Y}(t,f) \mathbf{Y}(t,f)^{\mathrm{H}}, \tag{3.7}$$

where $\mathbf{Y}(t,f) \in \mathbb{C}^{D\times 1}$ stands for a vector that concatenates the time-frequency frames $Y_d(t,f) \in \mathbb{C}$ of all microphones, and $(\dots)^H$ stands for the Hermitian operator.

The beamforming vector $\mathbf{F}_{\text{GEV}}(f) \in \mathbb{C}^{D\times 1}$ is then obtained by performing the eigenvalue decomposition of the target and interference SCMs :

$$\mathbf{F}_{\text{GEV}}(f) = \mathcal{P}\{\mathbf{\Phi}_{\mathbf{NN}}^{-1}(f)\mathbf{\Phi}_{\mathbf{XX}}(f)\}. \tag{3.8}$$

While the target and noise masks estimated with the filterbank have a poor frequency resolution, this remains acceptable for estimating the target and noise SCMs, as the eigenvalue decomposition leads to the same eigenvector as long as the target SCM captures more energy from the target signal than the noise signal, and vice-versa.

Finally, as demonstrated by Heymann et al. [25], a Blind Analytic Normalization (BAN) gain, defined as :

$$g_{\text{BAN}}(f) = \frac{\sqrt{\mathbf{F}_{\text{GEV}}^{\text{H}}(f)\mathbf{\Phi}_{\mathbf{NN}}(f)\mathbf{\Phi}_{\mathbf{NN}}(f)\mathbf{F}_{\text{GEV}}(f)}}{D^2\mathbf{F}_{\text{GEV}}^{\text{H}}(f)\mathbf{\Phi}_{\mathbf{NN}}(f)\mathbf{F}_{\text{GEV}}(f)}, \tag{3.9}$$

can be applied as a post-filter in order to reduce the distortion that can be introduced by the GEV beamformer in the target direction. The enhanced STFT representation of the target signal is then obtained as follows :

$$Z(t,f) = g_{\text{BAN}}(f)\mathbf{F}_{\text{GEV}}^{H}(f)\mathbf{Y}(t,f). \tag{3.10}$$

3.4 Results

KISS-GEV's performance is evaluated by calculating the Signal-to-Distortion Ratio (SDR) improvement of the enhanced signal on a dataset containing 20 target speech tracks, 20 different speech tracks to use as interference, 20 ambient noise tracks with both stationary and non-stationary noise and 20 music tracks. Every combination of target and interference was simulated on 5 different room impulse responses (RIRs). Each simulation was performed with a new RIR, with none being reused between configurations, for a total of 6000 unique RIRs. The RIRs were simulated using the image method [1] with the geometry of a ReSpeaker Core v2 microphone array as the receiver, as well as varying uniformly distributed parameters as in the BIRD dataset [19], such as the room dimensions (width between 5 and 15 m, length between 5 and 15 m, and height between 3 and 4 m), absorption coefficients (between 0.2 and 0.8), speed of sound (between 340 and 355 m/sec), and target and interference source positions. The speech segments are from the Librispeech [38] dataset, while the noise and music tracks were selected from the Musan [50] dataset.

A value of $\gamma = 100$ is defined to separate between voiced and unvoiced spectrum segments (which represents a separation at 3125 Hz), and $\alpha = 25$ is used as an adequate compromise between good target/interference separation and having enough data to properly calculate the SCM.

The mean SDR was calculated for the four following outputs : unprocessed mixture (from channel 1), DS beamforming, KISS-GEV beamforming, and GEV beamforming with Oracle ideal ratio mask (IRM) [37]. Table 3.1 presents the results of this experiment.

Tableau 3.1 SDR (in dB) in various noise conditions

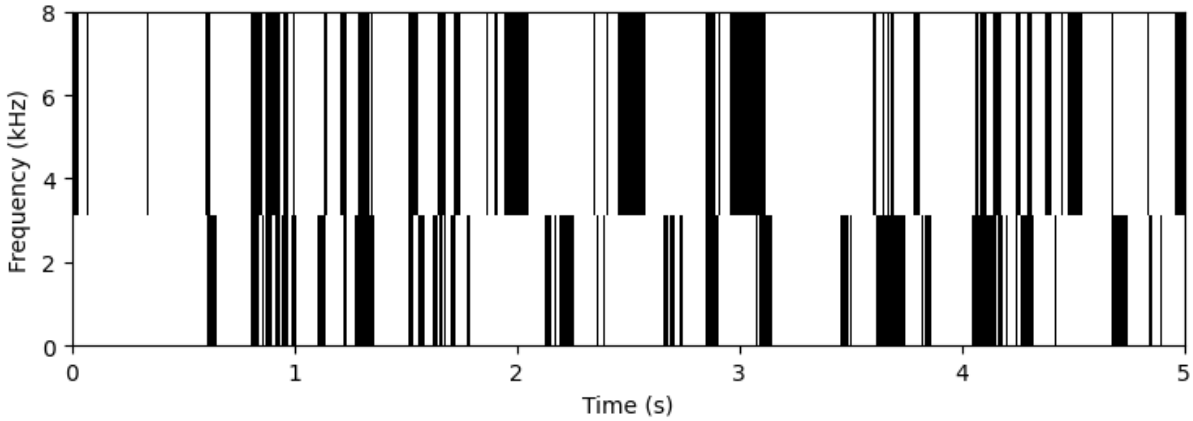
Method	Types of interference		
	Ambiant	Music	Speech
Unprocessed	4.05	3.92	8.76
DS	5.69	4.36	9.06
KISS-GEV	10.31	10.42	13.09
GEV with Oracle IRM	16.98	17.40	18.58

KISS-GEV significantly outperforms DS thanks to its ability to null interference. The remaining gap between KISS and Oracle IRM is likely explained by the loss of spectral information to interference.

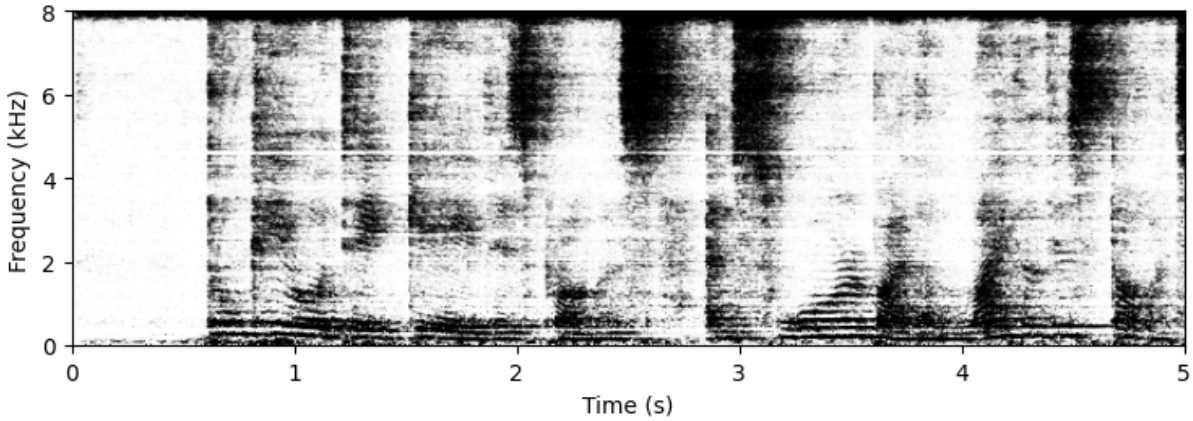
Figure 3.3 contrasts the time-frequency mask generated with KISS-GEV and Oracle IRM, on a single simulation with stationary noise. Apart from showing how a coarse mask still leads to effective results for SCM estimation, it also suggests dereverberation can be achieved, as it only takes into account the direct path from the target and ignores the late reverberation.

Figure 3.4 shows example spectrograms of the four outputs on the same simulation as Figure 3.3. It can be observed that although it gives a similar noise floor to DS, KISS-GEV yields a much clearer resolution of the harmonics in voiced segments, as well as a more robust dereverberation than the other methods, which is especially visible on the fricatives. It also removes more white noise in the lower frequencies than DS.

Figure 3.5 shows spectrograms of a simulation using speech as interference on the same target utterance as Figure 3.4, where a sharp dereverberation is also demonstrated, while offering a significantly better attenuation of the interference voiced segments than DS. This demonstrates that using KISS-GEV with informed DoA can enhance speech in the direction of interest and deal with the permutation issue observed in speech separation based on neural networks.



(a) Binary mask generated with KISS-GEV

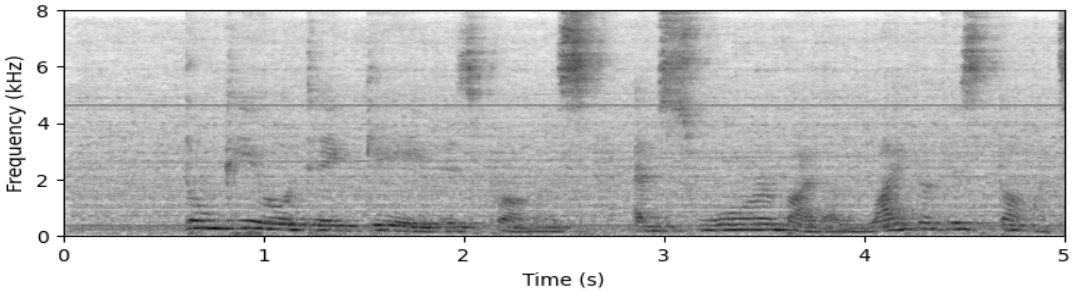


(b) GEV with Oracle IRM

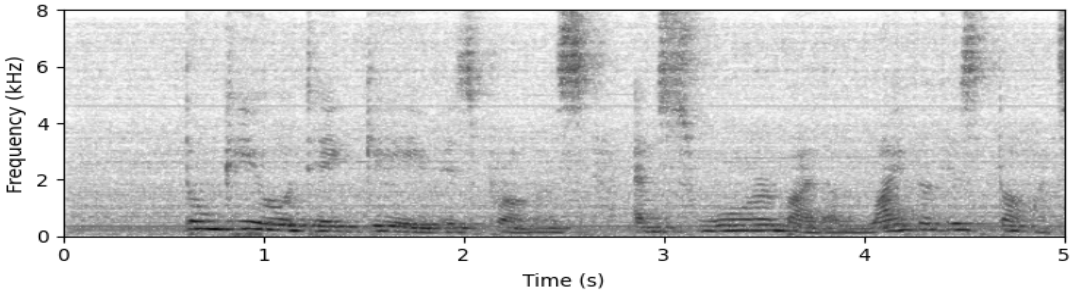
FIGURE 3.3 Time-frequency masks (the color black stands for a value of 1, and white for a value of 0)

3.5 Conclusion

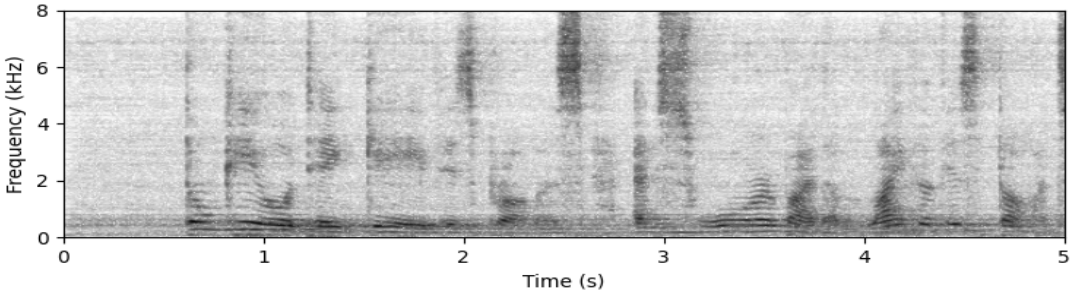
This paper presents KISS-GEV, a lightweight mask estimation front-end to generate target and interference SCMs for GEV-BAN beamforming that requires no training, hence performs well when enhancing against unseen interference. Results show a significantly better SDR than the popular DS beamformer while relying only on the same target DoA assumption. This approach is thus ideal for low-cost embedded hardware deployed in real-life environments. As the end use of this method is speech processing, Word Error Rate (WER) improvement should also be evaluated with an ASR backend, such as Kaldi [41]. Future work could also involve estimating the masks and SCMs in an online manner for real-time applications, as well as comparing enhancement and computational performance with a state-of-the-art DNN-based approach.



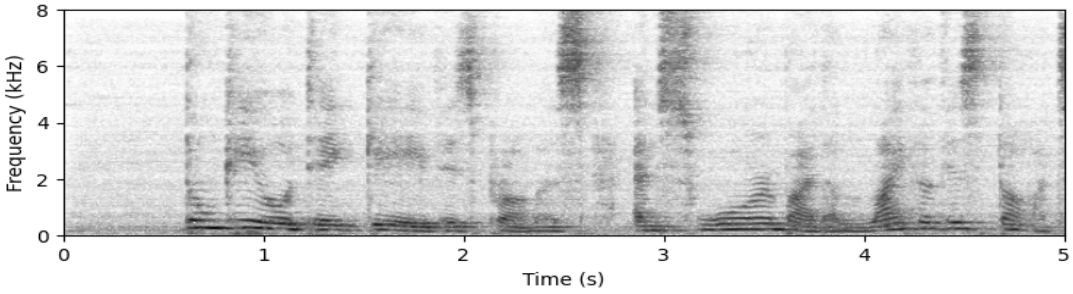
(a) Unprocessed signal from channel 1 (SDR 11.65 dB)



(b) DS beamforming (SDR 13.54 dB)

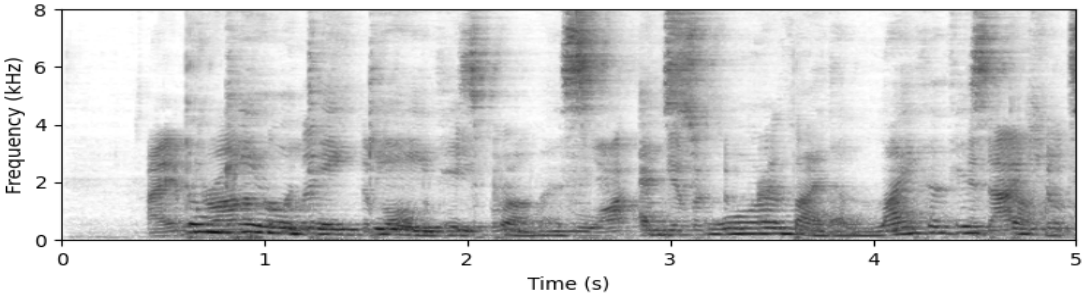


(c) KISS-GEV beamforming (SDR 19.33 dB)

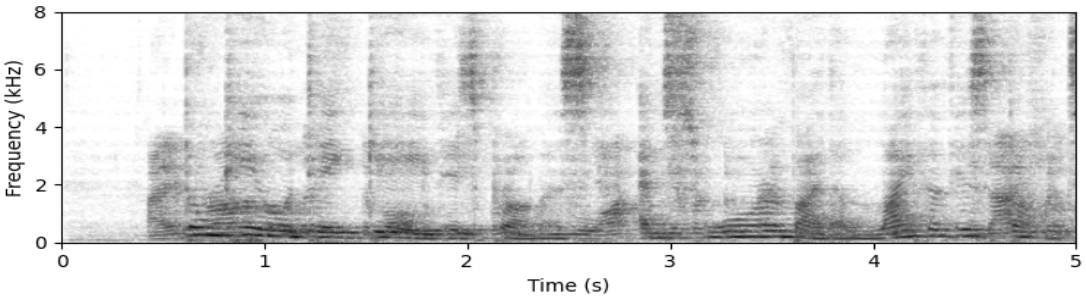


(d) GEV with Oracle IRM (SDR 24.28 dB)

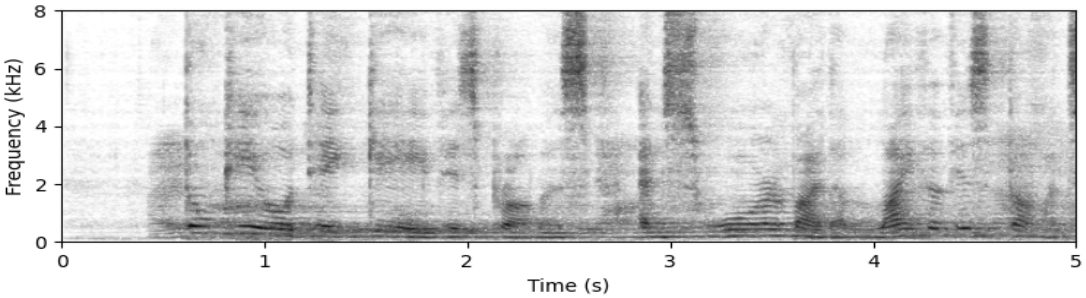
FIGURE 3.4 Spectrograms of the enhanced mixture, noise interference. Note the variation in the sharpness of upper harmonics (1.0 - 2.5 kHz) around 1 s mark and of fricative (4.0 - 8.0 kHz) at 2.5 s.



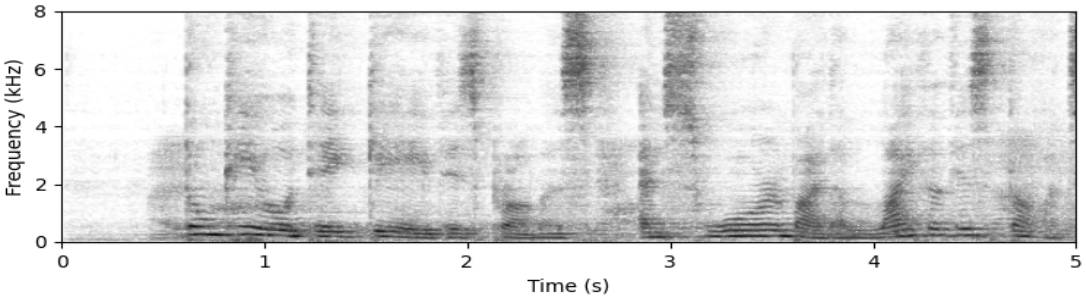
(a) Unprocessed signal from channel 1 (SDR 7.86 dB)



(b) DS beamforming (SDR 9.77 dB)



(c) KISS-GEV beamforming (SDR 14.94 dB)



(d) GEV with Oracle IRM (SDR 21.18 dB)

FIGURE 3.5 Spectrograms of the enhanced mixture, speech interference. Note the attenuation of voiced interference (0.0 - 4.0 kHz) segment before target utterance starts, around 0.6 s, and at 4.8 sec.

CHAPITRE 4

PERFORMANCES D'ASR DANS UN ENVIRONNEMENT INDUSTRIEL

Alors que les résultats présentés au chapitre 3 démontrent la performance de KISS-GEV en rehaussement de la parole en terme de SDR dans un environnement simulé, le présent chapitre détaille l'évaluation des performances de la méthode KISS avec GEV et MVDR, en-ligne et hors-ligne. L'expérimentation est réalisée en environnement industriel réel, en évaluant le WER des transcriptions fournies par *Google Speech-to-text* avant et après rehaussement.

4.1 Méthodologie

L'expérimentation est réalisée sur la plateforme matérielle ReSpeaker Core v2.0 [1], illustrée à la figure 4.1, qui comporte une matrice circulaire de six microphones d'une ouverture de 92,6 mm ainsi qu'un système embarqué (*System on a Chip*) Rockchip RK3229 (cœur ARM Cortex-A7), opéré avec un noyau Linux (Debian 10).

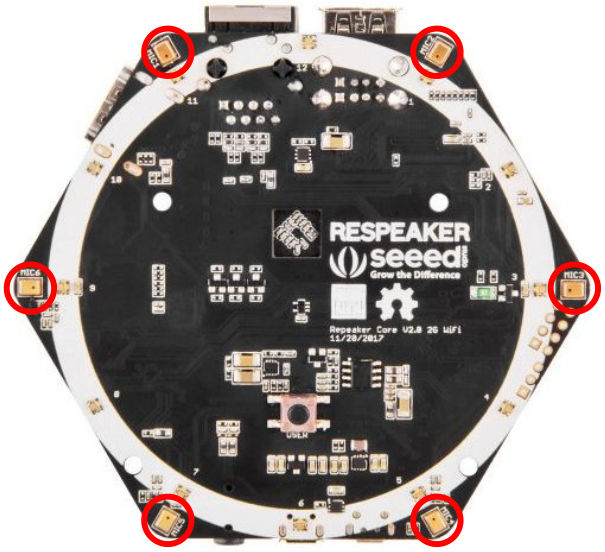
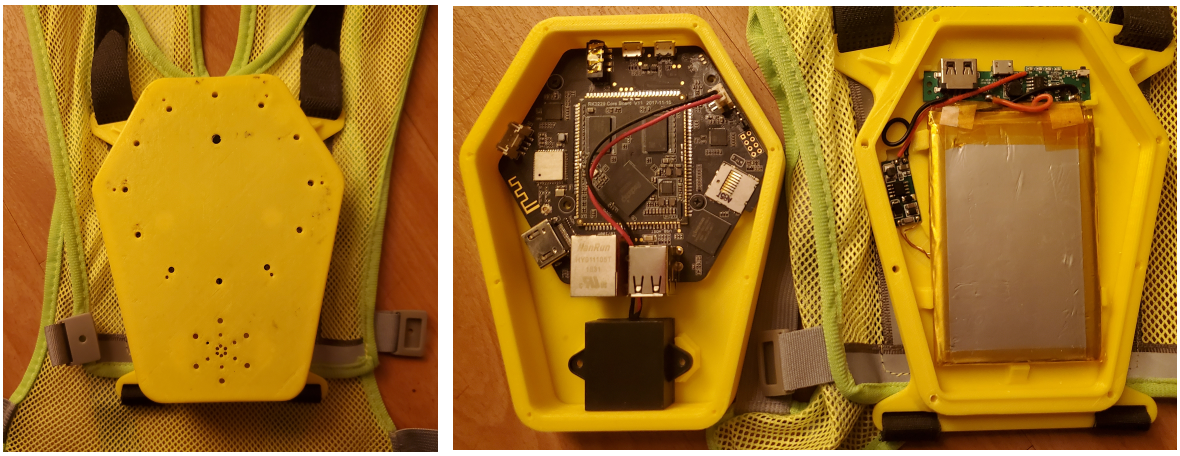


FIGURE 4.1 ReSpeaker Core v2.0
(microphones encerclés en rouge)

1. https://wiki.seeedstudio.com/ReSpeaker_Core_v2.0/

La matrice est contenue dans un boîtier imprimé en 3D, montré à la figure 4.2, conçu pour s'attacher à l'avant d'un gilet de sécurité réfléchissant.



(a) Extérieur (b) Intérieur

FIGURE 4.2 Boîtier pour gilet de sécurité imprimé en 3D

Le gilet portant le boîtier est placé sur un mannequin d'une grandeur originale de 1950 mm, dont la tête a été remplacée par un haut-parleur. Le centre du haut-parleur se trouve à une distance de 1735 mm du sol, soit 385 mm au-dessus du centre de la matrice de microphones. Le système est montré à la figure 4.3.

Un ensemble d'extraits audio de l'ensemble LibriSpeech [38] est sélectionné, regroupant quatre locuteurs différents. Chacun des quatre groupes contient exactement 60 secondes de parole dont les caractéristiques sont les suivantes :

- Homme, lexique complexe (littéraire) ;
- Homme, lexique simple (descriptions touristiques) ;
- Femme, lexique complexe (littéraire) ;
- Femme, lexique simple (recettes).

Quatre emplacements d'un complexe industriel sont sélectionnés pour effectuer les enregistrements, dont les spécifications sont présentées au tableau 4.1. Bien qu'une représentation plus formelle de la réverbération serait idéale, l'impossibilité d'éteindre les équipements bruyants lors des enregistrements a empêché de prendre les mesures nécessaires. Les valeurs présentées au tableau 4.1 sont donc estimées qualitativement par l'écoute des enregistrements.

À chaque emplacement, le mannequin est placé dans une position plausible pour un ouvrier qui travaillerait à cet endroit. L'ensemble d'enregistrements est ensuite joué sur le haut-parleur à la tête du mannequin, puis enregistré à une fréquence de 16 000 échantillons par

FIGURE 4.3 Mannequin

TABLEAU 4.1 Caractéristiques des emplacements sélectionnés

Emplacement	Bruit stationnaire	Réverbération	Niveau de bruit (SPL)
Usine	Oui	Moyenne	72 dB
Ventilation	Partiellement	Faible	68 dB
Génératrice	Oui	Élevée	65 dB
Entrepôt	Oui	Faible	54 dB

seconde par la matrice attachée au gilet. Lorsque le niveau de bruit ambiant le permet, plusieurs niveaux de parole sont enregistrés en faisant varier le gain du haut-parleur. Plus le niveau ambiant est bas, plus de niveaux différents sont enregistrés. Dans le cas contraire, dans l'usine où le bruit est très élevé, un seul niveau est enregistré avec le gain du haut-parleur au maximum, ce qui résulte tout de même à un SNR négatif pour certains locuteurs.

4.1.1 Évaluation des performances

Puisque la méthode KISS présentée au chapitre 3 sert à estimer les matrices de covariance préalables à un filtrage spatial, elle n'est pas uniquement confinée à être utilisée avec GEV. En effet, une implémentation de l'algorithme MVDR se base sur les SCMs de la cible et de l'interférence, tout comme GEV. Ainsi, le même pré-traitement KISS peut être utilisé pour les deux algorithmes. Tous deux sont donc évalués et comparés dans ce chapitre.

La méthode KISS-GEV a été développée et validée en mode hors-ligne, c'est-à-dire qu'elle dispose du signal en entier avant de procéder au rehaussement. Selon l'implémentation désirée, il peut être nécessaire de procéder en-ligne pour obtenir un rehaussement à faible latence et, par exemple, téléverser le signal rehaussé en continu (*stream*) vers un agent conversationnel. La méthode d'estimation de masques KISS a été implémentée au sein de la librairie d'audition PyODAS[2] en mode temps réel, où les masques sont générés pour chaque trame de STFT (fenêtre sinus de 512 trames, saut 256 trames) à partir d'un tampon de longueur paramétrable, actuellement fixé à 30 trames. Les SCMs sont par la suite mises à jour à chaque trame, en ajoutant à la matrice précédente celle de la trame nouvellement calculée avec une pondération paramétrable. Ces matrices sont finalement utilisées pour calculer en temps réel les vecteurs GEV et MVDR.

La contribution de la méthode KISS à la problématique étudiée est donc évaluée à travers quatre algorithmes de rehaussement, soit :

- KISS-GEV hors-ligne ;
- KISS-MVDR hors-ligne ;
- KISS-GEV en-ligne ;
- KISS-MVDR en-ligne.

À titre de référence, tous les enregistrements originaux sont également évalués sans aucun rehaussement, de même qu'avec un filtrage spatial DS.

L'ensemble des bandes audio précédemment décrites sont analysés par le moteur de STT de la plateforme infonuagique *Google Cloud Platform*[3]. Les transcriptions retournées par le service sont ensuite comparées à celles fournies par LibriSpeech à l'aide de la librairie Python *JiWER*[4] pour en obtenir le WER.

Puisque les assistants vocaux peuvent reconnaître un lexique large mais ont une portée finie des commandes qu'ils peuvent traiter, une reconnaissance parfaite de chaque mot constituant une commande n'est pas essentielle pour que celle-ci soit traitée correctement.

2. https://github.com/introlab/pyodas
3. https://cloud.google.com/speech-to-text
4. https://github.com/jitsi/jiwer

Les agents conversationnels réussissent souvent à extrapoler à partir du contexte et de la portée des commandes disponibles pour corriger plusieurs mots incorrectement reconnus à l'étape de STT. Ainsi, pour mesurer rigoureusement la performance d'un système de rehaussement dans ce contexte, il faudrait également évaluer le taux de succès des commandes (*Command Success Rate* – CSR) [36], soit la proportion des commandes correctement traitées. Contrairement au WER pour lequel une certaine marge d'erreur est presque inévitable et peut être mitigée, le CSR doit être presque parfait pour juger qu'un système est adéquat, particulièrement en contexte industriel où une erreur peut engendrer des coûts importants. Ceci dit, la relation entre le WER et le CSR est entièrement dépendante de la nature de l'agent conversationnel employé. Un WER évalué sur du texte générique, comme dans le cas présent, aurait possiblement peu de signification comparativement à un ensemble d'élocutions de commandes destinées à un agent conversationnel spécifique. Puisqu'un tel agent n'était pas disponible au moment des enregistrements et de la rédaction, un seuil arbitraire de WER de 20 % a été fixé pour l'analyse des résultats, en supposant qu'un agent bien entraîné pourrait généralement compléter cette proportion des mots erronés si le reste de la commande est adéquat, et donc tendre vers un CSR de 100 %.

4.2 Résultats

Les figures 4.4 et 4.5 présentent le taux de reconnaissance des mots (*Word Recognition Rate* – WRR) moyen en fonction de l'emplacement et du locuteur, respectivement. Le WRR est choisi plutôt que son inverse, le WER, pour mieux souligner la relation avec le SNR de l'enregistrement original. Ainsi, l'objectif est d'atteindre un WRR s'approchant de 100 %. L'ensemble des résultats obtenus sont présentés à l'annexe A.

La figure 4.4 souligne la relation entre le SNR du signal (et donc le niveau de bruit ambiant présenté au tableau 4.1) et la tendance générale du WER, tel qu'attendu. De surcroît, elle fait également ressortir le rendement décroissant du rehaussement en proportion du SNR d'origine, comme quoi plus le signal est déjà propre pré-rehaussement, et donc le WER déjà élevé, plus l'amélioration du WER par rehaussement est difficile. La figure illustre également que les deux algorithmes hors-ligne offre une amélioration significative dans toutes les situations, et sont les seuls à dépasser le seuil de WRR 80 % même dans les conditions les plus adverses. Dans la même lignée, les résultats à l'emplacement *Usine* donnent un aperçu de l'ampleur de la dégradation de performances qu'occasionne un niveau de bruit aussi fort avec les algorithmes de rehaussement populaires comme DS, alors que même sans rehaussement les enregistrements avec un bruit plus faible surpassent le seuil. D'ailleurs, dans la majorité des cas, la performance de DS est inférieure à celle

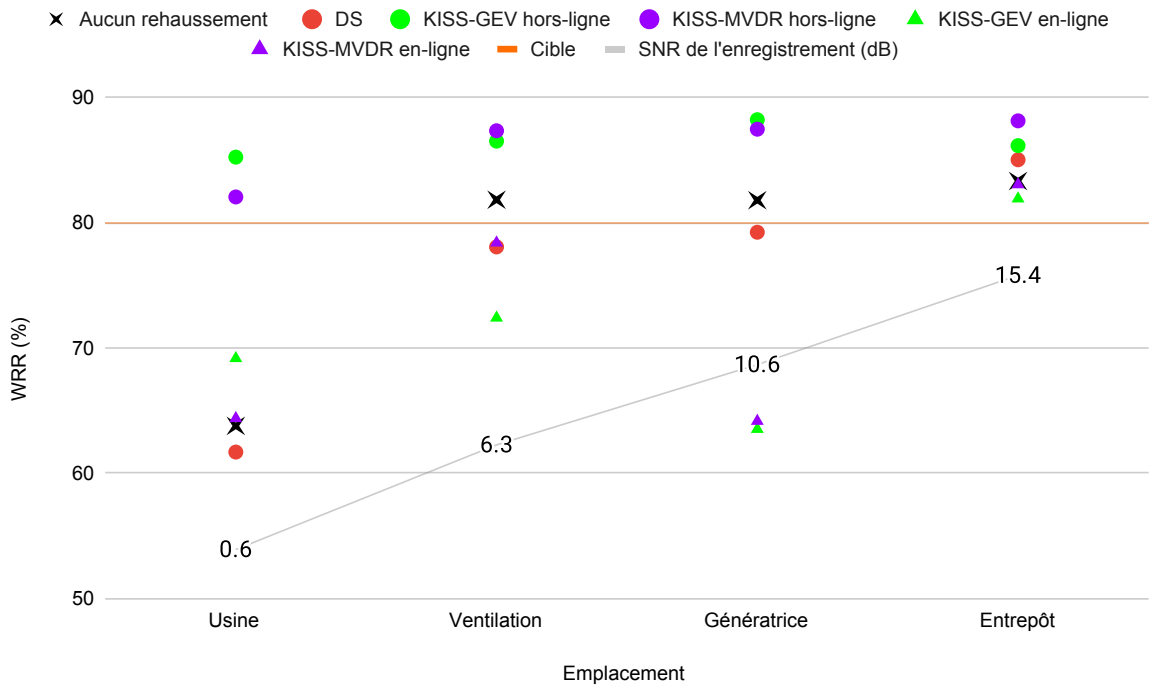


FIGURE 4.4 WRR par algorithme en fonction de l'emplacement

de l'enregistrement original. Ce phénomène s'explique par l'apparition d'artéfacts dans le signal rehaussé, provenant de réflexions qui sont ignorées par l'hypothèse anéchoïque de DS.

Les résultats démontrent également l'augmentation du niveau de difficulté de l'estimation des caractéristiques spectrales nécessaires au filtrage spatial plus avancé lorsqu'il est nécessaire de procéder en-ligne avec un contexte temporel réduit. Bien que la perception à l'oreille humaine des enregistrements filtrés par GEV et MVDR en-ligne soit d'une qualité acceptable, le changement au fil du temps du vecteur de filtrage spatial occasionne des changements de phase qui introduisent des distorsions qui nuisent au STT. Ce problème est particulièrement visible à l'emplacement *Génératrice*, où la forte réverbération et le manque de contexte temporel causent une coupure trop agressive dans les sections voisées.

La figure 4.5, quant à elle, révèle encore une fois la fiabilité de l'amélioration offerte par KISS-GEV et KISS-MVDR en mode hors-ligne. Leurs contreparties en-ligne, cependant, semblent se dégrader significativement avec les locuteurs masculins. Cette lacune peut en partie s'expliquer par le phénomène décrit ci-haut et illustré à la figure 4.6, où la coupure excessive se situe principalement entre 0 et 1 kHz. Puisque les harmoniques des voix masculines sont en plus basses fréquences, il est possible qu'elles se fassent couper en

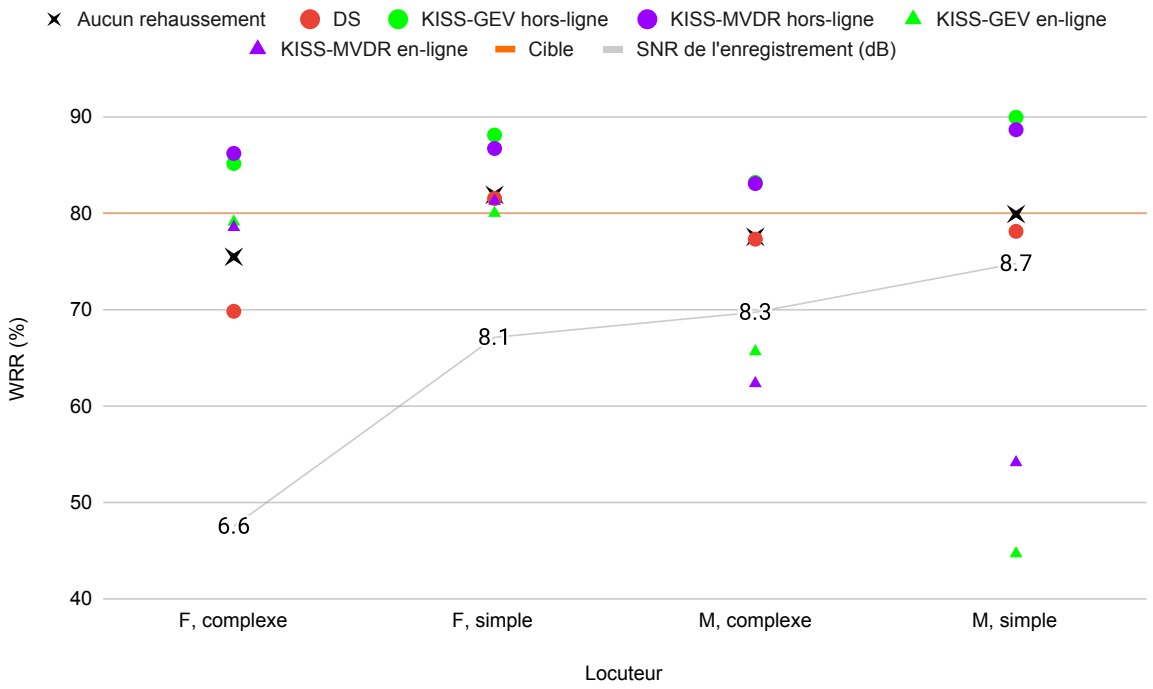


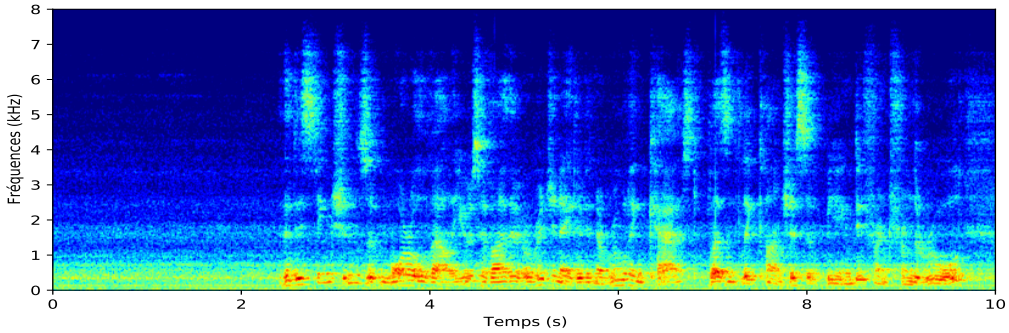
FIGURE 4.5 WRR par algorithme en fonction du locuteur

plus grande partie que celles des voix féminines, qui conservent davantage d'informations au-dessus de 1 kHz.

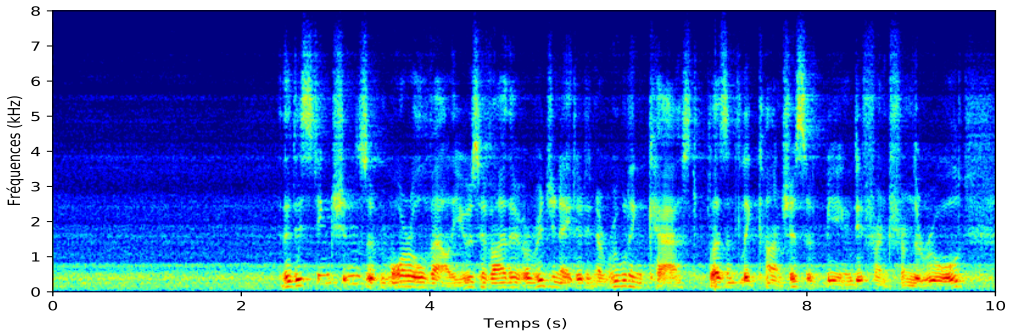
En somme, les enregistrements rehaussés par les algorithmes KISS-GEV et KISS-MVDR hors-ligne surpassent le seuil de WRR de 80% sans exception, et présentent une amélioration notable des performances en reconnaissance de la parole même dans les conditions les plus difficiles. Leurs analogues en-ligne, en contrepartie, fournissent une amélioration non-négligeable en conditions très bruités, mais souffrent dans certaines conditions spécifiques, comme un faible niveau de bruit mélangé à une forte réverbération ou encore pour les locuteurs masculins. Ces résultats font ressortir la difficulté de mettre en œuvre un algorithme qui rehausse en temps réel avec peu de contexte temporel, et laisse place à de nouvelles perspectives de recherche.

Puisque l'objectif du présent projet est de mettre en œuvre une méthode de rehaussement pour permettre l'usage d'un agent conversationnel, l'usage des algorithmes hors-ligne demeure une option viable dont les performances sont prometteuses. L'écoute est normalement activée par une approche par mot de réveil dont le taux de reconnaissance est très élevé même dans les pires conditions, les commandes envoyées sont relativement courtes, et il est commun chez les assistants vocaux populaires actuels d'avoir un délai de réponse

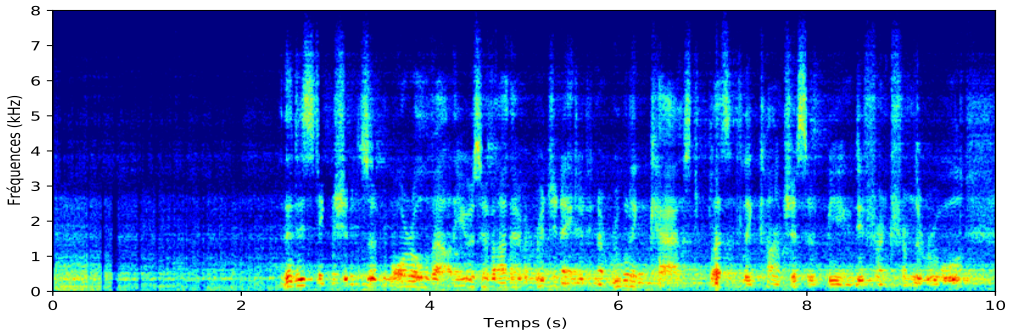
pouvant atteindre deux secondes avant de répondre à une commande. Le traitement par KISS-GEV ou KISS-MVDR hors-ligne d'une commande typique de moins de cinq secondes étant négligeable comparativement à ce délai, il est raisonnable d'utiliser un tel algorithme sur la commande complète sans encourir de délai supplémentaire perceptible à l'usager.



(a) Signal non-traité



(b) Signal rehaussé par KISS-GEV hors-ligne



(c) Signal rehaussé par KISS-GEV en-ligne

FIGURE 4.6 Spectrogrammes des 10 premières secondes d'une élocution du lecteur *masculin-complexe*, dans l'environnement *Génératrice*. Noter le très faible niveau de bruit ambiant dans (a), la déréverbération visible en région fricative (4 à 8 kHz) dans (b) et (c), et la coupure agressive de KISS-GEV en-ligne (c) entre 0 et 1 kHz.

CHAPITRE 5

CONCLUSION

Ce mémoire présente la méthode de pré-traitement KISS jumelée aux méthodes de filtrage spatial reconnues GEV et MVDR pour rehausser un signal de parole multicanal contre du bruit inconnu. KISS est une solution simple et intuitive, entièrement basée sur une approche de traitement de signal et ne nécessitant aucun entraînement, qui se résume en trois étapes. D'abord, un filtrage DS est effectué dans la direction d'intérêt et le ratio entre le signal filtré et la moyenne des canaux est calculé. Ce ratio est par la suite projeté sur un simple banc de deux filtres séparés à 3250 Hz, représentant les zones voisées et non-voisées, respectivement. Finalement les 25 % des valeurs les plus élevés sont conservées comme un « 1 » et le reste est coupé, résultant en un masque binaire représentant le signal cible. Le masque d'interférence, quant à lui, est produit de la même façon à partir des 25 % les plus faibles. Les masques temps-fréquences ainsi obtenus sont appliqués au signal bruité pour obtenir les matrices de covariance représentant la cible et l'interférence. Ces matrices peuvent ensuite être employées dans le calcul d'un vecteur de filtrage spatial GEV ou MVDR pour obtenir le signal rehaussé.

Le chapitre 3 présente la définition théorique de la méthode KISS et évalue ses performances de rehaussement lorsqu'utilisée avec le filtrage GEV en environnement simulé, en comparant le SDR obtenu par KISS-GEV à celui du filtrage populaire DS ainsi qu'au signal original. Ces tests constituent un ensemble de 6000 simulations, regroupant 20 pistes de parole cible de locuteurs variés, 20 autres pistes de parole utilisées en interférence, 20 pistes de bruit ambiant et 20 pistes de musique variée, chacune étant simulée avec 5 RIR différentes. En moyenne, pour les trois types d'interférence étudiés, KISS-GEV fournit une amélioration de SDR de 5,7 dB et de 4,9 dB comparativement au signal original et à celui rehaussé par DS, respectivement.

Ensuite, le chapitre 4 étudie la combinaison de KISS avec MVDR en plus de GEV, en faisant également la comparaison des algorithmes hors-ligne avec leurs équivalents en-ligne. Cette fois, les expérimentations sont réalisées sur des enregistrements réels en environnement industriel, réalisés par une matrice de microphones et un système embarqué portable fixé à un gilet de sécurité porté par un mannequin, pour simuler le plus réalistement possible des cas d'utilisation réels. Les enregistrements sont rehaussés par les quatre algorithmes mentionnés précédemment, en plus de DS en guise de référence. Les enregis-

trements avant et après rehaussement sont transcrits par le service *Google Speech-to-Text* et le WRR est comparé entre les méthodes. Les algorithmes GEV et MVDR hors-ligne dépassent aisément la cible de WER fixée à 80 % avec une moyenne de 86,5 % pour GEV et de 86,2 % pour MVDR. Ils dépassent cette cible pour chaque enregistement, sans exception, même dans les conditions les plus difficiles où les autres méthodes rendent une performance parfois complètement inutilisable. Les deux contreparties en-ligne de ces algorithmes fournissent en moyenne un rendement meilleur à DS, mais la dégradation de leur rendement dans certaines conditions spécifiques souligne l'augmentation de la difficulté de la tâche de rehaussement lorsqu'il est nécessaire de procéder en-ligne, laisse entrevoir des travaux futurs à perfectionner leur application plus profondément.

La complexité de calcul de la méthode KISS jumelée à GEV ou à MVDR étant faible, le délai de traitement d'une commande typique de quelques secondes est négligeable comparativement au délai de traitement de la commande par un agent conversationnel infonuagique populaire. Ainsi, l'utilisation d'un des algorithmes hors-ligne pour rehausser une commande entière qui serait délimitée préalablement par un mot de réveil et un VAD est une option viable et performante face à du bruit inconnu selon les résultats présentés par ce mémoire.

Les résultats présentés au chapitre 4 démontrent que l'utilisation en-ligne de KISS-GEV et de KISS-MVDR mériterait d'être perfectionnée davantage pour être fiable dans un grand éventail de situations. Dans leur état actuel, ils surpassent DS dès que le niveau de bruit ambiant est relativement élevé, mais leur WER se dégrade lorsque le bruit diminue. Une explication possible à ce phénomène est que les données d'entraînement des systèmes d'ASR permettent de filtrer un faible niveau de bruit, mais sont très sensibles aux distorsions en phase introduites par le filtrage par GEV ou MVDR en ligne lorsque le vecteur change de direction. Cela expliquerait pourquoi ces derniers sont dépassés par DS lorsque le bruit est faible. Une piste de mitigation qui pourrait être explorée pour cette problématique serait de sélectionner dynamiquement l'algorithme de rehaussement selon le niveau de bruit ambiant : en deçà d'un certain seuil, le signal serait simplement rehaussé par DS, pour changer pour KISS-GEV ou KISS-MVDR lorsque le bruit est plus fort.

Finalement, il serait intéressant d'étudier une application réelle en industrie, moyennant la disponibilité d'un agent conversationnel adapté à un cas d'utilisation industriel, ce qui n'était pas le cas au moment de la rédaction. Une telle étude de cas permettrait d'évaluer le gain en CSR fourni par les méthodes présentées, et soulever des améliorations potentielles plus spécifiques à l'utilisation avec un agent conversationnel qui n'ont pas pu être explorées au cours du présent projet.

ANNEXE A

RÉSULTATS DE L'ANALYSE DU WRR EN ENVIRONNEMENT INDUSTRIEL

Le tableau A.1 et la figure A.1 présentent les résultats individuels de WRR obtenus avec Google STT pour les enregistrements réalisés en environnement industriel présentés au chapitre 4.

Tableau A.1 WRR par algorithme pour chaque enregistrement

Enregistrement	**Aucun rehaussement**	**DS**	**KISS-GEV hors-ligne**	**KISS-MVDR hors-ligne**	**KISS-GEV en-ligne**	**KISS-MVDR en-ligne**	**SNR de l'enregistrement**
M, complexe, Génératrice, 68dB	74.85	73.62	84.66	84.05	40.49	40.49	9.77
M, complexe, Génératrice, 71dB	82.82	82.82	82.82	83.44	63.80	61.96	14.18
M, complexe, Usine, 72dB	68.10	69.33	82.21	80.37	74.85	61.35	1.62
M, complexe, Ventilation, 68dB	77.30	78.53	84.66	84.05	77.30	72.39	4.83
M, complexe, Ventilation, 71dB	84.66	82.21	81.60	83.44	71.78	75.46	11.03
M, simple, Génératrice, 68dB	79.10	79.10	91.54	90.05	60.70	47.76	10.92
M, simple, Génératrice, 71dB	83.08	82.09	89.05	86.57	18.91	32.34	15.94
M, simple, Usine, 72dB	76.12	69.65	91.04	88.06	55.22	55.22	1.96
M, simple, Ventilation, 68dB	72.14	72.64	87.56	88.56	59.70	71.64	2.34
M, simple, Ventilation, 71dB	89.05	87.06	90.55	90.05	28.86	63.68	12.21
F, complexe, Entrepôt, 63dB	80.23	83.62	84.75	89.83	81.92	83.62	13.63
F, complexe, Entrepôt, 66dB	86.44	86.44	87.57	86.44	81.92	82.49	17.21
F, complexe, Génératrice, 68dB	76.27	67.23	85.31	85.31	81.36	74.58	3.02
F, complexe, Génératrice, 71dB	85.88	82.49	88.14	87.57	68.93	74.01	10.22
F, complexe, Usine, 72dB	40.68	35.03	79.10	76.27	70.06	64.41	-2.48
F, complexe, Ventilation, 68dB	71.19	53.11	84.75	88.14	83.62	84.18	-2.89
F, complexe, Ventilation, 71dB	87.57	80.79	86.44	89.83	85.88	86.44	7.34
F, simple, Génératrice, 68dB	80.38	78.48	87.34	87.34	81.01	76.58	9.33
F, simple, Génératrice, 71dB	86.08	86.08	87.97	87.97	70.25	79.75	14.21
F, simple, Usine, 72dB	70.25	72.78	88.61	83.54	76.58	76.58	1.48
F, simple, Ventilation, 68dB	84.81	81.65	87.34	87.34	86.71	85.44	4.48
F, simple, Ventilation, 71dB	87.97	88.61	89.24	87.34	85.44	87.97	10.93

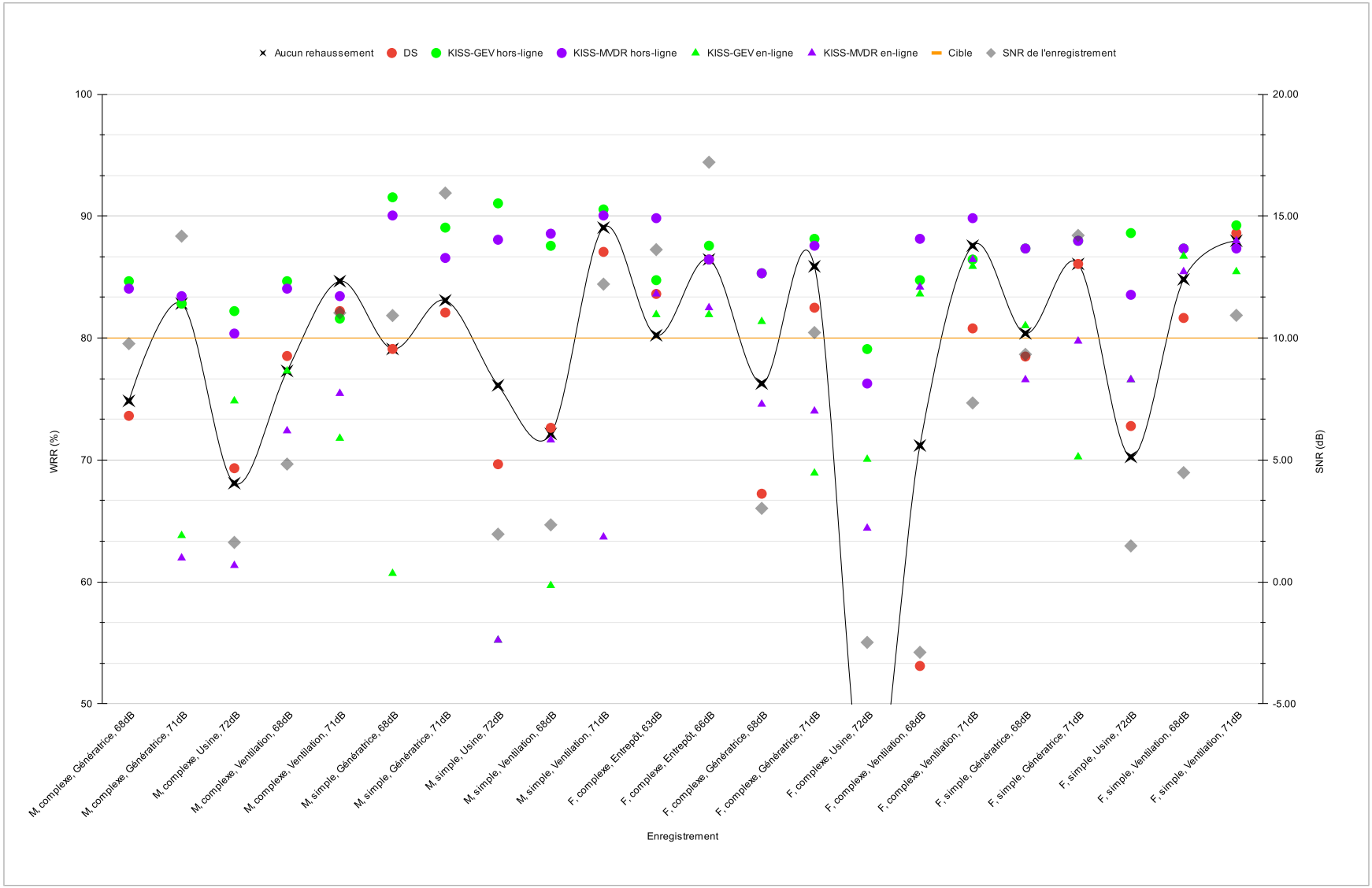


FIGURE A.1 WRR par algorithme pour chaque enregistrement

LISTE DES RÉFÉRENCES

[1] J. Allen et D. Berkley: Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65 :943–950, 1979.

[2] A. AlShehhi, M. Luai Hammadih, M. Sami Zitouni, S. AlKindi, N. Ali et L. Weruaga: Linear and circular microphone array for remote surveillance : Simulated performance analysis. *arXiv :1703.02318*, 2017.

[3] F. Bao, H. Dou, M. Jia et C. Bao: A novel speech enhancement method using power spectra smooth in Wiener filtering. Dans *Proc. Signal and Information Processing Association Annual Summit and Conference Asia-Pacific*, pages 1–4, 2014.

[4] J. Barker, R. Marxer, E. Vincent et S. Watanabe: The third 'CHiME' speech separation and recognition challenge : Dataset, task and baselines. Dans *IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 504–511, 2015.

[5] J. Barker, E. Vincent, N. Ma, H. Christensen et P. Green: The PASCAL CHiME Speech Separation and Recognition Challenge. *Computer Speech and Language*, 27(3) :621–633, 2013.

[6] S. Boll: Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27 :113–120, 1979.

[7] M.S. Brandstein et H.F. Silverman: A robust method for speech signal time-delay estimation in reverberant rooms. Dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, tome 1, pages 375–378, 1997.

[8] Centre canadien d'hygiène et de sécurité au travail Gouvernement du Canada: Limites d'exposition au bruit au Canada : Réponses SST, 2021 (visité le 2021-12-17). https://www.cchst.ca/oshanswers/phys_agents/exposure_can.html, visité le 2021-12-17.

[9] J. Chen, K. Phua, L. Shue et H. Sun: A robust adaptive cross microphone array. Dans *Proc. IEEE International Symposium on Circuits and Systems*, pages 1682–1685 Vol. 2, 2005.

[10] S. Chen, A.S. Subramanian, H. Xu et S. Watanabe: Building state-of-the-art distant speech recognition using the CHiME-4 challenge with a setup of speech enhancement baseline. Dans *Proc. Interspeech*, pages 1571–1575, 2018.

[11] Z. Chen, X. Xiao, T. Yoshioka, H. Erdogan, J. Li et Y. Gong: Multi-channel overlapped speech recognition with location guided speech extraction network. Dans *Proc. IEEE Spoken Language Technology Workshop*, pages 558–565, 2018.

[12] I. Cohen: Noise spectrum estimation in adverse environments : Improved minima controlled recursive averaging. *IEEE Transactions on Speech and Audio Processing*, 11(5) :466–475, 2003, ISSN 1063-6676.

[13] I. Cohen et B. Berdugo: Noise estimation by minima controlled recursive averaging for robust speech enhancement. *IEEE Signal Processing Letters*, 9(1) :12–15, 2002.

[14] H. Erdogan, J.R. Hershey, S. Watanabe, M.I. Mandel et J. Le Roux: Improved MVDR beamforming using single-channel mask prediction networks. Dans *Proc. Interspeech*, pages 1981–1985, 2016.

[15] N. Fan, J. Rosca et R. Balan: Speech noise estimation using enhanced minima controlled recursive averaging. Dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pages IV–581–IV–584, 2007.

[16] Z. Goh, K.C. Tan et T.G. Tan: Postprocessing method for suppressing musical noise generated by spectral subtraction. *IEEE Transactions on Speech and Audio Processing*, 6(3) :287–292, 1998.

[17] E.M. Grais, M.U. Sen et H. Erdogan: Deep neural networks for single channel source separation. Dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3734–3738, 2014.

[18] F. Grondin et J. Glass: SVD-PHAT : A fast sound source localization method. Dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4140–4144, 2019.

[19] F. Grondin, J. S. Lauzon, S. Michaud, M. Ravanelli et F. Michaud: BIRD : Big Impulse Response Dataset. *arXiv :2010.09930*, 2020.

[20] F. Grondin, J. S. Lauzon, J. Vincent et F. Michaud: GEV Beamforming supported by DOA-based masks generated on pairs of microphones. Dans *Proc. Interspeech*, pages 3341–3345, 2020.

[21] F. Grondin, D. Létourneau, F. Ferland, V. Rousseau et F. Michaud: The ManyEars open framework : Microphone array open software and open hardware system for robotic applications. *Autonomous Robots*, 34(3) :217–232, 2013.

[22] F. Grondin, H. Tang et J. Glass: Audio-visual calibration with polynomial regression for 2-D projection using SVD-PHAT. Dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4856–4860, 2020.

[23] E.A.P. Habets, J. Benesty, I. Cohen, S. Gannot et J. Dmochowski: New insights into the MVDR beamformer in room acoustics. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 18(1) :158–170, 2009.

[24] J. Heymann, L. Drude, C. Boeddeker, P. Hanebrink et R. Haeb-Umbach: Beamnet : End-to-end training of a beamformer-supported multi-channel ASR system. Dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5325–5329, 2017.

[25] J. Heymann, L. Drude, A. Chinaev et R. Haeb-Umbach: BLSTM supported GEV beamformer front-end for the 3rd CHiME challenge. Dans *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, pages 444–451, 2015.

[26] J. Heymann, L. Drude et R. Haeb-Umbach: Neural network based spectral mask estimation for acoustic beamforming. Dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 196–200, 2016.

[27] X. Huang, A. Acero, H. W. Hon et R. Reddy: Spoken Language Processing : A Guide to Theory, Algorithm, and System Development. Prentice Hall PTR, 2001, ISBN 978-0-13-022616-7.

[28] IntRoLab: 16SoundsUSB. GitHub, 2017 (visité le 2020-09-15). https://github.com/introlab/16SoundsUSB.

[29] H. G. Kang, M. Graczyk et J. Skoglund: On pre-filtering strategies for the GCC-PHAT algorithm. Dans *Proc. IEEE International Workshop on Acoustic Signal Enhancement*, pages 1–5, 2016.

[30] E. Kim, K. Nakadai et H. Okuno: Improved sound source localization in horizontal plane for binaural robot audition. *Applied Intelligence*, 42 :63–74, 2014.

[31] S. Kim, T. Hori et S. Watanabe: Joint CTC-attention based end-to-end speech recognition using multi-task learning. Dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4835–4839, 2017.

[32] K. Kinoshita, M. Delcroix, S. Gannot, E.A. P. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, A. Sehr et T. Yoshioka: A summary of the REVERB challenge : State-of-the-art and remaining challenges in reverberant speech processing research. *EURASIP Journal on Advances in Signal Processing*, (1) :7, 2016.

[33] D. Lee, B. Jang, S. Im et J. Song: A new sound source localization approach using stereo directional microphones. Dans *Proc. IEEE International Conference on Information Communication and Signal Processing*, pages 504–508, 2019.

[34] M. Liang, L. Xi-Hai, Z. Wan-Gang et L. Dai-Zhi: The generalized cross-correlation method for time delay estimation of infrasound signal. Dans *Proc. International Conference on Instrumentation and Measurement, Computer, Communication and Control*, pages 1320–1323, 2015.

[35] D.J. Maguire et R.L. St-Pierre: The impact of A-weighting sound pressure level measurements during the evaluation of noise exposure. Dans *Proc. National Conference on Noise Control Engineering*, pages 702–708, 2004.

[36] A.C. Morris, V. Maier et P. Green: From WER and RIL to MER and WIL : Improved evaluation measures for connected speech recognition. Dans *Proc. Interspeech*, pages 2765–2768, 2004.

[37] A. Narayanan et D. Wang: Ideal ratio mask estimation using deep neural networks for robust speech recognition. Dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7092–7096, 2013.

[38] V. Panayotov, G. Chen, D. Povey et S. Khudanpur: Librispeech : An ASR corpus based on public domain audio books. Dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5206–5210, 2015.

[39] Y. Park, S. Patwardhan, K. Visweswariah et S.C. Gates: An empirical analysis of word error rate and keyword error rate. Dans *Proc. Interspeech*, pages 22–26, 2008.

[40] N. Peters et A.W. Schmeder: Beamforming using a spherical microphone array based on legacy microphone characteristics. Dans *Proc. International Conference on Spatial Audio*, page 7, 2011.

[41] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer et K. Vesely: The

Kaldi Speech Recognition Toolkit. Dans *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, 2011.

[42] L.R. Rabiner et R.W. Schafer: Theory and Applications of Digital Speech Processing. Pearson, Upper Saddle River, 1st édition, 2011, ISBN 978-0-13-603428-5. 1056 pages.

[43] D.V. Rabinkin, R.J. Renomeron, J.C. French et J.L. Flanagan: Optimum microphone placement for array sound capture. *The Journal of the Acoustical Society of America*, 101(5) :3114–3114, 1997.

[44] D.R. Reddy: Speech recognition by machine : A review. *Proceedings of the IEEE*, 64(4) :501–531, 1976.

[45] H. Sak, A. Senior, K.a Rao, F. Beaufays et J. Schalkwyk: Google Voice Search : Faster and More Accurate. Google AI Blog, 2015 (visité le 2020-09-15). http://ai.googleblog.com/2015/09/google-voice-search-faster-and-more.html.

[46] M.N. Schmidt, J. Larsen et F. T. Hsiao: Wind noise reduction using non-negative sparse coding. Dans *Proc. IEEE Workshop on Machine Learning for Signal Processing*, pages 431–436, 2007.

[47] J. Schröder, T. Rohdenburg, V. Hohmann et S.D. Ewert: Classification of reverberant acoustic situations. Dans *Proc. International Conference on Acoustics*, page 4, 2009.

[48] M.L. Seltzer, D. Yu et Y. Wang: An investigation of deep neural networks for noise robust speech recognition. Dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.

[49] K. Shimada, Y. Bando, M. Mimura, K. Itoyama, K. Yoshii et T. Kawahara: Unsupervised speech enhancement based on multichannel NMF-informed beamforming for noise-robust automatic speech recognition. *Proc. IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(5) :960–971, 2019.

[50] D. Snyder, G. Chen et D. Povey: Musan : A music, speech, and noise corpus. *arXiv :1510.08484*, 2015.

[51] I. Szoke, M. Skacel, L. Mosner, J. Paliesek et J. Cernocky: Building and evaluation of a real room impulse response dataset. *IEEE Journal of Selected Topics in Signal Processing*, 13 :863–876, 2019.

[52] H. Tang, W. N. Hsu, F. Grondin et J. Glass: A study of enhancement, augmentation, and autoencoder methods for domain adaptation in distant speech recognition. Dans *Proc. Interspeech*, page 2928–2932, 2018.

[53] K. Varma: Time-Delay-Estimate Based Direction-of-Arrival Estimation for Speech in Reverberant Environments. Thèse de doctorat, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA, 2002. 127 pages.

[54] M. Wax et A. Leshem: Joint estimation of time delays and directions of arrival of multiple reflections of a known signal. *IEEE Transactions on Signal Processing*, 45(10) :2477–2484, 1997.

[55] F. Weninger, J. Le Roux, J.R. Hershey et S. Watanabe: Discriminative NMF and its application to single-channel source separation. Dans *Proc. Interspeech*, pages 865–869, 2014.

[56] X. Xiao, S. Zhao, D.L. Jones, E.S. Chng et H. Li: On time-frequency mask estimation for MVDR beamforming with application in robust speech recognition. Dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3246–3250, 2017.

[57] D. Ying, J. Li, Y. Feng et Y. Yan: Direction of arrival estimation based on weighted minimum mean square error. Dans *Proc. IEEE China Summit and International Conference on Signal and Information Processing*, pages 318–321, 2013.